# EXPLORATORY STUDY INTO THE RELATIONSHIP BETWEEN STATISTICAL MEASURES OF KEYWORD QUALITY AND CITATION RATES FOR OPEN ACCESS ACADEMIC PAPERS

This dissertation was submitted in part fulfilment of requirements for the degree of MSc Information Management

DEPT. OF COMPUTER AND INFORMATION SCIENCES UNIVERSITY OF STRATHCLYDE

August 2017

Neil Wells

# DECLARATION

This dissertation is submitted in part fulfilment of the requirements for the degree of MSc of the University of Strathclyde.

I declare that this dissertation embodies the results of my own work and that it has been composed by myself. Following normal academic conventions, I have made due acknowledgement to the work of others.

I declare that I have sought, and received, ethics approval via the Departmental Ethics Committee as appropriate to my research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to provide copies of the dissertation, at cost, to those who may in the future request a copy of the dissertation for private study or research.

I give permission to the University of Strathclyde, Department of Computer and Information Sciences, to place a copy of the dissertation in a publicly available archive.
(please tick)

Yes [ x ]
No [ ]

I declare that the word count for this dissertation (excluding title page, declaration, abstract, acknowledgements, table of contents, list of illustrations, references and appendices is .

I confirm that I wish this to be assessed as a

Type 1 2 3 4 5

Dissertation (please circle)

Signature:

Date:

## ABSTRACT

An exploratory study investigating the relationship between two different versions of a particular statistical measure of metadata quality applied to Author Keywords (Ochoa and Duval's *QtInfo*) and citation rates for Open Access papers relating to diabetes, recorded in the Scopus citation index.

17,451 articles were harvested from the CORE.ac.uk Open Access aggregator and citation counts and keywords for 3,588 of these were taken from the Scopus citation index and matched to the records from CORE, and analysed using R.

A variety of linear, quasi-Poisson and Negative Binomial regression models were applied, and most notably a highly statistically significant ($p = 0.000109$ ) and somewhat predictive ($R^2 = 0.01075$) positive relationship was found between the $Qtinto_{kw}$ measure and citation rates, outperforming the related, previously investigated measure of Author Keyword count by an order of magnitude both in terms of significance and variance described.

# TABLE OF CONTENTS

# TABLE OF FIGURES

# 1 INTRODUCTION

The rate of publication of scientific and other academic research is currently increasing at an exponential rate(Bornmann and Mutz, 2015)- While in the 17th and 18th Centuries it was possible for an educated person to keep abreast of all major scientific developments, so much is now published that it is impossible for one person, even within relatively specialised academic fields, to comprehensively survey and keep abreast of all research- for example, the Elsevier citation index Scopus lists a total of 2,318 articles published in 2016 relating to 'nephrology', and 2,502 relating to 'photovoltaics'. With 365 days in a year, it is difficult to imagine even the most assiduous researcher reading 6 or 7 papers per day every day (or having much time left to do anything else).

When searching for background material and other relevant material in the course of conducting research and literature reviews, the use of academic database search, using services such as Web of Science, Google Scholar and Scopus, is therefore one increasingly important means of discovery for relevant background material, with bibliographic and citation databases used as a primary search tool for 45% of researchers surveyed in 2007 (Hemminger *et al.*, 2007). While database search is very far from the only means of discovery for academic work (Medoff, 2006) it is, and will likely remain, a very important means of exploring and surveying an increasingly complex and difficult-to-navigate academic information space.

Metadata, or 'data about data' in bibliographic citation indexes contains a wealth of descriptive information on articles in the database, ranging from title, date of publication, publishing journal, authors' names, the authors' associated institutions, identifiers of various types including DOI (Digital Object Identifier) and internal unique document identifiers specific to the database, as well as associate keywords describing and disambiguating articles from their peers, as well as citation information which bibliometricians increasingly use to map networks of intellectual influence.

Initially academic databases were accessed via Boolean Search alone, requiring mastery of search syntax and exact matching to return any results. While academic databases differ from Web search engines such as Google or Bing in having a user base who can reasonably be expected to learn the comparatively complex and inflexible search syntax required to retrieve results from these databases, ranking algorithms, such as the tf-idf measure utilised in this study (although the precise nature of the ranking algorithms are proprietary), are also used to prioritise and rank search results in a manner analogous in presentation (if not functionally) to the results returned by Web search engines such as Google.

It will be the everyday experience of everyone familiar with academic work that metadata attached to documents, whether it be accurate titles and author information or accurate document identifiers are invaluable in retrieving documents- one need only imagine the nightmare of an academic library with no filing system to appreciate the importance of properly labelled and ordered documents to any kind of academic endeavour.

When searching an academic database such as Scopus for relevant material, the match between the keywords attached to the article and the search terms input into the database is, it is assumed, of importance in ensuring that relevant material is retrieved.

*Hypothesis 1:* Higher-quality (more descriptive) keywords enable articles to be found more efficiently, both by enabling articles to be retrieved, and by boosting the position of retrieved results in search rankings when many papers are retrieved.

Citation rates are variously defined as signifying various qualities of the individuals, institutions and articles to which they are attached, but in our current study its most pertinent quality is as a *signifier of successful retrieval*- that is, a citation of an academic paper in another indicates that, during the information search procedure carried out in the writing of the citing paper, the cited paper *was successfully retrieved* (although it does not tell us the means by which it was retrieved). Simply put, a paper cannot be cited if nobody has ever found it.

If articles are found more efficiently by academics performing search, then greater quality keywords this should also have a measurable (if modest) positive impact on citation rates for those papers which boast highly descriptive keywords:

*Hypothesis 2:* Higher-quality keywords results in more efficient discovery of articles of interest, and therefore result in higher citation rates for those articles.

The assessment of metadata quality, however, is not straightforward- what constitutes high-quality metadata is relative not only to the individual documents, but to the collections of which they form part, the needs of those trying to find documents, the historical context in which this all takes place, and many other factors. A reductive, if slightly circular definition might be that 'metadata is high quality if it reliably allows successful retrieval of an object'.

Although identifying high-quality metadata is a difficult and subjective task, that is not to say that the task is completely subjective- inaccurate, garbled, or simply missing metadata is objectively *bad* metadata, so there is at least the potential for the establishment of an objective measure of a minimum standard of acceptable metadata quality, even if the assessment of higher-quality metadata remains a subjective matter.

Manual assessments of metadata quality have by and large been relatively low-volume due to the highly labour-intensive nature of the work- studies conducted thus far have involved the recruitment of dozens of volunteers to spend many hours reading and appraising documents, and the volume of these studies has consequently been very low in comparison to the enormous and increasing volume of academic literature appearing.

Since such a large variety of factors are to be expected to affect citation rates, the impact of even the highest-quality of metadata on citation rates can only be expected to account for a small proportion of the variance of any model investigating a link between these properties. Other factors, both social and bibliometric, as well (one hopes) as the quality of the research described in the article, will have a large effect on the citation rate. Since the effect is therefore likely to be relatively subtle, to definitively establish any correlation between a measure of metadata quality and citation counts will involve the examination of many thousands of documents, a fact which no doubt in large part accounts for the lack of any such investigation.

The systematic investigation of any possible link between keyword quality and citation rates therefore requires some form of automatic or statistical measure metadata quality assessment, which can be calculated in a relatively straightforward manner and applied to a large number of documents in order to derive a robust and repeatable measure of quality which can be reliably applied to large numbers of documents.

Statistical measures of metadata quality are relatively new, and the advent and increasing popularity of Open Access (OA) over proprietary (or paywalled) access to academic literature provides opportunities for text mining and statistical analysis of academic literature in a way which would hitherto have been difficult or impossible to achieve using paywalled literature due to the ease of access and manipulation of data.

The relationship between statistical measures of metadata quality and actual metadata quality is also not easy to define, and so a more limited, but easier to investigate, third hypothesis might be:

***Hypothesis 3:*** If higher quality keywords result in more citations, then a positive correlation will be found between *Qtinfo* and citation rates.

# 2 LITERATURE REVIEW

## 2.1 METADATA AND KEYWORD QUALITY IN DIGITAL REPOSITORIES

Because of the wide variety of academic information stored and accessed electronically, from large sets of observations and experimental data in disciplines such as astronomy or bioinformatics, to archived image libraries, videos, articles of various types, books among others, there exists a profusion of hundreds of metadata standards for academic information, broadly defined, with standards constantly evolving.

A metadata scheme consists of a set of standard metadata elements to be applied consistently across a repository or collection.

Dublin Core is a standardised metadata scheme consisting of a core of 15 metadata elements suitable for describing digital documents and other digital objects, which has been adopted across a wide variety of platforms and has been ratified as ISO Standard 15836:2009. Dublin Core elements have been adopted by the Open Archives Initiative as the basis for interoperable metadata standards for online archives, and form part of both the Scopus and Web of Science metadata schemes.

Ward (Ward, 2003) conducted a survey of the utilization of Dublin Core metadata elements across a number of digital repositories registered to the Open Archives Initiative, revealing enormous variation and a high degree of underutilization of the 15 main Dublin Core metadata elements. Due to the heterogeneous nature of both the documents and of the metadata sources (some provided and validated by information professionals, some by authors, and some automatically generated), the completeness of metadata varied enormously, with some fields such as 'subject' missing values entirely in over 20% of cases. Just two fields (creator and identifier) accounted for approximately 50% of metadata field utilization for approximately 50% of the data providers surveyed.

Windnagel(Windnagel, 2014) surveys use of Dublin Core elements in 3 Mathematics and Science-based digital repositories manually, again indicating a bias toward heavy utilisation of a small subset of metadata elements , although only a very small sample of records (c.75) were reviewed.

Hughes(Hughes, 2004) reports an early attempt at algorithmic metadata quality evaluation in the context of the Open Language Archives Community, concluding that such measures have the potential to assist manual metadata creation by assisting identification of areas where metadata quality is low.

There have been a number of attempts to systematically assess the quality of metadata in digital repositories: Bruce and Hillman (Hillmann, 2004) propose categorization of metadata quality according to seven characteristics: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility.

> *Completeness:* Metadata should be complete both in terms of its description of the object to which it applied, and in its application to the object.

*Accuracy:* Metadata should accurately describe the object, and be in itself accurate- free of typographical errors.

*Provenance:* a measure of quality of the source of the metadata, its compliance to standards and creation and handling methodologies.

*Conformance to expectations:* Metadata should be appropriate to the intended audience, structurally conform to the expectations and likely search behaviour of that audience, and

*Logical consistency and coherence:* A particular problem for digital repositories which may contain objects from a large variety of sources, the need for consistency and coherence of metadata within a collection. Without this, similar objects in a collection may only be retrievable using different search terms, leading to imperfect retrieval.

*Timeliness:* metadata over time can become detached from the objects to which it is applied. Retrieval is negatively affected if this occurs. Conversely, there may also be a 'lag' effect where metadata is not applied until some time after the object creation. In both of these cases, metadata utility, and hence retrieval, is negatively affected.

*Accessibility:* metadata must also be easily accessible and readable. Metadata which is physically or logically detached from the objects which it describes, or which is otherwise inaccessible to the user, is of little use.

Ochoa and Duval propose a series of statistical measures based on this set of criteria by which some of these measures can be automatically assessed, and assess the efficacy of these measures by comparison with a manual assessment of metadata quality. Although most of these statistical measures were found to correlate poorly with human assessment, a small number of measures (particularly their measure of metadata information content, *Qtinfo*) correlated well ($R^2 > 0.8$ ) with human judgements of metadata quality.

The *Qtinfo* statistic is based on the *tf-idf* measure, in turn based on the Term Frequency (the number of occurrences of a particular term in a document) and the Inverse Document Frequency (the inverse of the number of documents in a collection in which a term occurs) measure proposed by Sparck-Jones (Sparck-Jones, 1972), which has proven to be a very strong heuristic measure of term significance with wide application, although it lacks any compelling theoretical justification(Stephen, 2004)

Ochoa and Duval conclude from this work that although full assessment of metadata quality is not something which can at present be automated, automatic detection of instances of *low-quality* metadata according to these measures may be possible. Furthermore, since these measures are (comparatively) straightforward to calculate (since they are based on well-known and understood statistical measures, primarily term frequency-inverse document frequency) and are independent of a particular corpus these measures may be sensitive enough to automatically classify objects by metadata quality, even if the assessment is not up to human standards. (Ochoa and Duval, 2009)

Gavrilis *et al* explore the concept of metadata quality, noting that further develop these measures into a more comprehensive Metadata Quality Evaluation Model (MQEM), designed as a framework intended to enable automatic assessment of metadata quality, although the framework has not been validated. (Gavrilis *et al.*, 2015)

Most applications of statistical measures of metadata quality have been experimental in nature and intended eventually to supplement and validate manual and automatic metadata generation for repositories. Margaritopoulos *et al* propose automatically-evaluable measures of metadata completeness(Margaritopoulos *et al*., 2012*)*, while Tsiflidou and Manouselis assess the efficacy of

three software tools (Google Refine, MINT, and their own proprietary tool) for calculating these statistical measures. (Tsiflidou and Manouselis, 2013)

Inacio *et al* present an ontology-based automatic metadata analysis tool for bioinformatics metadata, taking advantage of the semantic features of bioinformatic databases to assess the quality of metadata in a more semantically rich fashion, analysing the quality of records based on two criteria: term coverage and term specificity, finding overall poor levels of semantic density in metadata and crediting this to poor author awareness of available ontologies (Inactio *et al,* 2017).

## 2.2 CITATION INDEXES- COVERAGE AND LIMITATIONS

A citation index is a type of bibliographic index allowing users to determine citation relations between documents, usually academic works. Initially published in paper form, citation indexes now take the form of searchable databases containing comprehensive records of article names, authors, abstracts, keywords, citations, and other metadata for a wide range of academic journals, although no citation index claims complete coverage of all published academic literature. The two dominant citation indexes extant today are Web of Science (WoS) and Scopus, the former the descendent of Eugene Garfield's 1963 Science Citation Index.

Mongeon and Paul-Hus compare the coverage of WoS (13,605 journals) with Scopus (20,346 journals) and conclude that both exhibit a both marked bias toward STEM literature and toward English-language material, compared to Ulrich's periodical directory, and both exhibit differing biases which may lead to differing outcomes for bibliometric analyses depending on choice of citation index. Importantly, they also remark that these citation indexes focus primarily on journal articles and not on other forms of academic literature, such as books, which are more prevalent in the humanities. Citation analysis of impact using these indexes for humanities subjects may therefore be less relevant than for STEM subjects. (Mongeon and Paul-Hus, 2016)

De Groote *et al* examine the coverage of Scopus, WoS and Google Scholar in calculating the h-index citation metric in the field of nursing and conclude that multiple citation indexes should be used when calculating the h-index (and by extension other citation rate-based measures) due to the high variability in outcomes for the measured metric. (De Groote and Raszewski, 2012)

Similarly, Harzing, in a comparison of Google Scholar, WoS and Scopus, concludes that choice of citation index can have a large impact on the outcomes of bibliometric measures and that therefore combining results from multiple sources yields the best results. (Harzing and Alakangas, 2016)

The CORE.ac.uk aggregator harvests Open Access article fulltext in a variety of different fashions, a necessity due to the wide variety of manners in which publishers make Open Access fulltext content available, ranging from FTP access through programmatic API access, to simple presentation of article text in html or PDF format at a publicly available address. Furthermore, DOI references resolve either to article fulltext or to 'splash pages' in an inconsistent fashion. The current interoperability challenges presented by the lack of standardisation on the presentation of Open Access materials presents a considerable challenge for text and data mining (TDM) efforts. (Knoth and Pontika, 2016)

Access to proprietary, paywalled article text for TDM presents even greater challenges- as well as the considerable technical challenges presented by the need for authentication, also potentially

require significant reform of copyright law in order to allow 'fair use' applications of TDM for the purposes of research. (European Commission, 2014)

## 2.3 CITATION ANALYSIS

Citation analysis is the measurement and study of citation patterns between documents, usually academic journal papers. Gilbert(Gilbert, 1977) considers citation counts important as a measure of academic authoritativeness, whereas Martin and Irvine consider them to measure intellectual influence.

In introducing the concept of the citation index, Garfield introduced the notion that a reference is itself a form of index term or subject heading, semantically linking the cited work and placing it in a context with other research (Garfield, 1964)

Small (Small, 1978) extends this concept, viewing the process of citation as one of the construction of a symbolic meaning for the cited document, with each cited document representing a particular concept external to the paper's content consistently across different citations.

Whatever the precise defining, a high citation rate is, from the point of view of the author and the institutions associated with the author, a desirable characteristic for a paper to have and is highly correlated with other measures of academic excellence.

Citation rates and associated derived bibliometric measures have long been used as a proxy for the quality of scholarly journals (Garfield, 2006), and are now used as a proxy for research and productivity quality both for groups (Mryglod *et al.*, 2013) and individuals (Duffy *et al.*, 2008). Exercises such as the UK Research Excellence Framework (HM Government, 2014) (Taylor, 2011), Australia's Excellence in Research for Australia (ERA) (Australian Research Council, 2017) and the New Zealand Performance-Based Research Fund (PBRF) (Anderson,Smart and Tressler, 2013) use citation data in assessing the quality of publicly-funded research.

Metrics such as the h-index(Hirsch, 2005) and the g-index (Egghe, 2006) use citation information to quantify the research output of individuals, while article-level metrics (ALM) (Handel, 2014) aim to quantify the wider impact of an article beyond academic citations by including information about the wider cultural impact of an article including blog citations, Twitter mentions and other measures.

Investigations have been performed into the effect on citation count of various metadata and text-related properties of academic papers, particularly into the effect of the properties of the title on citation counts, on which many papers have been published (e.g.(Nair and Gibbert, 2016). Letchford *et al* found that papers with shorter titles received more citations. (Letchford,Moat and Preis, 2015)

Uddin and Khan (Uddin and Khan, 2016) investigated the effect of Author Keyword selection and various measures of their diversity, including keyword diversity, number of new keywords, and total keyword number, finding a positive correlation for many of these measures with citation counts.

Sohrabji and Iraj introduce two new keyword-related measures- abstract ratio (the sum of the repetition of keywords in abstract divided by abstract length) and the weight ratio (the frequency of a paper's keyword per journal) and find that both are positively correlated with citation rates in a study of education-related literature. (Sohrabi and Iraj, 2017)

Gazni (Gazni, 2011) analysed the Flesch readability score of journal article abstracts, concluding that increased abstract readability was negatively correlated with citation count (or in other words that a less readable, more complex abstract results in higher citation rates), whereas conversely Letchford *et al* found that a shorter abstract was positively correlated with citation counts (Letchford,Preis and Moat, 2016)

Haslam *et al* more broadly examined aspects affecting citation rates including institutional factors, author eminence and research approach, finding that eminence of first author, later author seniority, journal prestige, article length, and number and recency of references were all predictors of citation impact. (Haslam *et al.*, 2008)

Falagas *et al* also found a positive correlation between article length and citation count in their multivariate analysis of various variables (number of authors, article length, study design (interventional/observational and prospective/retrospective), title and abstract word count, number of author-affiliated institutions, and number of references) in general medicine journals. Over 50% of total variance of citation counts was explained by these factors, with article length and Journal Impact Factor the only variables to independently predict citation rates. (Falagas *et al.,* 2013 )

Moed (Moed, 2005) examines the accuracy of citation counts for Web of Science records between 1980-2004 and discovered an approximately 7% rate of 'discrepant' citations, with citations either not counted due to e.g. incorrect citation formatting, mis-spelling or transliteration of foreign names by researchers unfamiliar with naming conventions, discrepancies between digital and physical versions of articles, issues due to inconsistencies with journal issue numbering.

Citation rates tend to peak between 2 and 7 years after a paper's publication, aside from a very small number of highly influential papers which are highly-cited very soon after publication (Brzezinski, 2015). A number of different mathematical distributions have been proposed to fit the distribution of citation rates, including power law, discretised log-normal and hooked power-law, (Thelwall, 2016) and stopped-sum hybrid distributions modelling the effect of multiple causative processes on the distribution of citation rates over time. (Low,Wilson and Thelwall, 2016)

After 7 years the rate of citations tends to decline, with the exception of a small number of 'Sleeping Beauties' which gather a spike in citations after a long period of garnering few or no citations (van Raan, 2004). It is not entirely clear whether the forms of information-seeking behaviour by which these 'Sleeping Beauties' are retrieved differs from the methods used to retrieve other papers- however, if the hypothesis that higher-quality metadata aids retrieval and thus citation rates is true, then these 'sleeping beauty' papers may be worthy of study in this context since by virtue of their age (thus making 'monitoring' a less likely means of discovery) and lack of previous citations (which lessens the impact of citation chaining) they are candidates for 'success stories' of retrieval by direct database search aided by high-quality metadata.

As Garfield notes: "*Mendel's experiments with peas in his monastery garden, Fleming's observations of bacterial lysis in mold-contaminated petri dishes, Pressey's reports of "An Apparatus which . . . teaches" all lay buried in dusty tomes for decades before their vast significance for genetics, antimi- crobial therapy, and teaching machines became widely recognized. Indeed, the history of science abounds with ex amples indicating that the scientific community is incapable of quickly absorbing radically new ideas or information. If we assume that the papers which have never been cited include those which were ahead of their time, the citation index may afford a means of ferreting out those papers which might deserve reevaluation, redissemination, or even republication."* (Garfield, 1964)

Length of availability also impacts citation count in that citation count is an aggregative phenomenon and the longer a paper has been available, the more time it has to pick up citations. Uddin *et al* propose a normalised citation count measure to account for this and enable papers of different vintages to be compared. (Uddin *et al.*, 2012)

Rousseau and Vazirgiannis suggest an alternative to simple term frequency-based weighting systems, document representation based on an unweighted directed graph-of-words, thus allowing a richer depiction of documents in terms of relationships with relatively little computational overhead. (Rousseau and Vazirgiannis, 2013)

Shirakawa *et al* highlight difficulties in applying tf-idf weightings to word N-grams, highlighting a tendency disproportionately to favour 'awkward' or unusual term formulations when calculating weightings, and propose an alternative IDF formulation which applies a  (Shirakawa,Hara and Nishio, 2017).

## 2.4 ACADEMIC INFORMATION SEEKING AND RETRIEVAL BEHAVIOUR

For an article to be cited in a paper, the author of the paper must first successfully retrieve it. Academics use a variety of behaviours described by Ellis *et al* (Ellis,Cox and Hall, 1993) to find and retrieve, including 'monitoring' (in this instance monitoring journals for articles of relevance) and 'chaining' (following patterns of citation back through successive papers). Retrieval of papers for citation is likely to occur at least partly through these mechanisms (where metadata quality is likely to have less or no impact), and partly through database searching (where metadata quality would seem intuitively to be likely to have a greater impact). However, almost all academics use direct keyword-based database searching, either via Google or via their library interface, to access digital papers in repositories.

Keywords attached to articles provide a summary guide to the content of an article, and are provided both by authors (Author Keywords) and professional indexers (Index Keywords). In one study, a 46% overlap was found between Author Keywords and indexer-assigned descriptors. (Gil-Leiva and Alonso-Arroyo, 2007)

Keywords represent the authors' conception of the meaning of their work within the wider context of the academic discipline within which they work- they serve both to effectively describe the work and to distinguish it from other work in the area.

Few if any quantitative studies have been published on the effect of author keywords on document retrieval rates, but the single study suggested that author keywords outperformed applied user tags. (Lu and Kipp, 2014)

# 3 RESEARCH METHODS

## 3.1 EQUIPMENT AND SOFTWARE USED

Analysis was undertaken on a Core 2 Quad Q6600 Windows 10 PC with 8GB DDR3 RAM and a 1TB hard disk drive.

R 3.4.0. was used for data preparation and numerical analysis.

R Packages used include:

**dplyr:** used as part of **tidytext** for manipulation of tidy text tables.

**fulltext:** used as part of initial exploration to retrieve full text objects from various Open Access sources

**rjson :** used for parsing JSON files retrieved from CORE.ac.uk

**rscopus :** used to retrieve article metadata records from Scopus: citation counts and keywords

**rcoreoa:** used to retrieve article records and full text from CORE.ac.uk

**tm:** used for for stemming text

**tidytext:** used for conversion of full text into tidy text format, and for calculation of tf*idf weightings

## 3.2 CALCULATION OF VARIABLES

### 3.2.1 Independent variables: $Qtinfo_{token}$ and $Qtinfo_{kw}$.

The Information Content of a metadata field is calculated using the tf-idf values:

$$infoContent_{term} = \sum_{i=1}^{N} tf_i * \log\left(\frac{1}{df_i}\right) \quad (1)$$

The primary independent variable is *Qtinfo.* For a set of metadata with *N* fields,

$$Qtinfo = \log\left(\sum_{i=1}^{N} infoContent_i\right) \quad (2)$$

In Duval and Ochoa's paper (Ochoa and Duval,2009) there is some ambiguity as to the intention regarding the basis on which the *infoContent* measure should be calculated, due to a certain ambiguity inherent in the way the term 'Keyword' is used in bibliographic databases. 'Keyword' is in fact something of a misnomer, since many of the terms used in keyword fields consist of 2, 3 and in rare cases up to 9 word phrases.

The standard approach for applying tf-idf values here is to decompose the document into a 'bag of word' and calculate term frequencies on this basis: in other words, the tf-idf (*infoContent)* weighting (and hence for a keyword phrase such as 'Body Mass Index' should be calculated on the basis of the term frequencies and inverse document frequencies of the individual tokens 'Body', 'Mass', and 'Index' , rather than on the basis of the single string 'Body Mass Index'.

For the purposes of this analysis, the variable $Qtinfo_{token}$ refers to *Qtinfo* calculated on this per-token basis.

An alternative approach is to calculate the *infoContent* value on the basis of each keyword *field*- therefore if a keyword field contains the string 'Body Mass Index' the term and inverse document frequencies are calculated for that string rather than for the individual constituent tokens.

It should be apparent from this example that the resultant values obtained by calculating the same measure on this basis will be *very* different indeed- 'Body', 'Mass' and 'Index' are all general terms which might be expected to occur separately over a great variety of STEM literature in a huge variety of different contexts, while the latter is a highly specific term with potentially great discriminatory value- the *term independence* assumption which underlies the calculation of tf-idf values may in fact be particularly unhelpful in this context.

### *3.2.2* Independent variables: *token.count* and *Keyword.count*

These quantities simply represent the number of keyword tokens (*token.count)* and keyword field values (*Keyword.count)* for each article. For an article with two attached keywords, "Body Mass Index" and "Obesity", for example *token.count* has a value of 4 (tokens "Body", "Mass", "Index", "Obesity") while *Keyword.count* has a value of 2 ("Body Mass Index" and "Obesity").

### 3.2.3 Dependent Variable: *cited.by*

Dependent variable is Citation Count, as acquired from Scopus, and is a simple integer count measure of the number of citations received by an academic paper in other academic papers. The measure was used as is, except for various transformations, summarised in Table 10, applied in order to make the highly skewed distribution of citation counts approximate a normal distribution for the purposes of linear regression modelling.

## 3.3 DATA ACQUISITION- AVENUES EXPLORED AND ABORTIVE ATTEMPTS

### 3.3.1 General issues with full text acquisition

Since calculation of the *Qtinfo* parameter requires calculation of access to the full text of each article in order to calculate term and document frequencies, finding a broad and easy-to-acquire set of article full text was a primary consideration for this analysis. Ideally analysis would be performed on all extant published articles, whether paywalled or not. However, the technical and legal obstacles to this approach are considerable.

Corpuses prepared for the purpose of text mining do exist- for example, the Elsevier Text Mining Corpus  comprises a set of 110 STEM papers prepared as plain text for the purposes of text mining- however, for the purposes of the current analysis this sample size is at least an order of magnitude too small for

However, this introduces a level of extra complexity given the fact that each commercial publisher has their own authentication system to ensure access to their content.

For example, although the Scopus API does allow retrieval of Elsevier-published article fulltext, it does this by providing a link to the article text, which may be either a html page or a PDF. Choosing this route would also have limited the choice of papers to those from a single publisher.

### 3.3.2 Choice of Open Access papers as source of full text

For these reasons, it was therefore decided to use Open Access papers, since by definition these papers are freely available without publisher authentication requirements. (Knoth and Pontika, 2016) A means of acquiring article fulltext via direct download rather than by scraping web pages hosting articles, or decoding PDFs, was also a requirement for reasons of speed and efficiency, since the coding necessary to scrape and decode article URLs in the appropriate manner could easily have taken the project outside the limited time frame available.

### 3.3.3 CORE.ac.uk dataset download

The CORE.ac.uk aggregator provides regular dumps of its dataset (c. 6 million ~100GB) documents as fulltext in JSON format for download, for the purposes of text and data mining, and the initial approach envisioned was to use this dataset- the advantages envisioned being replicability of results due to a static dataset, and ease of extension and resampling of different fulltext samples once data was acquired.

Sampling from this dataset was envisaged as being achieved by determining the set of CORE ID numbers attached to the dataset and simply generating a set of random integers within that range to extract and use as a sample.

However, upon acquisition of the data, a number of problems with this approach were apparent- firstly, the JSON interpreters used in R (the package **rjson )** to load the datasets did not successfully interpret all of the several thousand individual JSON files as valid- the reason for this was not determined and it was not entirely clear whether the fault lay with the JSON interpreter or the data files themselves- since the JSON files were very large, trying to narrow down the cause of the problem by manually inspecting the data would have been extremely time-consuming and probably impossible. In the interests of time it was decided that instead of spending significant time and effort trying to track down and repair the fault it would be easier to retrieve article full text from another source.

A secondary problem was encountered in that the article metadata supplied as a separate set of files was not easily matched with the article full text using manual inspection.

Finally, the issue of RAM management emerged, which had not been fully considered prior to beginning. Since R was chosen as the toolset for analysis, all objects to be analysed and manipulated needed to be stored in the 8GB of RAM available. It was initially envisioned that a sample could be extracted from the dataset and held in RAM, but the difficulties in parsing the JSON files meant that even extracting a list of CORE IDs from which to construct a sample proved impossible.

Given the large nature of the dataset, using a different approach, using a relational SQL database for storage of the data, would most likely be the most sensible course of action were this approach to be pursued, but given the necessity of changing either the analysis tools or the dataset, changing the dataset was judged the more sensible approach in this instance.

### 3.3.4 Attempt at fulltext acquisition from CORE.ac.uk via API

It was therefore decided to attempt a different approach:

This approach involved using the CORE.ac.uk aggregator API to directly download XML article full text using the **rcoreoa** R package- this was the approach later adopted (with modifications).

A search for a term (e.g. “diabetes”) was made through the entire CORE dataset via the CORE API, restricting the year of publication to limit the number of search terms returned (papers published afer 2015 were excluded on the basis that they will either not have picked up sufficient citations and that citations picked up soon after publication are intuitively more likely to be discovered predominantly by other means than by database searching (monitoring and browsing of recently published journals for example), using the following :

```
retrievearticleids = function(x) {

findata = data.frame(type = character(), id=character(), stringsAsFactors=FALSE)

for (i in 1:100) {

coreinit = core_search(paste(x," AND year:{1999 TO 2015} AND fullText=TRUE"), key=
"0KgyhMIxmwLQNUjfrDdFp3VnuYoSbJTG", limit=100, page = i)

retdata = data.frame(type = coreinit$data$type, id=coreinit$data$id,
stringsAsFactors=FALSE)

findata = rbind(findata, retdata)

}

return(findata)

}
```

The function is executed as follows:

```
> articleids = retrievearticleids("diabetes")
```

Returning a data frame with the following format:

| | type | id |
|---|---|---|
| 1 | article | 13976229 |
| 2 | article | 13986846 |
| 3 | article | 13986843 |
| 4 | article | 19593335 |
| 5 | article | 13993000 |
| 6 | article | 19593666 |
| 7 | article | 47195579 |
| 8 | article | 13319989 |
| 9 | article | 13970165 |
| 10 | article | 30817470 |

*Table 1: Example output for retrievearticleids() function*

Where 'type' indicates CORE object type (since all objects retrieved are articles this is always 'article') and 'id' is the CORE ID.

Once CORE IDs were returned, the remaining relevant metadata for the articles was downloaded and inspected using the following function . The initial approach envisaged (obtaining article keywords from the CORE metadata) was proved not to be feasible since the metadata returned by

the CORE fulltext did not include keywords, and although there is a citation count field this is not used (see appendix for example metadata schema):

```
#takes as input the list of CORE IDs, and returns a data frame with title, DOI, and
full text for each article.

fulltextandDOIret = function(x) {

#initialises empty data frame

fulltextart = data.frame(DOI = character(), title=character(),
fulltext=character(), stringsAsFactors=FALSE)

#loops through each line in the input table and submits each CORE ID n turn

for (i in 1:length(x$id)) {

searchres = core_articles(x$id[i], fulltext=TRUE,
key="0KgyhMIxmwLQNUjfrDdFp3VnuYoSbJTG")

fulltextadd = data.frame(DOI= ifelse(is.null(searchres$data$identifiers[2]),"NA",
searchres$data$identifiers[2]), title= ifelse(is.null(searchres$data$title),"NA",
searchres$data$title), fulltext = ifelse(is.null(searchres$data$fullText),"NA",
searchres$data$fullText), stringsAsFactors=FALSE)

fulltextart = rbind(fulltextart, fulltextadd)

}

return(fulltextart)

}
```

The above function was executed thus:

```
> articles =   fulltextandDOIret(articids)
```

Which returns a table like the following (fulltext field has been omitted for reasons of space and clarity)

| DOI | title |
|---|---|
| <NA> | Care of people with diabetes : a manual of nursing practice |
| <NA> | A novel gene and uses therefor |
| <NA> | Gene and uses therefor |
| <NA> | Book Review: Diabetes and wellbeing: Managing psychological and emotional challenges of diabetes types 1 and 2 |
| <NA> | Psychological care and structured education for people with diabetes. |
| <NA> | Spirituality:an essential aspect of holistic, individualised diabetes care and education |
| <NA> | Socioeconomic status and diabetes among urban Indigenous Australians aged 15-64 years in the DRUID study |
| <NA> | Preventing, delaying, or masking type 2 diabetes with metformin in the diabetes prevention program? |
| 9 10.1002/9780470057438.ch1 | Introduction to diabetes |
| <NA> | Indiciense and risk factors for type 2 diabetes in a general population : the Tromsø Study |

*Table 2:Example output from fulltextandDOIret() function*

This approach was successful in terms of retrieving article fulltext, although many of the returned articles were very heterogeneous and many returned fulltext fields were empty.

This set of results was then cleaned by applying the following to remove any entries with empty 'fulltext' and 'DOI' fields:

```
> articleswithdoiandfulltext = articles[!is.na(articles$DOI) & articles$fulltext
!="", ]
```

Next, this list was submitted to the Scopus API on the basis of a DOI query using the following function scopusdataretrieve():

```
scopustitleretrieve = function(x, y) {

docudata = data.frame(DOI=character(), cited.by=character(),
Author.Keywords=character(), stringsAsFactors=FALSE

)

for (i in y:length(x$title))

{

#strips out punctuation from query

titlename = x$title[i]

titlestrip  = gsub('[[:punct:] ]+',' ',titlename)

scopusquery = paste("DOI(",x$DOI[i],")"))

es = generic_elsevier_api(query= scopusquery, type ="search",
api_key="13bc04931f6e4ebb90637857c919345a", oa = TRUE, view = "COMPLETE", #change
to COMPLETE to retrieve keywords

search_type="scopus")

doitest = ifelse(is.null(es$content$`search-results`$entry[[1]]$`prism:doi`),"No
DOI Found", es$content$`search-results`$entry[[1]]$`prism:doi`)


citetest = ifelse(is.null(es$content$`search-results`$entry[[1]]$`citedby-count`),
"No Citation Count", es$content$`search-results`$entry[[1]]$`citedby-count`)

kwordtest = ifelse(is.null(es$content$`search-results`$entry[[1]]$`authkeywords`),
"N/A", es$content$`search-results`$entry[[1]]$`authkeywords`)

titletest = ifelse(is.null(es$content$`search-results`$entry[[1]]$`dc:title`), "No
Title", es$content$`search-results`$entry[[1]]$`dc:title`)

tempdata = data.frame(DOI = doitest ,

cited.by = citetest,

Author.Keywords = kwordtest,

title = titletest,

stringsAsFactors=FALSE

)

write.csv(tempdata, paste(i,".csv"))

print(paste(i," of ",length(x$title)))

docudata = rbind(docudata, tempdata)

}

return(docudata)

}
```

Executed thus:

```
> scopusarticles = scopstitleretrieve(articleswithdoiandfulltext, 1)
```

When running this collection of article DOIs into Scopus (see next section for justification of the selection of Scopus as a citation index) the number of titles retrieved proved to be very poor indeed- on the order of c. 1% - and the likelihood of obtaining a sample of usable size via this method was therefore judged to be too low to achieve a usable result, given the weekly Scopus API retrieval limit of 10,000 results, only approximately 120 results were retrieved, and of these only 30 results were found to have Author Keywords- a sample far too small to allow analysis.

### 3.3.5 Scopus search and fulltext retrieval using R 'fulltext' package

An alternative means of obtaining OA full text was therefore sought, and the R **fulltext** package was assessed as a means of obtaining the necessary text.

This package provides access to a wide variety of Open Access fulltext resources, including 'Biomed Central', Public Library of Science, 'Pubmed Central', 'eLife', 'F1000Research', 'PeerJ', 'Pensoft', 'Hindawi', 'arXiv' 'Preprints', and others via CrossRef.

Rather than obtaining article fulltext and then matching to Scopus data, the approach taken was to attempt to search Scopus and then pass the list of DOIs to the **fulltext** ft_get() function in order to retrieve the article fulltext.

```
#function takes string as search term

elsearch= function(x)

{

es = generic_elsevier_api(query=paste(x, "AND doctype(ar) AND  PUBYEAR < 2015"),
type ="search", api_key="13bc04931f6e4ebb90637857c919345a", oa = TRUE, view =
"COMPLETE",search_type="scopus")

return(es)

}
```

However, this function proved inadequate to the purpose, returning complete metadata records and only the first 25 results of each query. Therefore, an improved function was required to retrieve and reformat each page of results into a usable format with only the required data:

```
elretrieve= function(x)

{

scopquery = paste(x, "AND doctype(ar) AND  PUBYEAR < 2016")

es = generic_elsevier_api(query= scopquery, type ="search",
api_key="13bc04931f6e4ebb90637857c919345a", oa = TRUE, view = "STANDARD", #change
to COMPLETE to retrieve keywords

search_type="scopus")

resultslen = as.integer(es$content$`search-results`$`opensearch:totalResults`)

unpacked = unpack(es)

for (i in seq(26, resultslen-25, 25)) {

ez = generic_elsevier_api(query= scopquery, type ="search",
api_key="13bc04931f6e4ebb90637857c919345a", oa = TRUE, view = "STANDARD", #change
to COMPLETE to retrieve keywords

search_type="scopus", start = i)

unpackez = unpack(ez)

unpacked = rbind(unpacked, unpackez)

}

return(unpacked)

}
```

The function is called thus:

```
> scopusdata = elretrieve("diabetes")
```

The above function is dependent on the function unpack() to reformat the retrieved metadata into an appropriate data frame. The function below relies on ‘for’ loop, which would be more efficiently performed using a vector-based method- however for this small amount of data the overhead in terms of time is negligible.

```
unpack = function(x) {

docudata = data.frame(DOI=character(), cited.by=character(),
Author.Keywords=character(), stringsAsFactors=FALSE

#, type=character()

)

for (i in 1:25) {

doitest = ifelse(is.null(x$content$`search-results`$entry[[i]]$`prism:doi`),"No DOI
Found", x$content$`search-results`$entry[[i]]$`prism:doi`)

citetest = ifelse(is.null(x$content$`search-results`$entry[[i]]$`citedby-count`),
"No Citation Count", x$content$`search-results`$entry[[i]]$`citedby-count`)

kwordtest = ifelse(is.null(x$content$`search-results`$entry[[i]]$`authkeywords`),
"N/A", x$content$`search-results`$entry[[i]]$`authkeywords`)

tempdata = data.frame(DOI = doitest ,

cited.by = citetest,

Author.Keywords = kwordtest

)

docudata = rbind(docudata, tempdata)

}

return(docudata)

}
```

This returns a  data frame of the following format:

| | DOI | cited.by | Author.Keywords |
|---|---|---|---|
| 1 | 10.1371/journal.pone.0115436 | 3 | N/A |
| 2 | 10.1371/journal.pone.0116039 | 7 | N/A |
| 3 | 10.1186/2193-1801-3-562 | 0 | N/A |
| 4 | 10.1186/2193-1801-3-4 | 2 | N/A |
| 5 | 10.1186/2047-217X-3-34 | 41 | N/A |
| 6 | 10.1021/jm500902x | 10 | N/A |
| 7 | 10.1073/pnas.1409507111 | 13 | N/A |
| 8 | 10.1073/pnas.1411450111 | 20 | N/A |
| 9 | 10.1038/srep07600 | 23 | N/A |
| 10 | 10.1073/pnas.1411959111 | 5 | N/A |

*Table 3: Example output from elretrieve() function*

This is then passed to the function retrievetext() which uses the ft_get() function from the **fulltext** package to retrieve article fulltext:

```
retrievetext = function(x) {

#initialises empty data frame

ftcontain = data.frame(DOI=character(), fulltext=character(),
stringsAsFactors=FALSE)

#loops through each line and submits the DOI to ft_get(), unpacking the resulting
output using the chunks() function

for (i in 1:length(x$DOI)) {

#exception handling code to handle instances where objects cannot be retrieved or
unpacked successfully

text = tryCatch({chunks(ft_get(toString(x$DOI[i])), what="body")},

warning = function (wa) {

return(paste("WARNING:", wa))

},

error = function(er){

return(paste("ERROR:", er))

}

)

textunwrap = toString(unlist(text[1]))

temptext = data.frame(DOI= x$DOI[i], fulltext= textunwrap, stringsAsFactors=FALSE)

ftcontain = rbind(ftcontain, temptext)

}

return(ftcontain)

}
```

While this approach was successful in successfully retrieving fulltext-containing objects for approximately 50% of Scopus articles, the objects returned proved to be very heterogeneous in nature- the ft_get() function returns an object of class *ft_data* (an S3 type object) with slots for each individual publisher- and there is no single unified path to extraction of text. Therefore, the proportion of

Some text objects can be directly extracted using the fulltext chunks() function called in the function above, but this transpired only to work for a small proportion of results returned, from a relatively small number of publishers.

Other queries merely return links to PDFs which then have to be extracted using PDF extraction tools. While a function ( ft_extract() ) exists for this purpose, it became evident that the intended use of the tool was text mining of individual publishers' output, and that the data structures output by the same query for different publishers were extremely heterogeneous.

Extracting fulltext for all publishers would therefore have necessitated writing bespoke text extraction scripts for each publisher's content, which while certainly possible would most likely be very time-consuming for multiple publishers and due to the potential for this task taking up the entirety of the time available it was decided to pursue a different course of action. It seems likely that the further development of the **fulltext** package will eventually result in analyses of this nature becoming extremely simple as more publishers' metadata types are added to the library.

Nevertheless, it was hoped that enough fulltext data could be extracted using the chunks() method to allow a large enough sample suitable for analysis to be obtained via this method, but sadly when this method was trialled with a large amount of data it transpired that the vast majority of fulltext obtained was from Public Library Of Science (PLOS) , which does sadly not feature Author Keyword data, and so a usable sample was not easily obtainable via this method either. At this stage the analysis focused

The final (and ultimately successful) approach taken involved searching Scopus manually to determine publishers with a large portfolio of Open Access Journals and good Scopus coverage, before cross-referencing this with Core data to determine if a substantial number of these publishers' publiations also appeared in Core. This process will be described in more detail in the 'Data Preparation' section.

## 3.4 CITATION INDEX SELECTION

Citation index selection is the other critical part of this work. Although a large number of citation indexes exist, four main candidates were assessed for suitability. The criteria used for assessment were: Breadth of coverage, existence and access to API, API documentation, data returned.

### 3.4.1 Google Scholar

Google's widely-used academic search engine is powerful and boast excellent coverage and powerful analysis tools via its web interface. It is capable of efficiently locating highly-cited documents (Martin-Martin *et al.*, 2017), but its breadth of coverage, including a much broader variety of documents than more traditional citation indexes than Scopus or Web of Science

was rejected for a surprising reason- there is no API, and in order to retrieve citation data it would therefore be necessary to scrape data from results pages. Although tools exist to perform these duties it was felt that this would potentially add another level of complexity to the process and it was therefore decided not to pursue this direction.

### 3.4.2 Microsoft Academic

Microsoft Academic is a relatively new citation index from Microsoft (not to be confused with the now defunct but similarly-named Microsoft Academic Search), which provides API access, However the API documentation available online is once again sparse.

### 3.4.3 Web of Science Science (WoS)

The Web of Science Citation Index presents one of the two most natural initial candidate citation index. An API exists, and the R package **bibliometrix** provides tools for obtaining data and text mining techniques but publicly available documentation is scarce and given the e constraints it was judged to be better to proceed in a direction which did not necessitate potentially lengthy negotiations with Thomson Reuters representatives when it appeared from publicly available information that Scopus had the capabilities required.

### 3.4.4 Scopus

Scopus, introduced in 2004 by Elsevier, is the largest specialised bibliographic database, currently covering over 21,500 journals and 4,000 Open Access journals. (Elsevier B.V., 2016)

Scopus was eventually selected for this study due to the existence of an API, robust support for which is provided via the **rscopus** package in R. The high rate of coverage,

The comprehensiveness of the API documentation, with numerous examples of returned data which showed that access to Author Keywords would be relatively straightforward via the Scopus search API, and that (at least some) Index Keywords would be retrievable via the Abstract Search API request. Scopus was therefore selected as the citation index which provided the path of least resistance as far as the current project is concerned- an API key was requested from the Elsevier Developers Portal.

## 3.5 DATA ACQUISITION

### 3.5.1 Process Overview

**Stage A:** Search CORE.ac.uk using a keyword and publisher data for papers published between 1999 and 2015, retrieving CORE article IDs

**Stage B:** Pass CORE IDs to CORE article retrieval function, retrieving article title, fulltext and some DOIs

**Stage C:** Pass article titles and DOIs to Scopus search function, retrieving Author Keywords, DOIs, titles and Citation Counts

**Stage D:** Pass article DOIs from Scopus to Scopus Abstract Retrieval, retrieving Index Keywords

**Stage E:** Merge Scopus data from Stages D and E with article fulltext from Stage B. Assign unique identifier to each article. Merge Author and Index keywords into one master list.

**Stage F:** Stem and convert article fulltext to Tidy format. Calculate tf-idf values.

**Stage G:** Stem and convert Keywords to Tidy format.

**Stage H:** Merge Tidy Keywords and Tidy Fulltext to produce list of keywords with tf-idf values.

**Stage I:** Use tf-idf values to calculate *Qtinfo* values for each ID

**Stage J:** Merge *Qtinfo* values from Stage I with article data from Stage E to produce final table with *Qtinfo,* citation counts and article metadata.

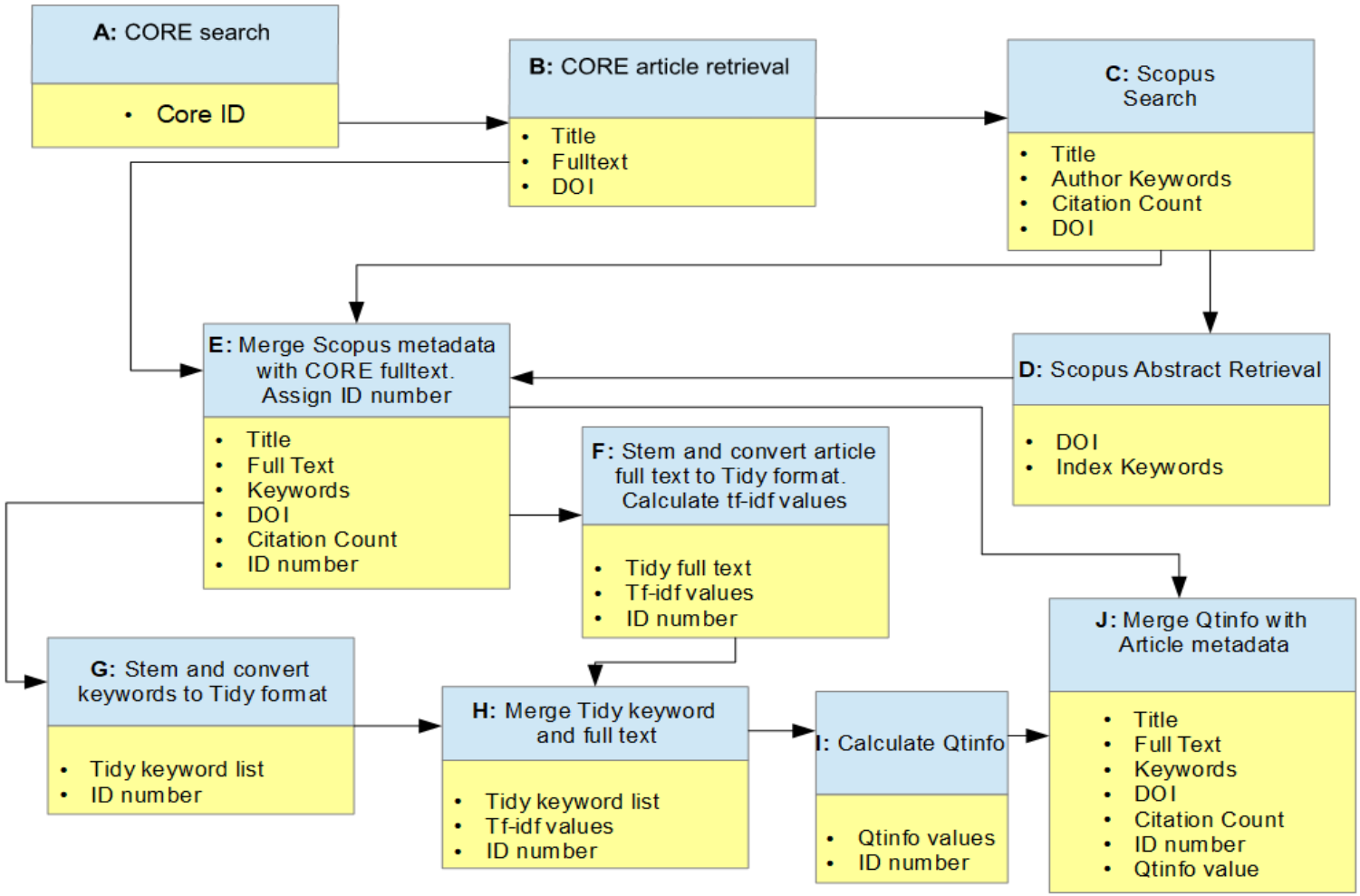


*Figure 1: Process overview flowchart*

| Stage | Process | Input Data | Output Data | Notes |
|---|---|---|---|---|
| **A** | CORE Search | Search keyword | CORE ID | |
| **B** | CORE Article Retrieval | CORE ID | Fulltext, Title, DOI | |
| **C** | Scopus Search | DOI (CORE), Title | Title, DOI (Scopus), Author Keywords, Citation Count | |
| **D** | Scopus Abstract Retrieval | DOI (Scopus) | | Only for Data Set C |
| **E** | Merge | Title, DOI (Core), DOI (Scopus), Author Keywords, Citation Count, Index Keywords, Fulltext | Title, DOI (Scopus), Keywords, Citation Count, ID, Fulltext | |
| **F** | Stem and convert fulltext, calculate tf-idf | ID Number, Fulltext | Tidy Fulltext list, ID, tf-idf values | |
| **G** | Stem and convert keywords | ID Number, keywords | Tidy Keyword list, ID | |
| **H** | Merge keywords and fulltext | Tidy Keyword, Tidy Fulltext, tf-idf values, ID | Tidy Keyword, Tidy Fulltext, tf-idf values, ID | |
| **I** | Calculate *Qtinfo* values | Tidy Keyword, Tidy Fulltext, tf-idf values, ID | ID, *Qtinfo* | |
| **J** | Final merge | Title, DOI (Scopus), Keywords, Citation Count, ID, Fulltext, *Qtinfo* | Title, DOI (Scopus), Keywords, Citation Count, ID, Fulltext, *Qtinfo* | |

*Table 4: Process Data inputs and outputs*

### 3.5.2 Stage A: CORE Search

A search was initially performed of the CORE.ac.uk aggregator for articles containing a particular word (e.g. “diabetes”, “obesity”), and under one of three particular publishers determined to have with good coverage in Scopus (BioMed Central, Springer and Nature Publishing). The purpose of this is twofold- firstly, to restrict the number of results to below the 10,000 hard limit imposed by the CORE API, and secondly to somewhat restrict the subject matter of the articles retrieved, using the function detailed below:

```
#variable x contains the string searched for, while y contains the name of the
publisher

retrievearticleids = function(x, y) {

findata = data.frame(type = character(), id=character(), stringsAsFactors=FALSE)

#loops through each page of 100 results returned by the search (up to the maximum
of 100 pages for 10,000 total results returned, then binds each result to the end
of the data frame

for (i in 1:100) {

coreinit = core_search(paste(x,"AND publisher:",y,"AND year:{1997 TO 2015} AND
fullText=TRUE"), key= "0KgyhMIxmwLQNUjfrDdFp3VnuYoSbJTG", limit=100, page = i)

retdata = data.frame(type = coreinit$data$type, id=coreinit$data$id,
stringsAsFactors=FALSE)

findata = rbind(findata, retdata)

}

#finally, returns the entire data frame

return(findata)

}
```

```
> articleidsbiomed = retrievearticleids(“diabetes”, “BioMed Central”)

> articleidsspringer = retrievearticleids(“diabetes”, “Springer”)

> articleidsnature = retrievearticleids(“diabetes”, “Nature Publishing Group”)
```

These function calls return a lists of CORE IDs (the CORE.ac.uk object identifier):

Which were then combined into one data frame using the rbind() function:

```
> articleids = rbind(articleidsbiomed, articleidsspringer, articleidsnature)
```

### 3.5.3 Stage B: CORE Article retrieval

These data were then passed to the following function in order to retrieve DOIs and article full text:

```
#takes as input an object consisting of CORE id numbers

fulltextandDOIret = function(x) {

fulltextart = data.frame(DOI = character(), title=character(),
fulltext=character(), stringsAsFactors=FALSE)

#loops through each CORE id and submits it as a query to core_articles()

for (i in 1:length(x$id)) {

searchres = core_articles(x$id[i], fulltext=TRUE,
key="0KgyhMIxmwLQNUjfrDdFp3VnuYoSbJTG")

#adds the query result to the data frame, testing for null values returned

fulltextadd = data.frame(

DOI= ifelse(is.null(searchres$data$identifiers[2]),"NA",
searchres$data$identifiers[2]),

title= ifelse(is.null(searchres$data$title),"NA", searchres$data$title),

fulltext = ifelse(is.null(searchres$data$fullText),"NA", searchres$data$fullText),
stringsAsFactors=FALSE)

#binds the results data frame to the previous iteration of the loop

fulltextart = rbind(fulltextart, fulltextadd)

}

#returns the final data frame

return(fulltextart)

}
```

```
> articlefulltext = fulltextandDOIret(articleids)
```

### 3.5.4 Stage C: Scopus Search

The output of this function is a data frame with DOIs, article title, and article full text. The next stage involves passing the DOI and article title to the following function, which searches Scopus for the DOI of an article. In cases where a DOI is not present (ie the DOI value is NA), the title (stripped of punctuation), is submitted as a search query instead. Author Keywords, EID (Scopus unique identifier), DOI, title and citation count for each article are retrieved and appended to a data frame, as well as writing each record to a CSV file in case the rather lengthy process of record retrieval is interrupted (as happened on several occasions due to power outages and network maintenance).

Submitting full titles as search queries to Scopus carries with it a risk of retrieving the wrong

document if the top search matching the search string is not the correct document (ie the original document is not present in Scopus.) A set of 100 records was manually examined for this eventuality and no discrepant records were found. Moreover, the process is designed to account for this possibility- because full text and Scopus records are later merged by title, any erroneously retrieved records will be discarded at this point.

```
scopustitleretrieve = function(x) {

docudata = data.frame(DOI=character(), cited.by=character(),
Author.Keywords=character(), stringsAsFactors=FALSE

)

for (i in 1:length(x$title))

{

titlename = x$title[i]

#strips out punctuation from query

titlestrip  = gsub('[[:punct:] ]+',' ',titlename)

scopusquery = ifelse(is.na(x$DOI[i]), paste("title(",titlestrip,")"),

paste("DOI(",x$DOI[i],")"))

es = generic_elsevier_api(query= scopusquery, type ="search",
api_key="13bc04931f6e4ebb90637857c919345a", oa = TRUE, view = "COMPLETE", #change
to COMPLETE to retrieve keywords

search_type="scopus")

doitest = ifelse(is.null(es$content$`search-results`$entry[[1]]$`prism:doi`),"No
DOI Found", es$content$`search-results`$entry[[1]]$`prism:doi`)

eidtest = ifelse(is.null(es$content$`search-results`$entry[[1]]$eid), "No EID
Found", es$content$`search-results`$entry[[1]]$eid)

citetest = ifelse(is.null(es$content$`search-results`$entry[[1]]$`citedby-count`),
"No Citation Count", es$content$`search-results`$entry[[1]]$`citedby-count`)

kwordtest = ifelse(is.null(es$content$`search-results`$entry[[1]]$`authkeywords`),
"N/A", es$content$`search-results`$entry[[1]]$`authkeywords`)

titletest = ifelse(is.null(es$content$`search-results`$entry[[1]]$`dc:title`), "No
Title", es$content$`search-results`$entry[[1]]$`dc:title`)

tempdata = data.frame(DOI = doitest ,

cited.by = citetest,

eid = eidtest,

Author.Keywords = kwordtest,

title = titletest,

stringsAsFactors=FALSE

)

write.csv(tempdata, paste(i,".csv"))

print(paste(i," of ",length(x$title)))

docudata = rbind(docudata, tempdata)

}

return(docudata)

}
```

This retrieves a data frame containing document DOIs, citation counts, Author Keywords and title:

| DOI | cited.by | Author.Keywords | title |
|---|---|---|---|
| 10.1007/s00125-008-1150-5 | 55 | Diabetes \| Epidemiology \| Incidence \| Life expectancy \| Lifetime risk \| Mortality \| Population \| Prevalence | Lifetime risk and projected population prevalence of diabetes |
| 10.1007/s00125-008-0982-3 | 67 | Association study \| BMI \| GWA \| Replication \| Single nucleotide polymorphism \| SNP \| Susceptibility gene \| Type 2 diabetes | Genetic analysis of recently identified type 2 diabetes loci in 1,638 unselected patients with type 2 diabetes and 1,858 control participants from a Norwegian population-based cohort (the HUNT study) |
| 10.1007/s10654-010-9540-7 | 17 | High metabolic score \| Low metabolic score \| Metabolic syndrome \| Non-metabolic risk factors \| Receiver operating characteristics \| Type 2 diabetes | Risk factors for type 2 diabetes in groups stratified according to metabolic syndrome: A 10-year follow-up of the Tromsø Study |
| 10.1007/s00125-009-1557-7 | 4 | Genetics \| LARS2 \| Mitochondria \| SNP \| Type 2 diabetes | Genetic association analysis of LARS2 with type 2 diabetes |
| 10.1007/s11695-013-0907-1 | 15 | Body mass Index <35 kg/m 2 \| Metabolic surgery \| Remission of diabetes \| Type 2 diabetes | Metabolic surgery for type 2 diabetes with BMI <35 kg/m2: An endocrinologist's perspective |
| 10.1007/s00592-009-0138-z | 40 | Brazil \| Diabetes mellitus \| Epidemiology \| Glycaemic control \| HbA 1c | Prevalence and correlates of inadequate glycaemic control: Results from a nationwide survey in 6,671 adults with diabetes in Brazil |

*Table 5:Example output from scopustitleretrieve() function*

### 3.5.5 Stage D: Index Keyword retrieval (Data Set C only)

For Data Set C, which sadly was not analysed due to hitting RAM limitations and a lack of time available to adapt the data preparation and analysis techniques sufficiently, the DOI was then passed to the following function, which uses the Scopus Abstract Retrieval function to retrieve the object representing the abstract, before reformatting the Index Keywords into a string formatted identically to the Author Keyword string, in order to ease subsequent manipulation.

```
#variable x is the name of the input object, y is the start point in the data
frame- because this operation takes many hours to retrieve all data from Scopus,
provision for restarting the operation if interrupted was necessary.

indexkwretrieve = function(x, y) {

#initialises an empty data frame for records to be bound to

docudata = data.frame(

DOI= character(),

Index.Keywords=character(),

stringsAsFactors=FALSE

)

#calls the Scopus API

for (i in y:length(x$DOI))

{

doist = x$DOI[i]

es = abstract_retrieval(doist, identifier="doi",
api_key="13bc04931f6e4ebb90637857c919345a")

idx = es$content$'abstracts-retrieval-response'$idxterms$mainterm

keyop = paste(lapply(idx, "[[", 3), collapse=" | ")

tempdata = data.frame(

DOI = doist,

Index.Keywords = keyop,

stringsAsFactors= FALSE

)

#returns current statuds and writes output to a series of csv files for process
resumption in the case of interruption

print(paste(i," of ",length(x$title)))

docudata = rbind(docudata, tempdata)

write.csv(tempdata, paste(i,"_indexkeyword.csv"))

}

return(docudata)

}
```

| | X | DOI | Index.Keywords |
|---|---|---|---|
| 47 | 1 | 10.1007/s40271-014-0068-x | Adult \| Aged \| Aged, 80 and over \| Decision Making \| Diabetes Mellitus, Type 2 \| Europe \| Female \| Focus Groups \| Health Knowledge, Attitudes, Practice \| Humans \| Hypoglycemic Agents \| Insulin \| Male \| Middle Aged \| Patient Education as Topic \| Patient Participation \| Risk Factors \| United States |
| 48 | 1 | 10.1007/s00125-008-0961-8 | NA |
| 49 | 1 | 10.1007/s12160-013-9498-2 | NA |
| 50 | 1 | 10.1007/s00125-009-1588-0 | NA |
| 51 | 1 | 10.1007/s00127-014-0974-1 | Adult \| Aged \| Aged, 80 and over \| Alcohols \| Anxiety Disorders \| Arthritis \| Cardiovascular Diseases \| Comorbidity \| Diabetes Mellitus \| Diagnostic and Statistical Manual of Mental Disorders \| Female \| Gastrointestinal Diseases \| Health Surveys \| Humans \| Male \| Middle Aged \| Personality Disorders \| Physical Examination \| Self Report \| Substance-Related Disorders \| United States \| Young Adult |
| 52 | 1 | 10.1007/s10552-014-0433-z | Australia \| Biomarkers \| Blood Glucose \| Body Mass Index \| Breast Neoplasms \| C-Reactive Protein \| Cohort Studies \| Comorbidity \| Cross-Sectional Studies \| Female \| Humans \| Inflammation \| Insulin Resistance \| Middle Aged \| Models, Statistical \| Postmenopause \| Sedentary Lifestyle \| Television \| Waist Circumference |
| 53 | 1 | 10.1007/s10552-009-9407-y | NA |
| 54 | 1 | 10.1007/s00125-014-3216-x | Adult \| Aged \| Biomarkers \| Blood Glucose \| Diabetes Mellitus, Type 2 \| Epidemiologic Studies \| Female \| Humans \| Hypoglycemic Agents \| Male \| Middle Aged \| Pregnancy \| Prospective Studies |
| 55 | 1 | 10.1007/s00192-009-0888-8 | NA |

*Table 6: Example output from indexkwretrieve() function*

### 3.5.6 Stage E: Merging Scopus information with CORE fulltext

#### *3.5.6.1 Author Keywords and Citation Counts*

Once the citation count and Author Keywords are retrieved, this information is merged with the fulltext information on the title value. Because discrepancies in character encoding and punctuation exist sometimes exist between title information retrieved from CORE and the same information retrieved from Scopus, both title columns are converted to lowercase, stripped of punctuation and a substring consisting of the first 50 characters of the title using the following function:

```
> striptitle = function(x) {
output = tolower(substr(gsub('[[:punct:] ]+',' ',x), 1, 50))
}
```

```
> fulltext$titleshort = striptitle(fulltext$title)
```

```
> keywords$titleshort = striptitle(keywords$title)
```

```
> fulltext = merge(fulltext, keywords, by="titleshort")
```

Before any duplicate rows were removed using:

```
>   fulltext = fulltext[!duplicated(fulltext$title), ]
```

It should be noted that upon later examination of the data, one instance of two papers with identical titles published by the same authors was discovered to have been omitted from the analysis at this point- a failure which was corrected in the preparation of the second data set by manually examination of the data, before using the following command to remove duplicates:

```
> fulltext = unique(fulltext)
```

Since DOI information is only present in a minority of documents (almost all records retrieved from Scopus, in contrast, have DOIs).

#### *3.5.6.2 Index Keywords*

For Data Set C, the Index Keyword data was then merged with the Author Keyword and citation data (object name fulltext) using an outer join on DOI (since both the Author Key, and both author keywords and index keywords are concatenated into one string for later processing.

```
> fulltext = merge(fulltext, indexkw, by="DOI")
```

#### *3.5.6.3 Assignment of Index value*

Next, an index value is generated for each row in order to reduce memory use in the tidy text stage to follow (since by default the unnest_tokens() function attaches all columns of the parent table to the output, resulting in a vastly inflated file size if all column values remain attached- estimated total size of this data would be on the order of 60GB!)

```
> fulltext$id = seq.int(nrow(fulltextkeywords))
```

### 3.5.7 Stage F: Fulltext preparation and calculation of tf-idf values (Data Set A)

Next, the fulltext is stemmed using the **tm** package's stemDocument() function, and converted into Tidy Text format using the **tidytext** package. This text format comprises one token per row, making analysis and manipulation using packages such as **dplyr** easy.

```
tidyfulltext = function(x) {

y = data.frame(id= x$id, stemmed = stemDocument(x$fulltext),
stringsAsFactors=FALSE)

g = unnest_tokens(y, stemmedtext, stemmed)

return(g)

}
```

Once the text is in tidy format, we count the instances of each token and then calculate the tf-idf weightings using the **tidytext** bind_tf_idf() function:

```
calculatetf_idf = function(x) {

bicop = ungroup(count(x, DOI, stemmedtext, sort=TRUE))

ttf = bind_tf_idf(bicop, stemmedtext, DOI, n)

besttg = arrange(ttf, desc(tf_idf))

return(besttg)

}
```

### 3.5.8 Stage F: Fulltext preparation and tf-idf value calculation (Data Set B)

For Data Set B, tf-idf values for N-grams with N of 2, 3, 4, and 5 were also calculated separately using the following function, and then separately merged with the keyword list using the rbind() function:

```
tidyfulltext = function(x, y) {

#prepares initial data frame

y = data.frame(id= x$id, stemmed = stemDocument(x$fulltext),
stringsAsFactors=FALSE)

#unnest_tokens takes a string of space-separated tokens of length y and returns a
tibble

g = unnest_tokens(y, keyword, stemmed, token="ngrams", n=y)

#calculates the raw frequency of each token per document

bicop = ungroup(count(g, id, keyword, sort=TRUE))


ttf = bind_tf_idf(bicop, keyword, id, n)

besttg = arrange(ttf, desc(tf_idf))

return(besttg)}
```

Thus (for object fulltext containing all of the article fulltext data)

```
> tokens = tidyfulltext(fulltext, 1)

> bigrams = tidyfulltext(fulltext, 2)

> trigrams = tidyfulltext(fulltext, 3)

> quadrams = tidyfulltext(fulltext, 4)

> quingrams= tidyfulltext(fulltext, 5)

> allkeywords = rbind(tokens, bigrams, trigrams, quadgrams,
quingrams)
```

RAM limits meant that creating the objects representing tidy fulltext N-grams where N > 5 was not possible on the hardware used for the analysis (each N-gram fulltext object had to be dropped from memory before the next one could be generated)- however, the number of these omitted keywords was very small (19) : they were calculated using the following method and listed in the table below:

Token numbers were calculated and bound to the keyword data frame using the following formula:

```
> keywords$tokennumber = sapply(gregexpr("[[:graph:]]+", keywords$keyword),
function(x) sum(x > 0))
```

| id | Keyword | Number of Tokens |
|---|---|---|
| 195 | insulin like growth factor 2 gene | 6 |
| 307 | heart failur with preserv eject fraction | 6 |
| 307 | heart failur with reduc eject fraction | 6 |
| 350 | 3d invers recoveri gradient echo puls sequenc | 7 |
| 456 | stroke depress minor risk factor psychosoci dispar | 7 |
| 474 | development origin of health and diseas | 6 |
| 479 | whole genom and whole exom sequenc | 6 |
| 495 | fulker associ model and lipid trait | 6 |
| 522 | geograph variat in medicar cost per episod | 7 |
| 576 | chronic allograft nephropathi /tubular atrophi with interstiti fibrosi (can/ifta) | 9 |
| 589 | tibetan wild barley (hordeum spontaneum l.) | 6 |
| 650 | carbohydr diet fatti acid f-fdg myocardium pet 18 | 8 |
| 869 | men who have sex with men | 6 |
| 919 | yogarandom control trialphys functionpsychosoci functionqu of life | 7 |
| 948 | hydrocephalus after spontan aneurysm subarachnoid hemorrhag | 6 |
| 1084 | singl photon emiss comput tomographi myocardi perfus imag | 8 |
| 1162 | acut physiolog and chronic health evalu (apache) ii score | 9 |
| 1167 | 1 % glucos acet ringer solut | 6 |
| 1202 | two-dimension fluoresc differ in gel electrophoresi | 6 |

*Table 7: Keywords with token number >5 omitted from analysis due to RAM restrictions*

### 3.5.9 Stage G: Stem and convert keywords to Tidy format (Data Set A)

The keyword list was then prepared in a similar fashion, stemming and unnesting into a tidy list of keyword tokens for model A:

```
keywordtokenise = function(x) {

stemmedsub= data.frame(index= x$index, stemmed.keywords=stemDocument(gsub(" \\| ",
" ", x$Author.Keywords)), stringsAsFactors=FALSE)

stemtokens = unnest_tokens(stemmedsub, tokens, stemmed.keywords, to_lower = TRUE,
drop=TRUE

)

return(stemtokens)

}
```

| | id | tokens |
|---|---|---|
| 1 | 1 | black |
| 1.1 | 1 | blood |
| 1.2 | 1 | coronari |
| 1.3 | 1 | arteri |
| 1.4 | 1 | dual |
| 1.5 | 1 | invers |
| 1.6 | 1 | inversion |
| 1.7 | 1 | recoveri |
| 1.8 | 1 | phasesensit |
| 1.9 | 1 | vessel |
| 1.1 | 1 | wall |
| 2 | 2 | |
| 3 | 3 | |
| 4 | 4 | |
| 5 | 5 | aneurysm |
| 5.1 | 5 | antithrombot |
| 5.2 | 5 | treatment |
| 5.3 | 5 | carotid |
| 5.4 | 5 | occlus |

*Table 8: Example output from keywordtokenise() function*

The keyword list was then merged with each tidied keyword list by id number and keyword, thus producing a data frame with the tf_idf weightings for each keyword.

```
> keywordtokens = merge(keywords, tokens, by=c("keyword","id")
```

| id | keyword | n | tf | idf | tf_idf |
|---|---|---|---|---|---|
| 10 | classif | 13 | 0.003947768 | 1.1020145427 | 0.0043504977 |
| 10 | condens | 11 | 0.0033404191 | 4.1556225653 | 0.0138815209 |
| 10 | edit | 11 | 0.0033404191 | 2.090871739 | 0.0069843878 |
| 10 | forget | 3 | 0.0009110234 | 3.1905416693 | 0.0029066581 |
| 10 | learn | 38 | 0.0115396295 | 1.4706135634 | 0.0169703357 |
| 100 | fhbl | 58 | 0.0084070155 | 8.9431143081 | 0.0751849007 |
| 1001 | divers | 75 | 0.0029972425 | 1.3136243917 | 0.0039372509 |
| 1001 | dysbiosi | 29 | 0.0011589338 | 6.4582076583 | 0.007484635 |
| 1001 | probiot | 100 | 0.0039963234 | 3.3446923491 | 0.0133664722 |
| 1002 | benfotiamin | 104 | 0.0118897908 | 8.2499671275 | 0.0980903831 |

*Table 9: Example of token data with bound tf-idf values. 'n' is the number of occurrences of the token in its document*

```
> keywords = merge(keywords, tokens, by=c("keyword","id")
```

The keyword list was then merged with each tidied keyword list by id number and keyword, thus producing a data frame with the tf_idf weightings for each keyword.

### 3.5.10 Stage G: Stem and convert keywords to Tidy format (Data Set B)

For Data Set B, a different approach was taken, taking each keyword phrase as an individual token and using the following function to decompose the author keywords field into a data frame with one author keyword per line:

```
keywordsplit = function(x) {

#prepares an intermediary data frame containing only the id number

y = data.frame(id= character(), keyword= character(), stringsAsFactors=FALSE)

interframe = data.frame(id= character(), keyword= character(),
stringsAsFactors=FALSE)

for (i in 1:length(x$id)) {

kspt = strsplit(as.character(stemDocument(tolower(x$keyword[i]))), split =" \\| ")

intermed = data.frame(id= x$id[i], keyword=
ifelse(length(kspt[[1]])==0,"NOKEYWORDS", kspt), stringsAsFactors=FALSE)

colnames(intermed)[2] <- "keyword"

interframe = rbind(interframe, intermed)

}

return(interframe)

}
```

| id | keyword |
|---|---|
| 1 | black-blood |
| 1 | coronari arteri |
| 1 | dual invers |
| 1 | invers recoveri |
| 1 | phasesensit |
| 1 | vessel wall |
| 2 | n/a |
| 3 | n/a |
| 4 | n/a |
| 5 | aneurysm |
| 5 | antithrombot treatment |
| 5 | carotid occlus |
| 5 | carotid stenosi |
| 5 | dissect |
| 5 | intern carotid arteri |
| 5 | stroke prevent |
| 5 | thrombolysi |
| 5 | vascular risk factor |
| 6 | classif use order statist (os) |
| 6 | moment of os |

*Table 10: Example Tidy Keyword output from keywordsplit() function*

### 3.5.11 Stage I: Calculation of *Qtinfo* values

To produce the final *Qtinfo* value for each article, we simply sum the tf-idf weightings for all keywords for each article and then take the logarithm.

This procedure is identical for $Qtinfo_{token}$ and $Qtinfo_{kw}$ - the difference between the two lies in the procedure for stage G.

```
qtinfocalc = function(x) {
#sums the Qtinfo values for each id number
aggregatesum = aggregate(list( tf_idf = x$tf_idf), list(id = x$id), sum)
#takes the base 10 logarithm of the sum Qtinfo value
qtinfotable = data.frame(index = aggregatesum$index, qtInfo =
log10(aggregatesum$tf_idf))
return(qtinfotable)
}
```

### 3.5.12 Stage J: Production of final table.

We then remerge this with the citation count data and end up with a table of paper id numbers, citation counts and Qtinfo parameter.

```
>  finaldata = merge(fulltext, qtinfovals, by="id")
```

# 4 RESULTS

## 4.1 SUMMARY OF DATA SETS GATHERED

| Data Set | CORE query keyword | Publisher fields submitted to CORE | Records retrieved from CORE | Records submitted to Scopus | Records retrieved with Author Keywords | Records Retrieved with Index Keywords | Sample size selected | Sample size after matching to CORE and duplicate removal | |
|---|---|---|---|---|---|---|---|---|---|
| A | diabetes | 'BioMed Central', 'Springer', 'Nature Publishing Group' | 17451 | 17451 | 3594 | 0 | 3594 | 3588 | |
| B | diabetes | 'BioMed Central', 'Springer', 'Nature Publishing Group' | 17451 | 17451 | 3594 | 0 | 1300 | 1291 | |
| C | diabetes | 'BioMed Central', 'Springer', 'Nature Publishing Group' | 17451 | 17451 | 3594 | 1987 | N/A | N/A | Not analysed |

*Table 11: Summary of Data Sets*

## 4.2 DATA SET A

### 4.2.1 Data Set A: $Qtinfo_{token}$ analysis. Results summary and distributions.

Initially, the documents were retrieved using the search term 'diabetes' from the CORE aggregator for the publishers Nature Publishing Group, Springer and BioMed Central between the years 1999 and 2015. This retrieved a total of 17,451 records, of which, when passed to Scopus, 3,588 were found to be unique articles with attached Author Keywords.

Keywords were decomposed into individual tokens and the $Qtinfo_{token}$ parameter calculated on this basis.

Initially, the distributions of dependent and independent variables were examined visually to determine their distributions.

| Parameter | Min. | 1st Quartile | Median | Mean | 3rd Quartile | Max. | Variance |
|---|---|---|---|---|---|---|---|
| $Qtinfo_{token}$ | -3.843 | -1.339 | -1.137 | -1.156 | -0.9375 | -0.1147 | 0.1046275 |
| cited.by | 0 | 5 | 9 | 19.72 | 20 | 1084 | 1851.939 |
| token.count | 2 | 7 | 10 | 10.3 | 12 | 59 | 18.4807 |

*Table 12: Basic statistical properties of independent and dependent variables*

*Qtinfo*$_{token}$ was determined visually to have a distribution approximating tolerably well to a normal distribution:

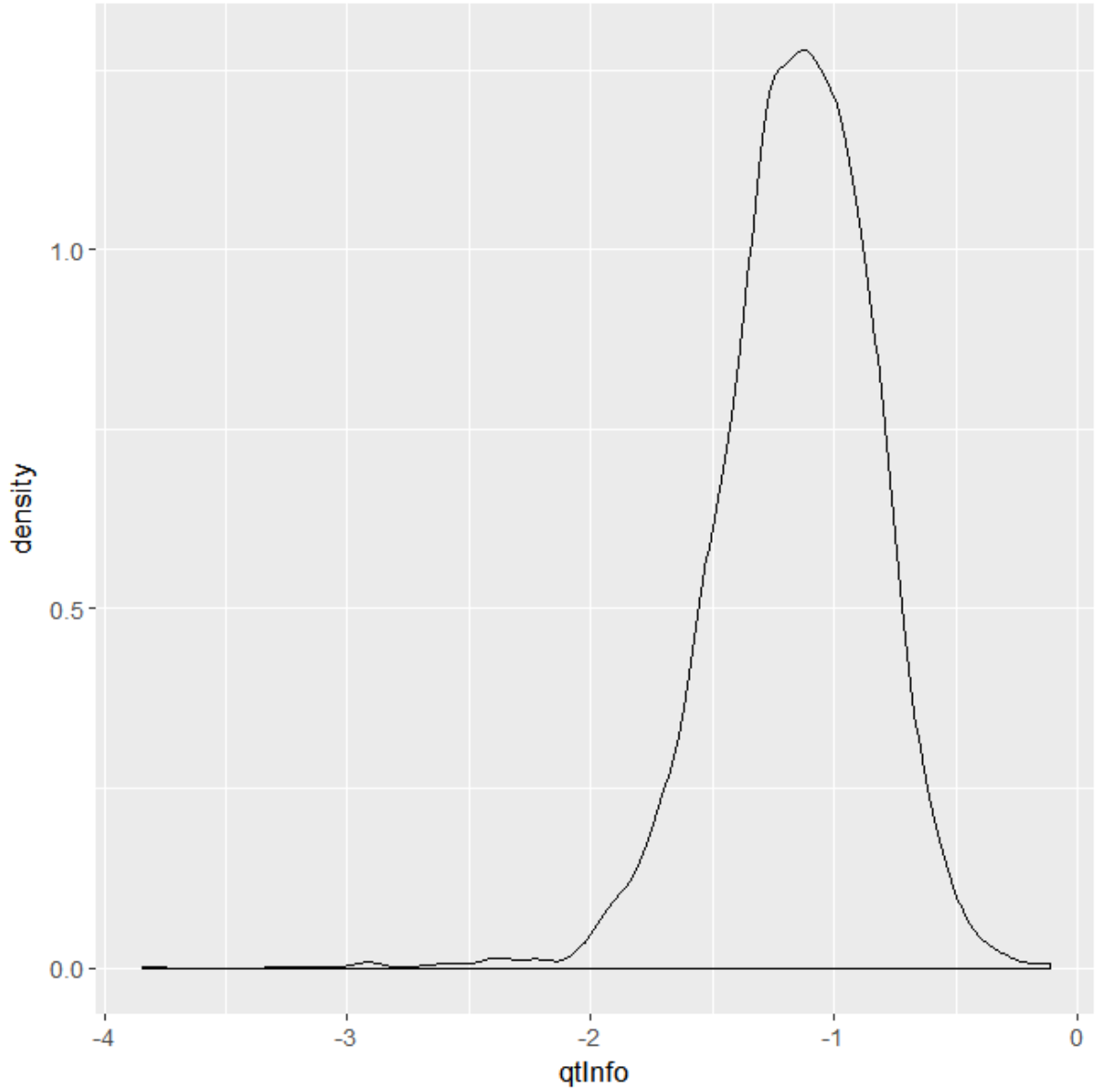


*Figure 2: Density Plot for Qtinfo*$_{token}$ *Data Set A*

Whereas the distribution of citation rates is highly right-skewed (it is notable that the mean is greater than the 3$^{rd}$ quartile value). The issue of how to appropriately model this distribution is a key

factor in the interpretation of these results.

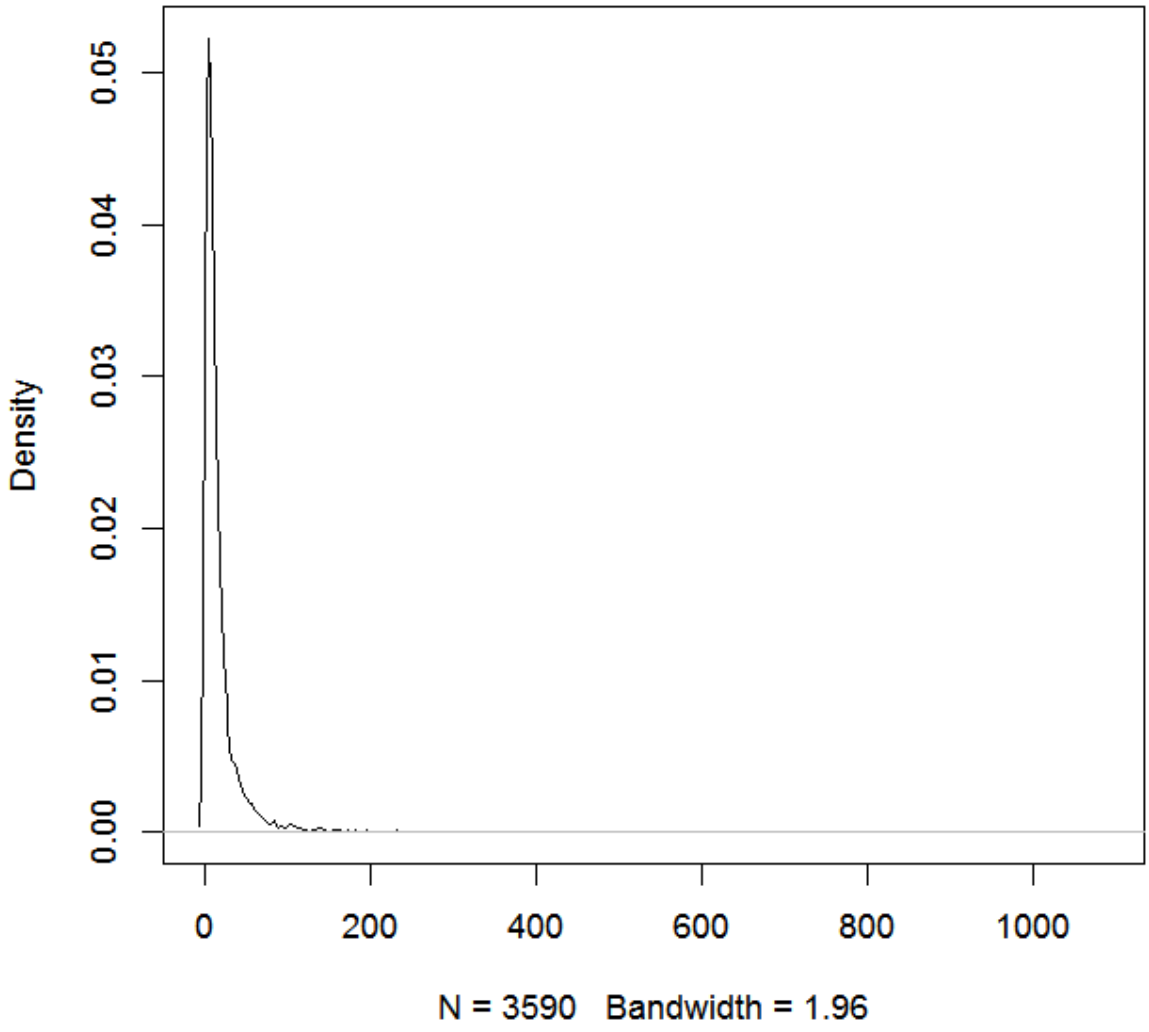


*Figure 3: Density plot for Citation Rates, Data Set A*

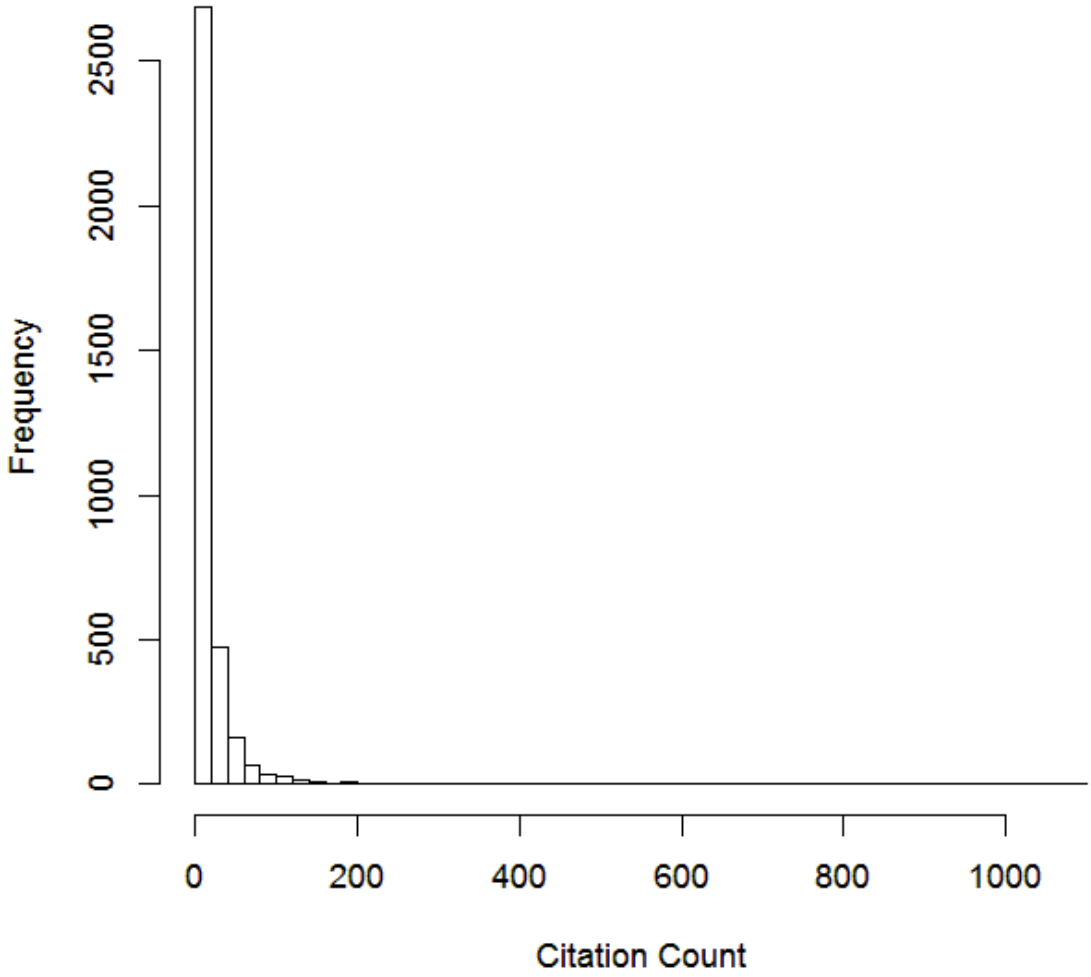


*Figure 4: Histogram for Citation Rates, Data Set A*

### *4.2.2* Model A1: $Qtinfo_{token}$, Log scaled *cited.by*

The initial approach attempted involved applying a logarithmic scaling to the citation count data, along with a small constant (0.0001) to avoid invalid values for log values of 0 counts.

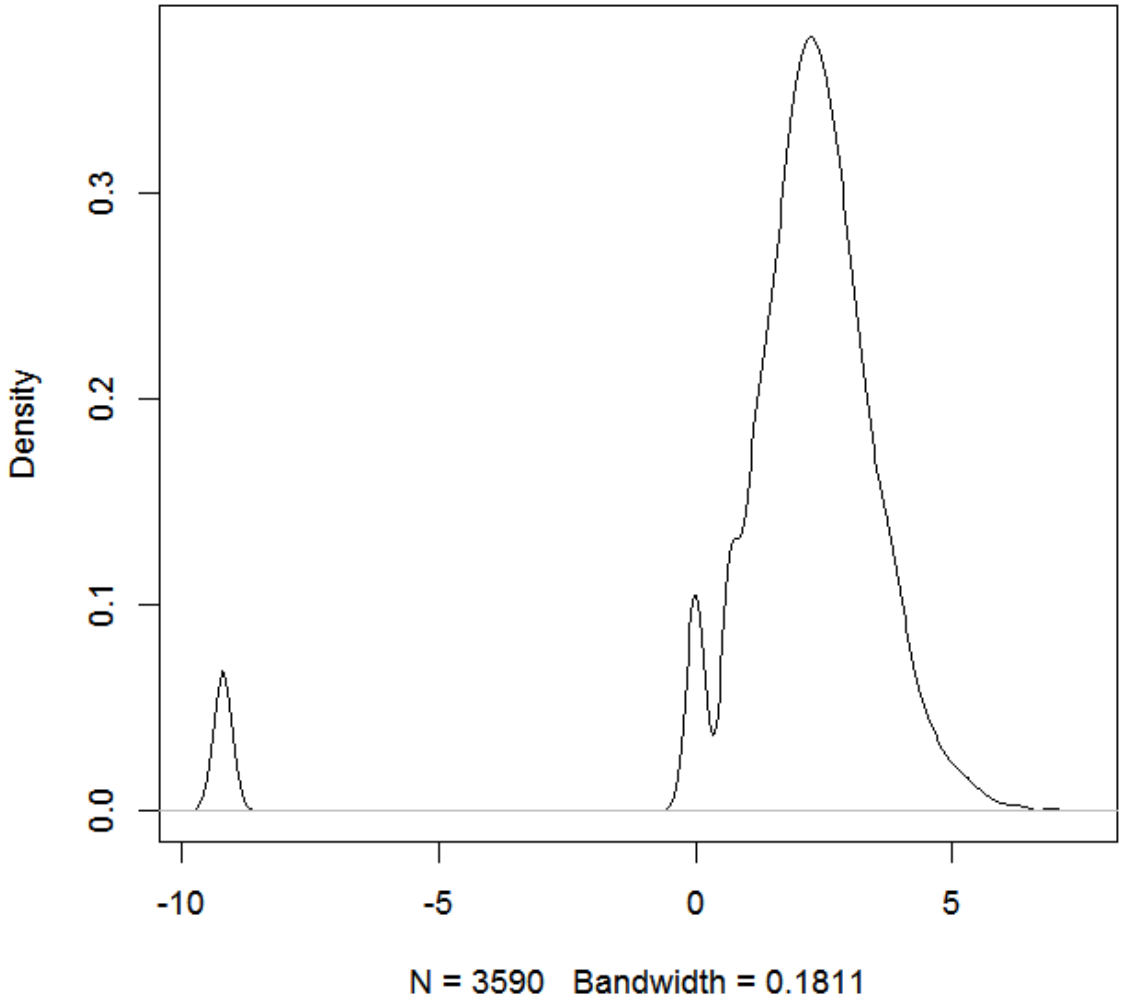


*Figure 5: Density Plot for Log-Scaled Citation Rates (Model A1)*

Once a log scaling for the citation rates was determined to be appropriate, the regression was run, with the following results:

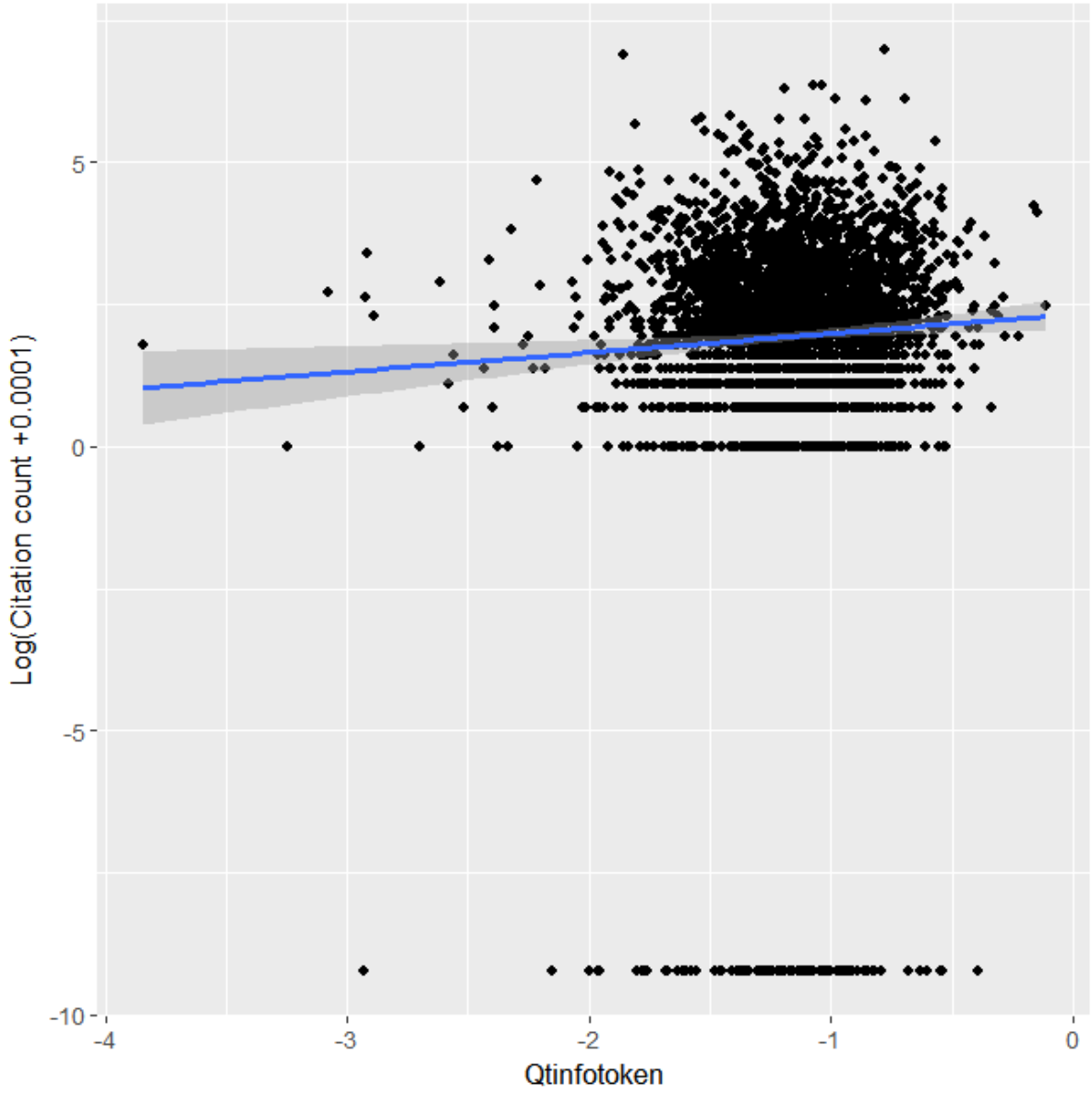


*Figure 6: Log-Scaled Citation Counts against $Qtinfo_{token}$ (Model A1)*

```
Call:
lm(formula = log(cited.by + 1e-04) ~ Qtinfo, data = theactualthing)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4194  -0.4385   0.2861   1.0265   5.1667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3414     0.1416  16.530  < 2e-16 ***
Qtinfo        0.3359     0.1180   2.847  0.00444 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.286 on 3588 degrees of freedom
Multiple R-squared:  0.002254,  Adjusted R-squared:  0.001976
F-statistic: 8.105 on 1 and 3588 DF,  p-value: 0.004438
```

*Model A1: Output*

A residuals plot was produced and examined (smoothing line added to indicate mean variances for each value of $Qtinfo_{token}$ since the large number of data points otherwise makes assessment somewhat difficult):

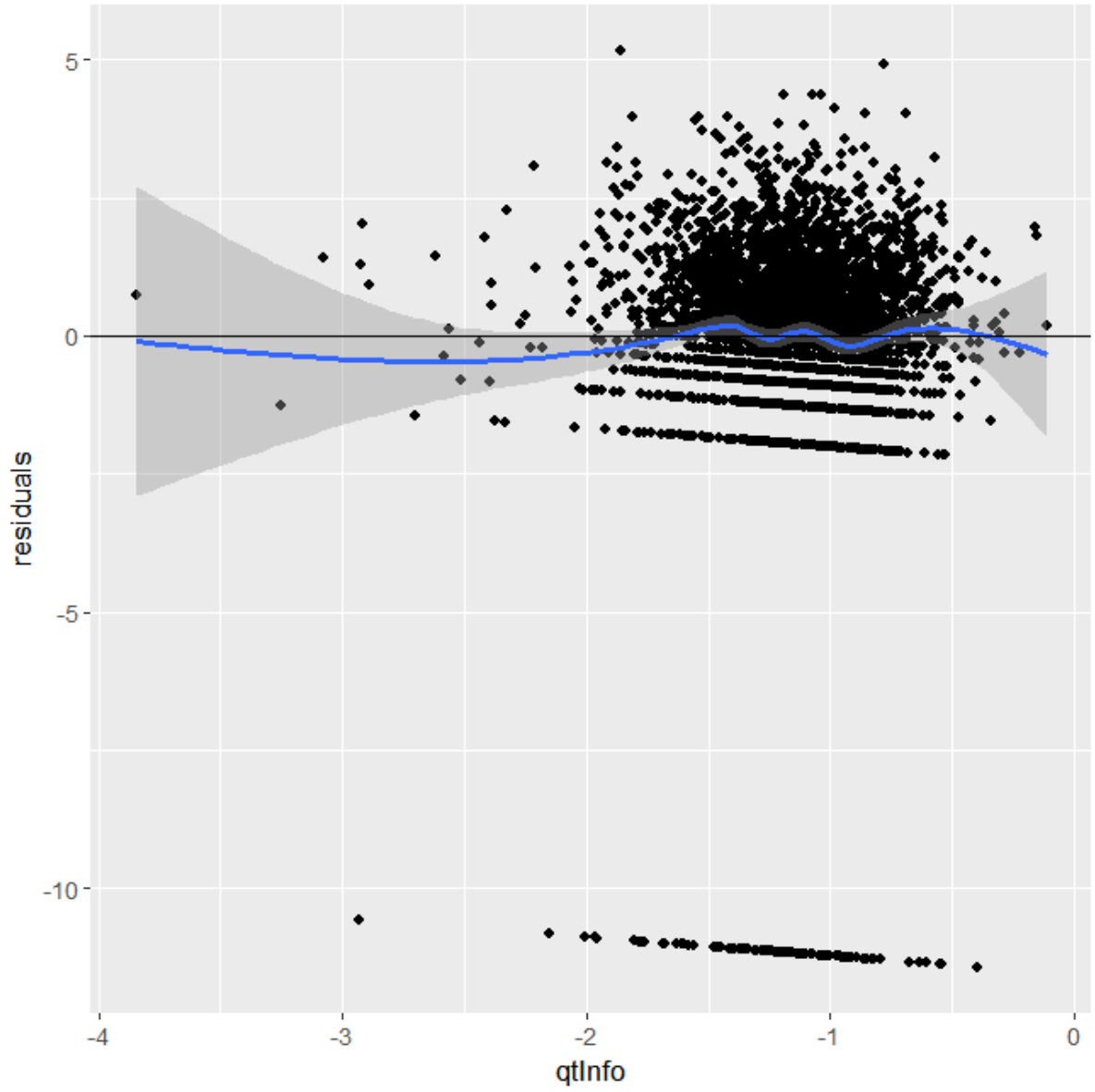


*Figure 7: Residuals plot, Model A1*

The 'banding' or stratification of the residuals is due to the fact that the data is count data and contains significant amounts of zero and one-count data (1,768 of 3,588 papers have a citation rate < 10). The distribution of residuals does not overall deviate systematically from the distribution about the 0 line.

Nonetheless, the distribution of residuals is far from random- given the small proportion of variance (ie the weakness of the overall effect) that the model predicts it was therefore judged worthwhile to

attempt other tests and investigation of the data before accepting the small p value as indicative of a genuine effect.

As well as this, the top most highly-cited papers were visually examined, and the highest and lowest $Qtinfo_{token}$ scores were examined, as well as atypical papers with high citation counts and low $Qtinfo_{token}$ scores.

| id | $Qtinfo_{token}$ | token.count | Author Keywords | Title | cited.by |
|---|---|---|---|---|---|
| 1437 | -0.7812114276 | 8 | Clinical risk factors \| Fracture probability \| Frax ⅈⅴ \| Osteoporotic fracture | FRAX ⅈⅴ and the assessment of fracture probability in men and women from the UK | 1084 |
| 1367 | -1.8599772004 | 7 | Classification of diseases \| disease \| epidemiology \| morbidity \| register | External review and validation of the Swedish national inpatient register | 976 |
| 3114 | -1.0405528818 | 7 | Behavior change interventions \| Behavior change techniques \| Taxonomy | The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions | 573 |
| 2092 | -1.0726032922 | 10 | Apoptotic body \| Autoimmune disease \| Biomarker \| Cancer \| Exosome \| Microparticle \| Microvesicle \| Platelet | Membrane vesicles, current state-of-the-art: Emerging role of extracellular vesicles | 570 |
| 2424 | -1.1916051596 | 8 | Dispersibility \| Microencapsulation \| Particle formation \| Respiratory drug delivery \| Stabilization | Pharmaceutical particle engineering via spray drying | 540 |
| 3149 | -0.6920402963 | 16 | Endothelial glycocalyx \| Endothelial surface layer \| Heparan sulfate \| Hyaluronic acid \| Optical imaging \| Two-photon microscopy \| Vascular disease | The endothelial glycocalyx: Composition, functions, and visualization | 460 |
| 1267 | -0.9788808773 | 15 | Bone mineral density \| Diagnosis of osteoporosis \| Fracture risk assessment \| FRAX \| Health economics \| Treatment of osteoporosis | European guidance for the diagnosis and management of osteoporosis in postmenopausal women | 451 |
| 2764 | -0.8547445926 | 12 | Aspart \| Cohort study \| Diabetes \| Glargine \| Human insulin \| Insulin analogue \| Lispro \| Mortality \| Neoplasm | Risk of malignancies in patients with diabetes treated with human insulin or insulin analogues: A cohort study | 439 |
| 66 | -1.4169376799 | 12 | genetics \| genome-wide association study \| major depressive disorder \| mega-Analysis \| meta-Analysis | A mega-Analysis of genome-wide association studies for major depressive disorder | 342 |
| 703 | -1.5366763859 | 5 | Diagnosis \| Guide \| Osteoporosis \| Prevention \| Treatment | Clinician ✗ s Guide to Prevention and Treatment of Osteoporosis | 330 |

*Table 13- Top 10 highly-cited papers (Data Set A)*

| index | $Qtinfo_{token}$ | Token.count | cited.by | Author.Keywords | Title | Cited By |
|---|---|---|---|---|---|---|
| 3185 | -0.1146827717 | 8 | 12 | Rollover footwear \| Rollover function \| Rollover shape \| Shoe radii | The effect of rollover footwear on the rollover function of walking | 12 |
| 2218 | -0.1509477133 | 21 | 61 | Coinfection \| Dual infection \| HBV/HCV \| HBV/HCV/HDV \| HBV/HCV/HIV \| Hepatitis B \| Hepatitis C \| Interferon \| Lamivudine \| Ribavirin \| Treatment \| Triple infection | Natural history and treatment of hepatitis B virus and hepatitis C virus coinfection | 61 |
| 806 | -0.1614512524 | 17 | 70 | Amylose-only starch \| Barley \| Resistant starch \| RNA interference \| Starch bioengineering \| Starch branching enzymes \| Starch crystallinity \| Starch granules | Concerted suppression of all starch branching enzyme genes in barley produces amylose-only starch granules | 70 |
| 2766 | -0.2231159587 | 8 | 7 | Extra-adrenal myelolipoma \| Myelolipoma \| Renal myelolipoma \| Retroperitoneal tumors | Renal myelolipoma: A rare extra-adrenal tumor in a rare site: A case report and review of the literature | 7 |
| 2198 | -0.2839281577 | 6 | 7 | Mucoepidermoid carcinoma \| Salivary gland \| Sialadenoma papilliferum | Mucoepidermoid carcinoma arising in a background of sialadenoma papilliferum: A case report | 7 |
| 223 | -0.2864838449 | 28 | 14 | Carcinoembryonic antigen (CEA) \| Cytokeratin 19 fragments (CYFRA 21-1) \| Cytological fluid \| Needle aspiration biopsy (NAB) \| Non-small cell lung cancer (NSCLC) \| Squamous cell carcinoma antigen (SCC) \| Tumor marker | Additional diagnostic value of tumor markers in cytological fluid for diagnosis of non-small-cell lung cancer | 14 |
| 3093 | -0.3069447302 | 11 | 10 | Cervical fracture \| Hip fracture \| Spinal deformity index \| Trochanteric fracture \| Vertebral fracture | The assessment of vertebral fractures in elderly women with recent hip fractures: The BREAK Study | 10 |
| 2452 | -0.3195616615 | 21 | 12 | executive coaching \| health coaching \| physician burnout \| physician coaching \| physician leadership coaching \| physician resilience \| physician wellness \| professional coaching \| resilience \| work-life balance | Physician Burnout: Coaching a Way Out | 12 |
| 14 | -0.321023766 | 16 | 25 | Apo10 \| Biomarker \| DNaseX \| Early detection and diagnosis \| EDIM (epitope detection in monocytes) \| EDIM-blood test \| TKTL1 | A biomarker based detection and characterization of carcinomas exploiting two fundamental biophysical mechanisms in mammalian cells | 25 |
| 1597 | -0.3356257209 | 13 | 11 | Hepatitis B \| Hepatitis C \| Hepatocellular Carcinoma \| Pakistan \| Viral marker negative HCC \| Viral-HCC | Hepatocellular carcinoma in Native South Asian Pakistani population; Trends, clinico-pathological characteristics & differences in viral marker negative & viral-hepatocellular carcinoma | 11 |

*Table 14- Top 10 highest $Qtinfo_{token}$ scores, Data Set A*

| id | Qtinfo | Token.count | Author.Keywords | title | cited.by |
|---|---|---|---|---|---|
| 1957 | -3.8429807358 | 2 | Clinical review | Left ventricular assist device implantation in high risk destination therapy patients: An alternative surgical approach | 6 |
| 3096 | -3.2498855513 | 4 | Epidemiology \| Oncology \| Public health | The association between Acute Lymphoblastic Leukemia in children and Helicobacter pylori as the marker for sanitation | 1 |
| 2047 | -3.0752954379 | 16 | Blindness \| Cataract \| Eye health \| Human resources \| Low vision \| Nursing \| Ophthalmology \| Optometry \| sub-Saharan Africa \| Vision 2020 | Mapping human resources for eye health in 21 countries of sub-Saharan Africa: Current progress towards VISION 2020 | 15 |
| 336 | -2.9281327666 | 7 | Anthropometry \| Azithromycin \| Mass drug administration \| Trachoma control | Anthropometric indices of Gambian children after one or three annual rounds of mass drug administration with azithromycin for trachoma control | 0 |
| 939 | -2.9250142425 | 4 | Atherosclerosis \| Hypertension \| Vascular effects | Detection of a and b waves in the acceleration photoplethysmogram | 14 |
| 3402 | -2.9144128663 | 8 | Adults \| Armspan \| BMI-armspan \| BMI-height \| Elderly \| Ethiopia | The use of armspan measurement to assess the nutritional status of adults in four Ethiopian ethnic groups | 30 |
| 1551 | -2.8870044583 | 7 | Anaemia \| Body mass index \| Haemoglobin \| Kazakh \| Women | Haemoglobin status of adult non-pregnant Kazakh women living in Kzyl-Orda region, Kazakhstan | 10 |
| 1180 | -2.7018348192 | 13 | Astrocytes \| Delayed neuronal death \| Hippocampus \| Ischemic damage \| Receptor for advanced glycation end products | Effect of transient cerebral ischemia on the expression of receptor for advanced glycation end products (RAGE) in the gerbil hippocampus proper | 1 |
| 38 | -2.6188332363 | 3 | environment \| walkabilityyouthobesitysocioeconomic status | A cross-sectional study of the individual, social, and built environmental correlates of pedometer-based physical activity among elementary school children | 18 |
| 2314 | -2.5827108864 | 10 | Clinical pharmacology \| Clinical research \| Pharmacokinetics and drug metabolism \| Pharmacology \| Physiology | Optimal back-extrapolation method for estimating plasma volume in humans using the indocyanine green dilution method | 3 |

*Table 15- 10 lowest $Qtinfo_{token}$ scores, Data Set A*

| index | qtInfotoken | n | cited.by | Author.Keywords | Title |
|---|---|---|---|---|---|
| 38 | -2.6188332363 | 3 | 18 | environment \| walkabilityyouthobesitysocioeconomic status | A cross-sectional study of the individual, social, and built environmental correlates of pedometer-based physical activity among elementary school children |
| 71 | -2.0418205283 | 5 | 10 | Epidemiologic transition \| Obesity \| Physical activity | A mixed ecologic-cohort comparison of physical activity & weight among young adults from five populations of African origin |
| 215 | -2.3905906028 | 9 | 8 | Acute kidney injury \| Cardiorenal syndrome \| Meta-analysis \| Type 1 | Acute kidney injury in cardiorenal syndrome type 1 patients: A systematic review and meta-analysis |
| 215 | -2.3905906028 | 9 | 8 | Acute kidney injury \| Cardiorenal syndrome \| Meta-analysis \| Type 1 | Acute kidney injury in cardiorenal syndrome type 1 patients: A systematic review and meta-analysis |
| 701 | -2.0029970354 | 5 | 27 | Classification \| Clinical staging \| Mental health | Clinical classification in mental health at the cross-roads: Which direction next? |
| 939 | -2.9250142425 | 4 | 14 | Atherosclerosis \| Hypertension \| Vascular effects | Detection of a and b waves in the acceleration photoplethysmogram |
| 1440 | -2.0652127404 | 8 | 18 | Comparative effectiveness research \| Estimation techniques \| Heterogeneity \| Risk adjustment | From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: A primer |
| 1446 | -2.2131735338 | 3 | 107 | Genome wide \| Heritability | From monogenic to polygenic obesity: Recent advances |
| 1551 | -2.8870044583 | 7 | 10 | Anaemia \| Body mass index \| Haemoglobin \| Kazakh \| Women | Haemoglobin status of adult non-pregnant Kazakh women living in Kzyl-Orda region, Kazakhstan |
| 1590 | -2.3913145619 | 5 | 12 | Anaemia \| Children \| Haemoglobin \| Kazakh \| Stunting | Height, weight and haemoglobin status of 6 to 59-month-old Kazakh children living in Kzyl-Orda region, Kazakhstan |
| 1679 | -2.2051830026 | 7 | 17 | Chronic disease \| Guideline adherence \| Healthcare \| Quality assurance | Identifying determinants of care for tailoring implementation in chronic diseases: An evaluation of different methods |
| 1751 | -2.0620637936 | 8 | 8 | Cardiovascular diseases \| Clinical inertia \| Diabetes mellitus \| Hyperlipidemia \| Hypertension | Improving treatment intensification to reduce cardiovascular disease risk: A cluster randomized trial |
| 1754 | -2.0538355844 | 8 | 14 | Biomedical research \| Global Justice \| John Stuart Mill \| Methodology | In favour of a Millian proposal to reform biomedical research |
| 2047 | -3.0752954379 | 16 | 15 | Blindness \| Cataract \| Eye health \| Human resources \| Low vision \| Nursing \| Ophthalmology \| Optometry \| sub-Saharan Africa \| Vision 2020 | Mapping human resources for eye health in 21 countries of sub-Saharan Africa: Current progress towards VISION 2020 |
| 2167 | -2.3232325215 | 4 | 46 | Gas chromatography \| Sample preparation | Modern methods of sample preparation for GC analysis |
| 3100 | -2.4139366894 | 9 | 27 | Diabetes mellitus \| General practice \| Primary care \| Urinary incontinence \| Women | The association between diabetes mellitus and urinary incontinence in adult women |
| 3402 | -2.9144128663 | 8 | 30 | Adults \| Armspan \| BMI-armspan \| BMI-height \| Elderly \| Ethiopia | The use of armspan measurement to assess the nutritional status of adults in four Ethiopian ethnic groups |

*Table 16- Outliers- Qtinfotoken < -2 AND cited.by > 7, Data Set A*

In particular, the considerably distance of the '0' count data points was a cause for concern- whether the constant value added to the citation counts before log-scaling was the cause of the apparent correlation was deemed worthy of investigation.

### 4.2.3 Model A2: log-scaled $Qtinfo_{token}$, *cited.by* unscaled excluding 0 citation count

A second regression was run, log-scaling the input but excluding the zero-count data:

```
lm(formula = qtInfo ~ log(cited.by), data = subset(theactualthing,
    cited.by > 0))

Residuals:
     Min       1Q   Median       3Q      Max
-2.68421 -0.18427  0.01568  0.21867  1.03794

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.174646   0.012490 -94.046   <2e-16 ***
log(cited.by)  0.008862   0.004859   1.824   0.0682 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3218 on 3476 degrees of freedom
Multiple R-squared:  0.0009563, Adjusted R-squared:  0.0006689
F-statistic: 3.327 on 1 and 3476 DF,  p-value: 0.06822
```

*Model A2: Output*

This model shows no significant correlation, indicating that the apparent significance of the model is possibly either an artefact of the linear regression used rather than a genuine predictive feature.

An alternative scaling method was therefore sought. A fractional exponent scaling of the citation count produces a density plot which visually approximates a normal distribution, and so a series of models was run using fractional exponent scaling:

### 4.2.4 Model A3: $Qtinfo_{token}$ Fractional exponent-scaled *cited.by*

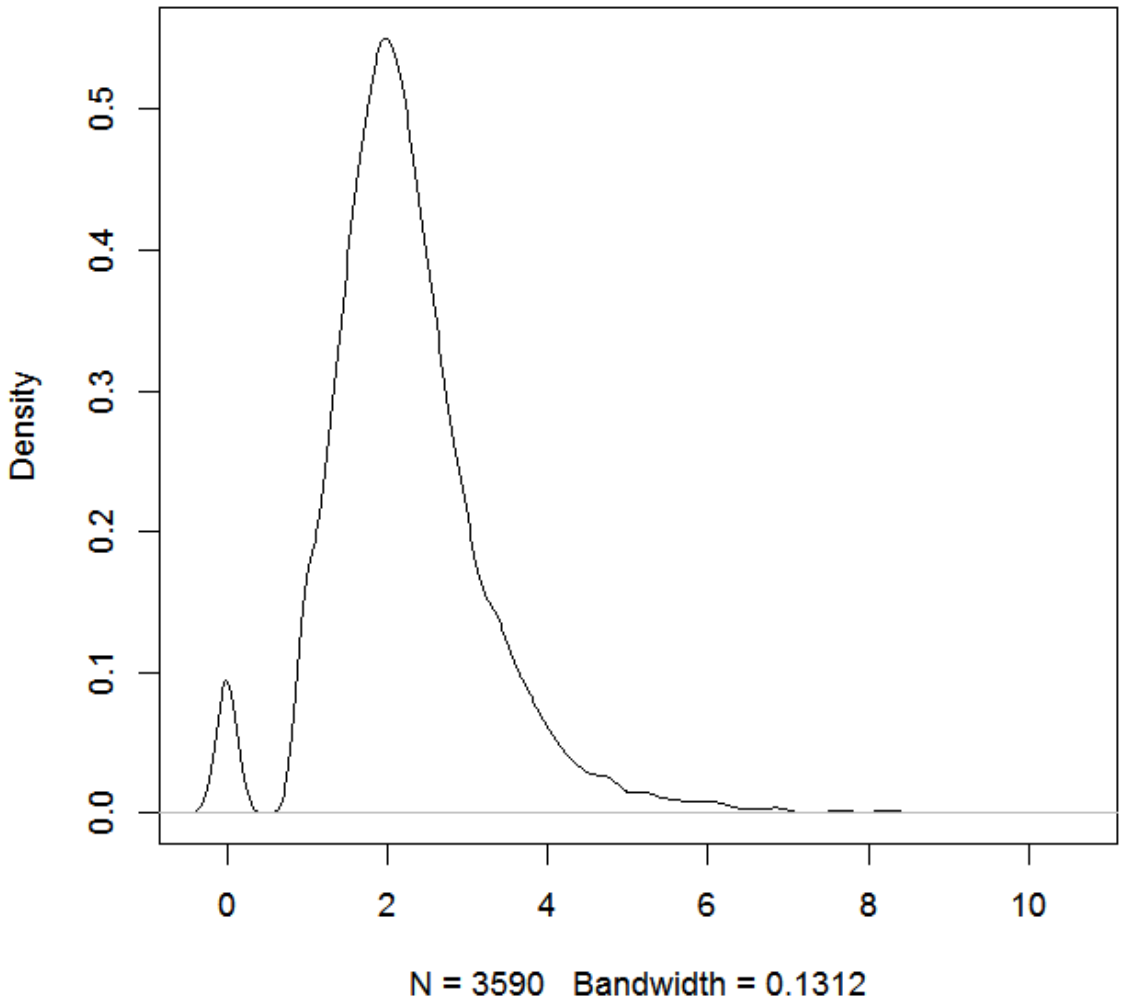


*Figure 8: Density plot for Model A3, Fractional exponent ($x^{1/10}$) scaled citation count (Data Set A)*

```
Call:
lm(formula = cited.by^(1/10) ~ qtInfo, data = theactualthing)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25747 -0.06521  0.02297  0.11852  0.78839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.27237    0.01625  78.313  < 2e-16 ***
qtInfo       0.03782    0.01353   2.795  0.00522 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2622 on 3588 degrees of freedom
Multiple R-squared:  0.002172,  Adjusted R-squared:  0.001894
F-statistic: 7.811 on 1 and 3588 DF,  p-value: 0.00522
```

*Model A3: Output*

A residuals plot was produced and visually examined for this model, which once again is very similar to the residuals distribution for the log-scaled linear model, showing many of the same features (visually asymmetric, banded, no systemic overall heteroskedastic trend).

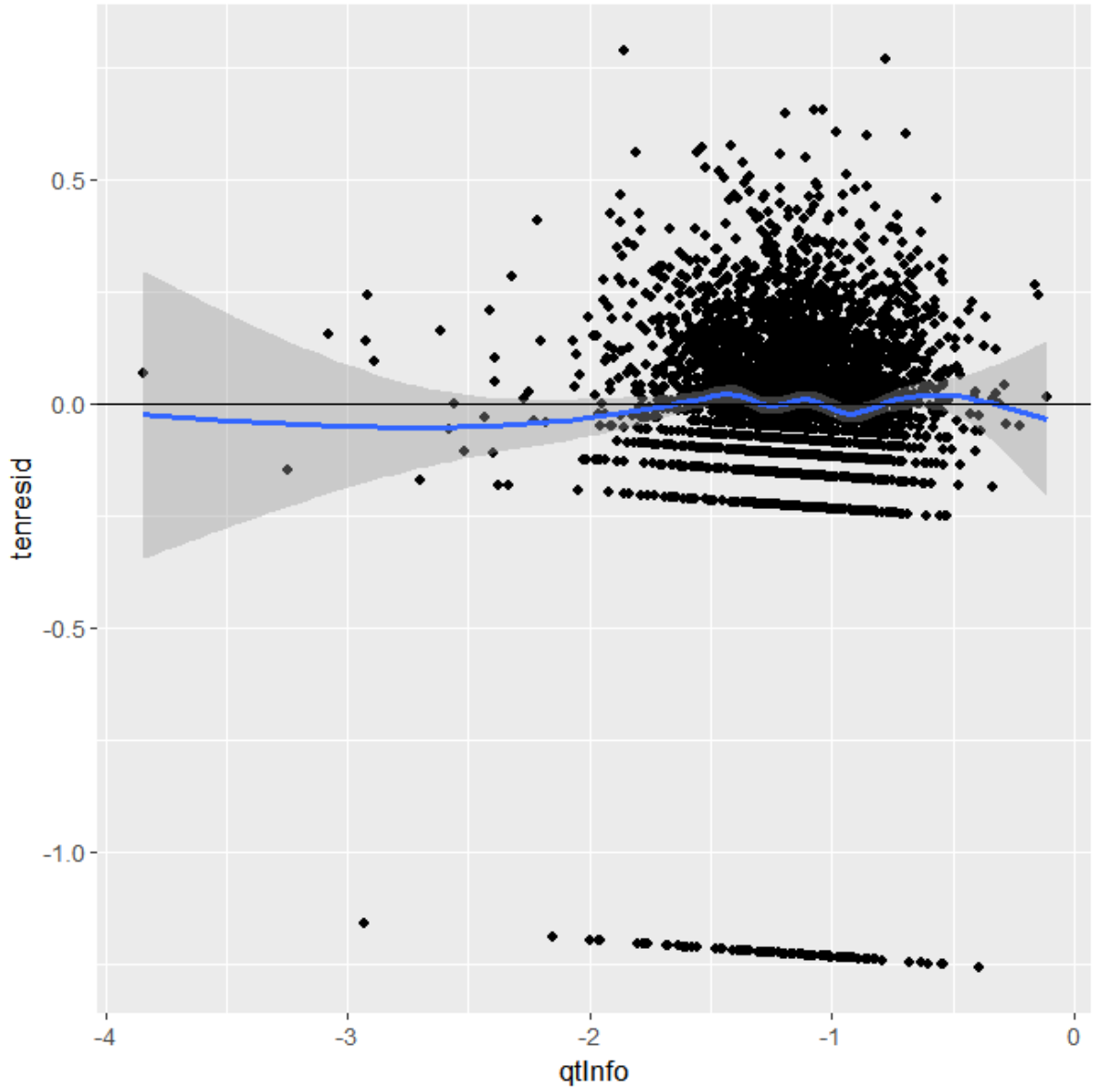


*Figure 9: Residuals Plot, Model A3*

### 4.2.5 Model A4: Association-randomised $Qtinfo_{token}$ (*RandQT*), fractional exponent scaled *cited.by*

In order to explore the possibility that the observed effect was an artefact of the distribution scaling, a 'null hypothesis' model was run manually, randomly reassigning $Qtinfo_{token}$ values by assigning a new ID value as a random permutation of the integer ID values:

```
> data$randid = sample(1:3588)
```

And then remerging the $Qtinfo_{token}$ values with the data set as a new column:

```
>data = merge(data, qtinfo, by.x ="randid", by.y="id")
```

Since all the $QtInfo_{token}$ values are retained and only the association between them and citation counts is changed, the overall distribution of the dependent and independent variables is identical to the other models featuring the same scaling.

As expected, this model shows no significant correlation at all, supporting the notion that the apparent correlation between the variables is real rather than a modelling artefact.

```
lm(formula = cited.by^(1/10) ~ randqt, data = theactualthing)


Residuals:
     Min       1Q   Median       3Q      Max
-1.22802 -0.07744  0.01883  0.11560  0.78383


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.228675   0.016528  74.340   <2e-16 ***
randqt      0.001058   0.013767   0.077    0.939
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.2641 on 3519 degrees of freedom
Multiple R-squared:  1.68e-06,  Adjusted R-squared:  -0.0002825
F-statistic: 0.005911 on 1 and 3519 DF,  p-value: 0.9387
```

*Model A4: Output*

### 4.2.6 Model A5: Independent variable *token.count,* fractional exponent scaled *cited.by*

Since the *Qtinfotoken* measure is derived by summing the number of keyword tokens, the predictive power of a model incorporating a measure of the keyword token count (*token.count)* per article was also assessed:

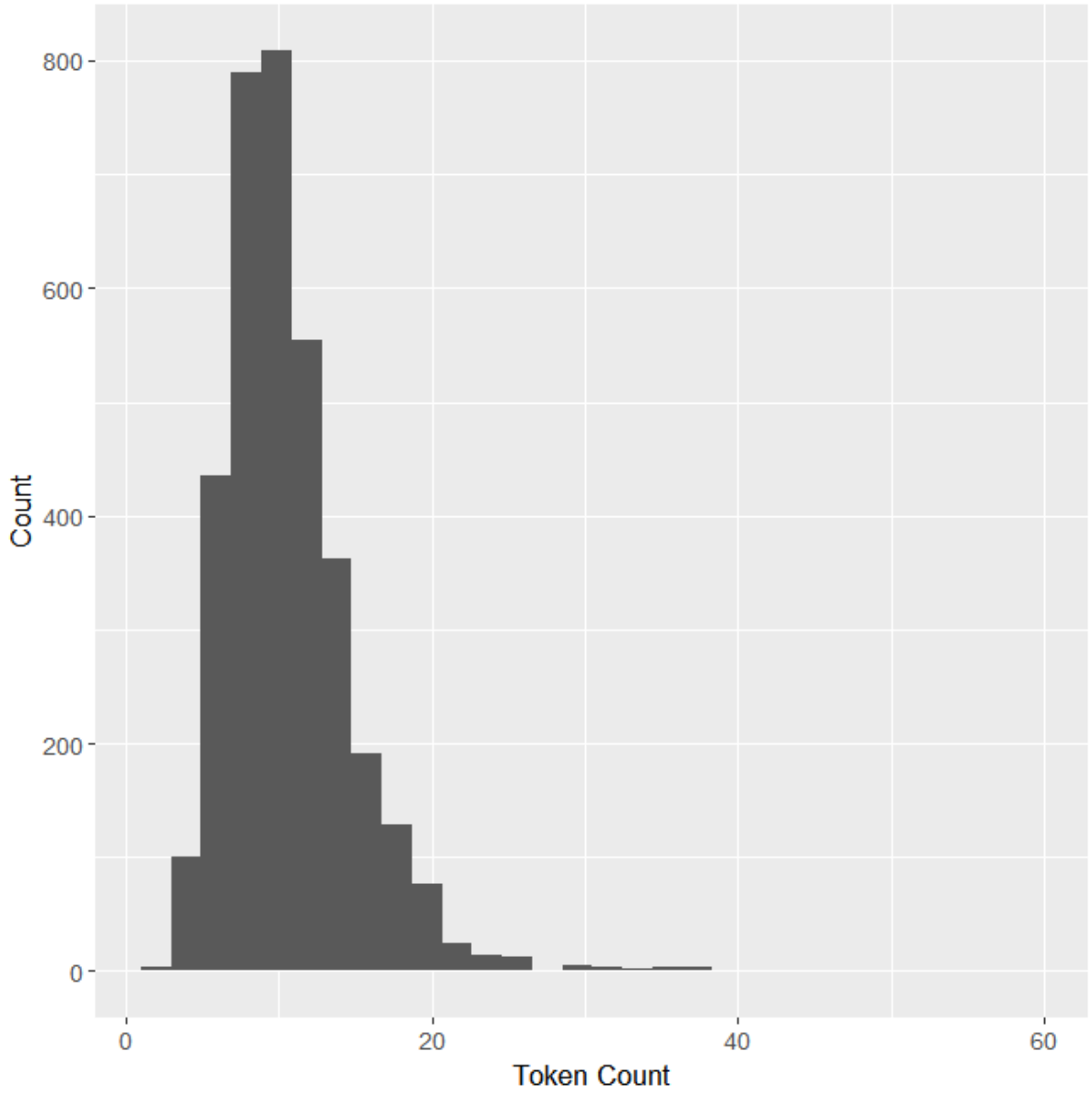


*Figure 10: Token.count Distribution Histogram*

```
lm(formula = cited.by^(1/10) ~ token.count, data = theactualthing)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25146 -0.06912  0.02332  0.11770  0.78748

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.211706   0.011550 104.906   <2e-16 ***
token.count 0.001529   0.001035   1.477     0.14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.264 on 3519 degrees of freedom
Multiple R-squared:  0.0006198, Adjusted R-squared:  0.0003358
F-statistic: 2.182 on 1 and 3519 DF,  p-value: 0.1397
```

*Model A5: Output*

### 4.2.7 Model A6: Multivariate model (*Qtinfo$_{token}$ + token.count*), fractional exponent scaled *cited.by*

```
Call:
lm(formula = cited.by^(1/10) ~ qtInfo + token.count, data = theactualthing)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25980 -0.06348  0.02385  0.11839  0.79026

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.2592157  0.0226238  55.659   <2e-16 ***
qtInfo      0.0348860  0.0142877   2.442   0.0147 *
token.count 0.0008326  0.0010730   0.776   0.4378
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2638 on 3518 degrees of freedom
Multiple R-squared:  0.00231,   Adjusted R-squared:  0.001743
F-statistic: 4.074 on 2 and 3518 DF,  p-value: 0.0171
```

*Model A6: Output*

### 4.2.8 Model A7: quasi-Poisson regression, independent variable *Qtinfo$_{token}$, cited by*

A standard Poisson model is not appropriate, since the citation count distribution is *highly* overdispersed, with the variance exceeding the mean by a factor of almost 100 (Table 1)- since distribution of the data violates the assumptions underlying Poisson regression, this approach was not attempted.

Negative binomial and quasi-Poisson models cope with overdispersion better than standard Poisson regression, and a quasi-Poisson model was therefore tested:

```
glm(formula = cited.by ~ qtInfo, family = quasipoisson, data = theactualthing)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.330  -4.013  -2.657   0.021  80.883

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.00554    0.13543  22.193   <2e-16 ***
qtInfo       0.02091    0.11301   0.185    0.853
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 94.01615)

    Null deviance: 116876  on 3589  degrees of freedom
Residual deviance: 116872  on 3588  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

*Model A7: Output*

### 4.2.9 Model A8: negative binomial model: *Qtinfo$_{token}$, cited.by*

A negative binomial model is also recommended for overdispersed count data, and this model was therefore run in addition to the linear regression:

```
glm.nb(formula = cited.by ~ qtInfo, data = theactualthing, init.theta =
0.778059901,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2598  -1.0585  -0.5878  -0.0234   8.7115

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.99984    0.07245  41.405   <2e-16 ***
qtInfo       0.01898    0.06035   0.314    0.753
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.7781) family taken to be 1)

    Null deviance: 4056.7  on 3520  degrees of freedom
Residual deviance: 4056.6  on 3519  degrees of freedom
AIC: 28063

Number of Fisher Scoring iterations: 1


              Theta:  0.7781
          Std. Err.:  0.0174

 2 x log-likelihood:  -28056.9340
```

*Model A8: Output*

### 4.2.10 Model A9: negative binomial regression model. *Qtinfo$_{token}$, cited.by* < 100.

To explore the possibility that high right skewness of the data may be obscuring the model, a negative binomial model was also run excluding any citation counts above 100 (see discussion for justification of this unorthodox practice):

```
Call:
glm.nb(formula = cited.by ~ qtInfo, data = subset(theactualthing,
    cited.by < 100), init.theta = 1.114049517, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4647  -1.0274  -0.4421   0.2253   2.8384

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.81190    0.06241  45.057   <2e-16 ***
qtInfo       0.11446    0.05205   2.199   0.0279 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.114) family taken to be 1)

    Null deviance: 3849.7  on 3426  degrees of freedom
Residual deviance: 3844.9  on 3425  degrees of freedom
AIC: 25437

Number of Fisher Scoring iterations: 1


              Theta:  1.1140
          Std. Err.:  0.0275

 2 x log-likelihood:  -25431.2290
```

*Model A9: Output*

## 4.3 DATA SET B

### 4.3.1 Data Set B summaries, distributions and general information

| Parameter | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Variance |
|---|---|---|---|---|---|---|---|
| **$Qtinfo_{kw}$** | -3.3266951823 | -1.7419802708 | -1.4408740276 | -1.499120062 | -1.1930741988 | -0.5511252547 | 0.1881167 |
| **Cited.by** | 0 | 4 | 10 | 20.3945945946 | 20 | 1094 | 2882.472 |
| **Keyword.count** | 1 | 4 | 5 | 5.2756756757 | 6 | 17 | 3.040637 |
| **$Qtinfo_{token}$** | -3.1598304612 | -1.3377258784 | -1.1401032847 | -1.1628656973 | -0.9612425329 | -0.3100872912 | 0.08385259 |

*Table 17:Summary of parameters, Data Set B*

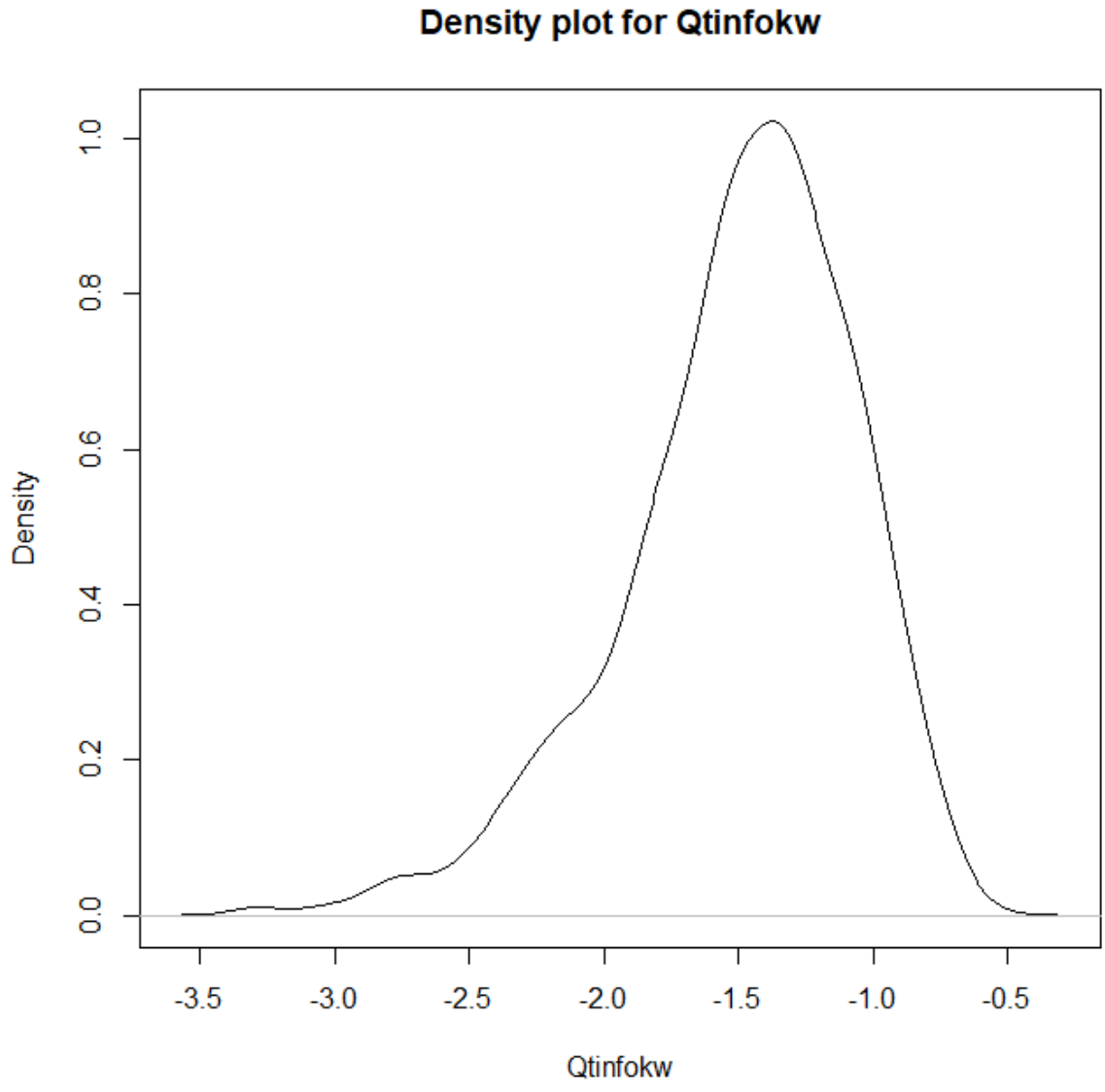


*Figure 11: Density plot for $Qtinfo_{kw}$, Model B*

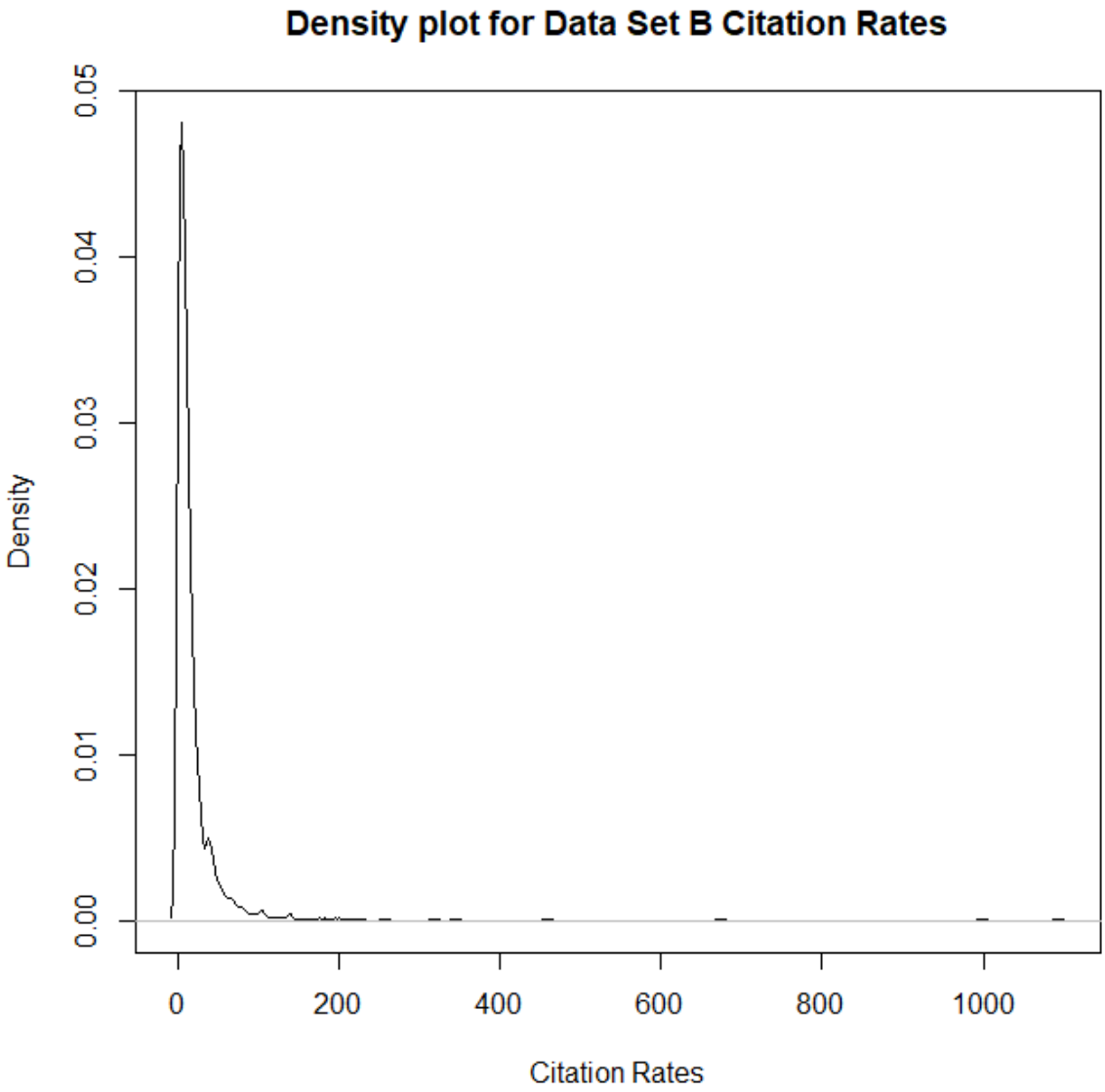


*Figure 12: Density plot for citation rates, Data Set B*

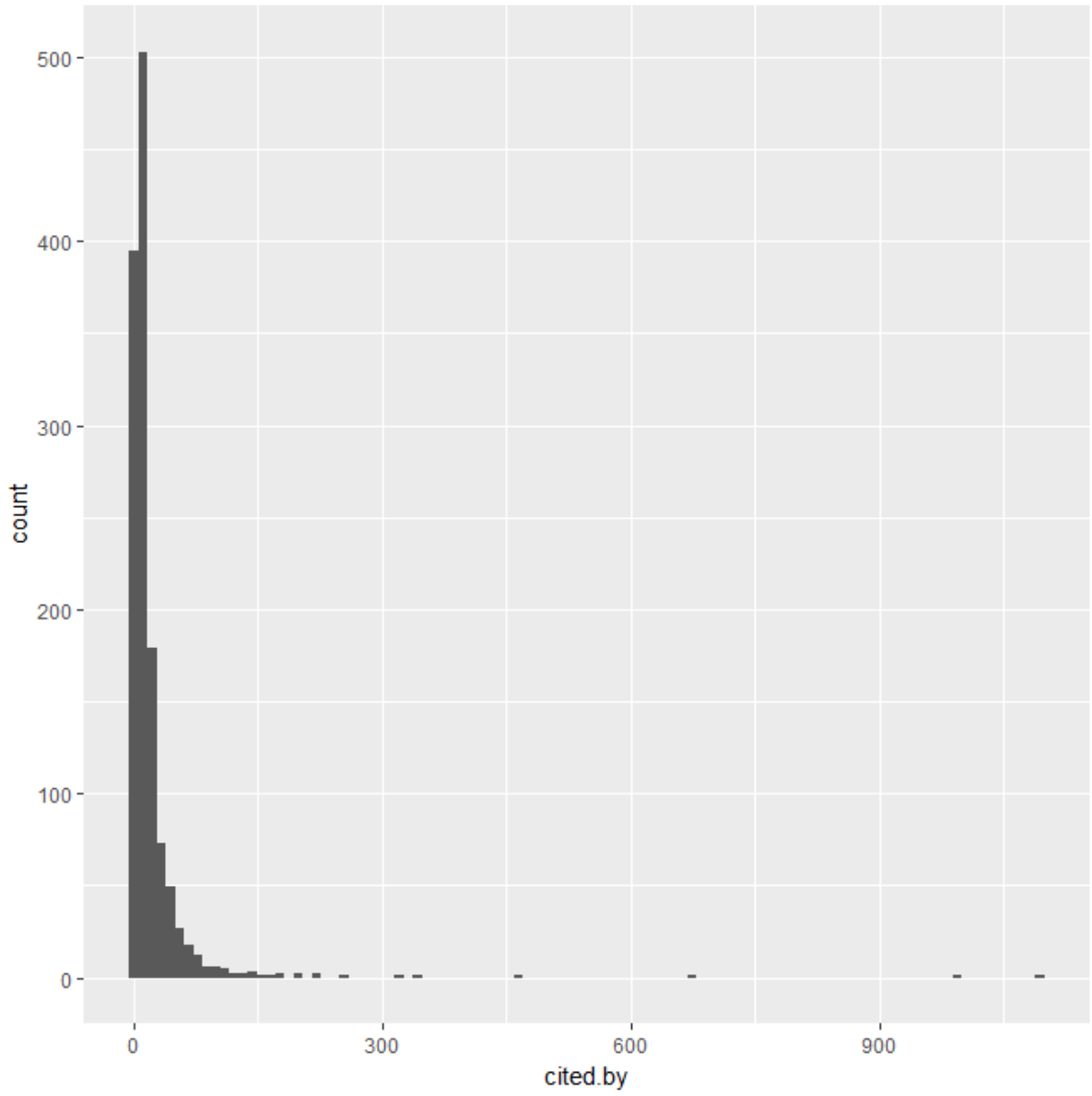


*Figure 13: Histogram for Citation Rates, Data Set B*

| id | qtInfokw | title | cited.by | Author.Keywords | DOI |
|---|---|---|---|---|---|
| 543 | -1.4249337877 | FRAX™ and the assessment of fracture probability in men and women from the UK | 1094 | Clinical risk factors \| Fracture probability \| Frax™ \| Osteoporotic fracture | 10.1007/s00198-007-0543-5 |
| 518 | -1.8740506201 | External review and validation of the Swedish national inpatient register | 999 | Classification of diseases \| disease \| epidemiology \| morbidity \| register | 10.1186/1471-2458-11-450 |
| 485 | -2.212056345 | European guidance for the diagnosis and management of osteoporosis in postmenopausal women | 673 | Bone mineral density \| Diagnosis of osteoporosis \| Fracture risk assessment \| Health economics \| Treatment of osteoporosis | 10.1007/s00198-008-0560-z |
| 484 | -2.108215856 | European guidance for the diagnosis and management of osteoporosis in postmenopausal women | 459 | Bone mineral density \| Diagnosis of osteoporosis \| Fracture risk assessment \| FRAX \| Health economics \| Treatment of osteoporosis | 10.1007/s00198-012-2074-y |
| 265 | -1.520580559 | Clinician's Guide to Prevention and Treatment of Osteoporosis | 344 | Diagnosis \| Guide \| Osteoporosis \| Prevention \| Treatment | 10.1007/s00198-014-2794-2 |
| 997 | -2.20175292 | Recent advances in 2D and 3D in vitro systems using primary hepatocytes, alternative hepatocyte sources and non-parenchymal liver cells and their use in investigating mechanisms of hepatotoxicity, cell signaling and ADME | 317 | Clearance \| Cryopreservation \| DILI*3D Models \| Mathematical modeling \| Mechanisms of gene regulation \| Non-parenchymal cells | 10.1007/s00204-013-1078-5 |
| 1197 | -1.5250551218 | The Oslo Health Study: The impact of self-selection in a large, population-based survey | 257 | Bias \| Disability benefit \| Epidemiological studies \| Equity \| Ethnicity \| Health surveys \| Non-response \| Response bias \| Response rate \| Self-selection | 10.1186/1475-9276-3-3 |
| 208 | -1.1259991559 | Born Too Soon: The global epidemiology of 15 million preterm births | 226 | epidemiology \| neonatal mortality \| Preterm birth | 10.1186/1742-4755-10-S1-S2 |
| 475 | -0.7314223089 | Epigenetic regulation of PPARGC1A in human type 2 diabetic islets and effect on insulin secretion | 216 | DNA methylation \| Epigenetic \| Gene expression \| Genetic \| Human \| Pancreatic islets \| PGC-1a \| PPARGC1A \| Type 2 diabetes | 10.1007/s00125-007-0916-5 |
| 1091 | -1.8719004606 | Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (SCMR) Board of Trustees Task Force on Standardized Post Processing | 202 | Heart \| Image interpretation \| Magnetic resonance imaging \| Post processing \| Recommendations | 10.1186/1532-429X-15-35 |
| 528 | -1.1473322514 | FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0 | 195 | FDG \| Imaging procedure \| Oncology \| PET/CT \| Quantification \| Tumour | 10.1007/s00259-014-2961-x |

*Table 18: Top 10 most-cited papers, Data Set B*

### 4.3.2 Model B1: *Qtinfo$_{token}$* , fractional exponent-scaled *cited.by*

The *Qtinfo$_{token}$* value was recalculated and run as a variable for this dataset:

```
Call:
lm(formula = cited.by^(1/10) ~ qtInfotoken, data = final)


Residuals:
     Min       1Q   Median       3Q      Max
-1.27237 -0.06790  0.02642  0.11836  0.81528


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.31636    0.03071  42.866  < 2e-16 ***
qtInfotoken  0.07513    0.02563   2.932  0.00343 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.2669 on 1293 degrees of freedom
Multiple R-squared:  0.006603,  Adjusted R-squared:  0.005835
F-statistic: 8.595 on 1 and 1293 DF,  p-value: 0.003431
```

*Model B1: Output*

### 4.3.3 Model B2: $Qtinfo_{kw}$, fractional exponent-scaled *cited.by*

```
Call:
lm(formula = cited.by^(1/10) ~ qtInfokw, data = final)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26261 -0.07153  0.03094  0.12160  0.79089

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.32830    0.02663  49.874  < 2e-16 ***
qtInfokw     0.06624    0.01707   3.881 0.000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2663 on 1293 degrees of freedom
Multiple R-squared:  0.01152,   Adjusted R-squared:  0.01075
F-statistic: 15.06 on 1 and 1293 DF,  p-value: 0.0001092
```

*Model B2: Output*

### 4.3.4 Model B3: *Keyword.count*, fractional exponent-scaled *cited.by*

```
lm(formula = cited.by^(1/10) ~ Keyword.count, data = final)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30853 -0.06522  0.03182  0.11664  0.79935

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.166605   0.023652  49.323  < 2e-16 ***
Keyword.count   0.011827   0.004257   2.778  0.00554 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.267 on 1293 degrees of freedom
Multiple R-squared:  0.005935,  Adjusted R-squared:  0.005166
F-statistic: 7.719 on 1 and 1293 DF,  p-value: 0.005542
```

*Model B3: Output*

A multivariate model was run, using 1/10$^{th}$ power scaling for the output variable and incorporating the variable *Keyword.count* (the number of Author Keywords per article).

The Pearson correlation coefficient between the variables was calculated as 0.2598211.

### *4.3.5* Model B4: Bivariate model *Keyword.count,* fractional exponent-scaled *cited.by*

```
Call:
lm(formula = cited.by^(1/10) ~ Freq, data = final)


Residuals:
     Min       1Q   Median       3Q      Max
-1.30853 -0.06522  0.03182  0.11664  0.79935


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.166605   0.023652  49.323  < 2e-16 ***
Freq        0.011827   0.004257   2.778  0.00554 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.267 on 1293 degrees of freedom
Multiple R-squared:  0.005935,  Adjusted R-squared:  0.005166
F-statistic: 7.719 on 1 and 1293 DF,  p-value: 0.005542
```

*Model B4: Output*

### 4.3.6 Model B5- Multivariate model, $Qtinfo_{kw}$ + *Keyword.count*, fractional exponent-scaled *cited.by*.

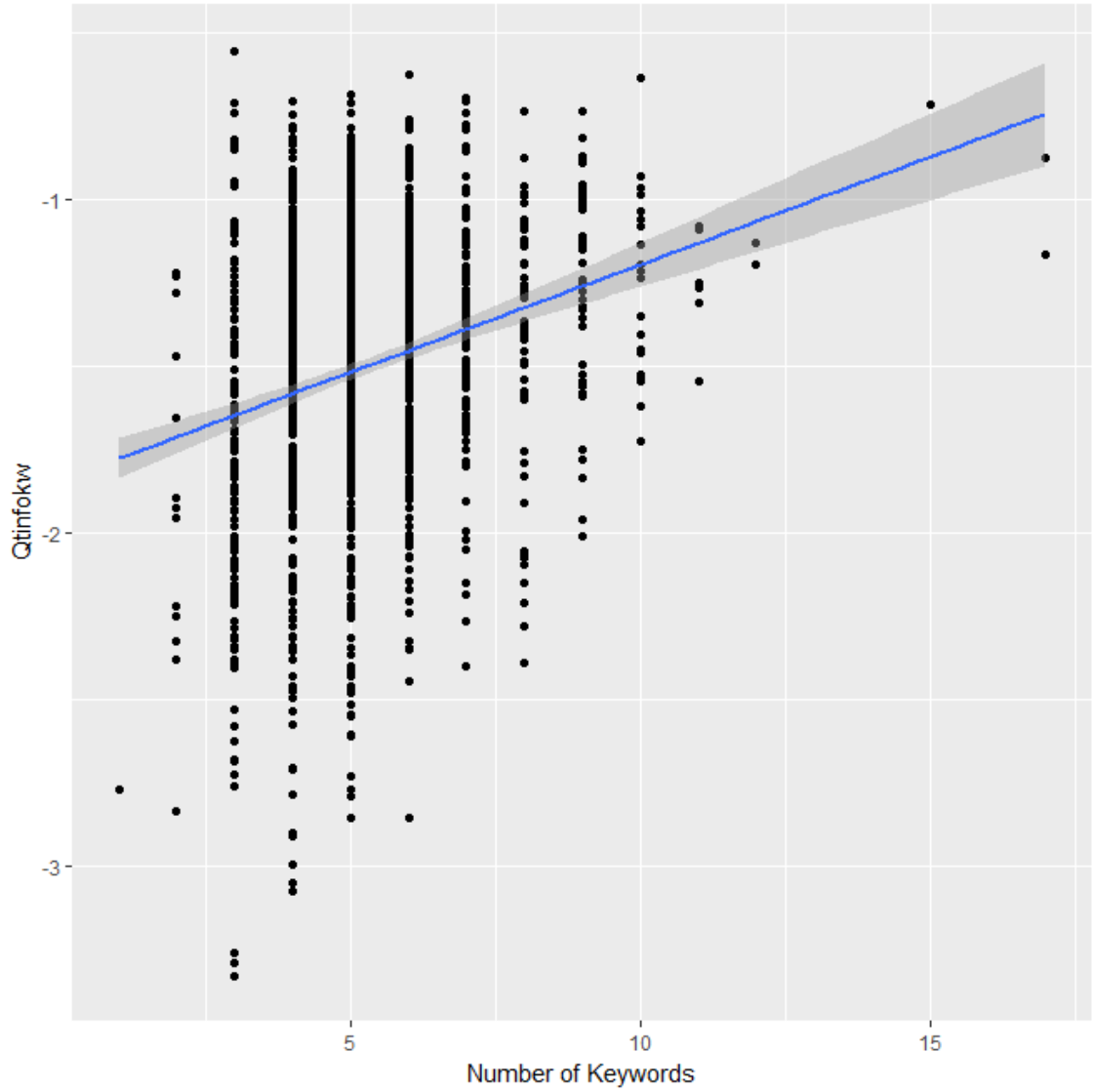


*Figure 14: Plot for Number of Author Keywords per article against $Qtinfo_{kw}$*

```
lm(formula = cited.by^(1/10) ~ Keyword.count + qtInfo, data = final)


Residuals:

     Min       1Q   Median       3Q      Max

-1.30106 -0.07170  0.02989  0.12172  0.79030


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.272927   0.040136  31.716  < 2e-16 ***

Kwno         0.008093   0.004392   1.843  0.06559 .

qtInfo      0.057782   0.017657   3.272  0.00109 **

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.266 on 1292 degrees of freedom

Multiple R-squared:  0.01411,   Adjusted R-squared:  0.01258

F-statistic: 9.243 on 2 and 1292 DF,  p-value: 0.0001033
```

*Model B5: Output*

### 4.3.7 Model B6: Multivariate model $Qtinfo_{token}$,+ $Qtinfo_{kw}$, + *Keyword.count,* fractional exponent-scaled *cited.by*

*4.3.8*

Finally, a multivariate linear regression was run, incorporating $Qtinfo_{token}$, $Qtinfo_{kw}$, and *Keyword.count*:

| | $Qtinfo_{kw}$ | Keyword.count | $Qtinfo_{token}$ |
|---|---|---|---|
| **$Qtinfo_{kw}$** | 1 | 0.2598211 | 0.616551 |
| **Keyword.count** | 0.2598211 | 1 | 0.2168864 |
| **$Qtinfo_{token}$** | 0.616551 | 0.2168864 | 1 |

*Table 19: Pearson Correlation Coefficients for model B6 independent variables*

```
Call:
lm(formula = cited.by^(1/10) ~ qtInfo + qtInfotoken + Keyword.count, data = final)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30154 -0.06995  0.02981  0.12114  0.79907

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.284094   0.044855  28.628   <2e-16 ***
qtInfo         0.050497   0.021960   2.299   0.0216 *
qtInfotoken    0.018163   0.032537   0.558   0.5768
Keyword.count  0.007910   0.004405   1.796   0.0728 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2661 on 1291 degrees of freedom
Multiple R-squared:  0.01434,   Adjusted R-squared:  0.01205
F-statistic: 6.263 on 3 and 1291 DF,  p-value: 0.0003214
```

*Model B6: Output*

| Model | Data Set | Regression type | Independent Variable 1 | Independent Variable 2 | Independent Variable 3 | Transformation of Dependent Variable (cited.by) | p-value | Adjusted $R^2$ | F-statistic | Estimates |
|---|---|---|---|---|---|---|---|---|---|---|
| A1 | A | Linear | *Qtinfotoken* | None | None | Log ($x + 10^{-4}$) | 0.00444 | 0.001976 | 8.105 | 0.3359 |
| A2 | A, cited.by > 0 | Linear | *Qtinfotoken* | None | None | Log | 0.0682 | 0.0006689 | 3.327 | 0.008862 |
| A3 | A | Linear | *Qtinfotoken* | None | None | $x^{1/10}$ | 0.00522 | 0.001894 | 7.811 | 0.03782 |
| A4 | A | Linear | *Qtinfotoken* (association randomised) | None | None | $x^{1/10}$ | 0.939 | -0.0002825 | 0.005911 | 0.001058 |
| A5 | A | Linear | *Token.count* | None | None | Log | 0.14 | 0.0003358 | 2.182 | 0.001529 |
| A6 | A | Linear | *Qtinfotoken* | Token.count | None | $x^{1/10}$ | 0.0147 | 0.001743 | 4.074 | 0.0348860<br>0.0008326 |
| A7 | A | Quasi-Poisson | *Qtinfotoken* | None | None | None | 0.853 | N/A | N/A | 0.02091 |
| A8 | A | Negative Binomial | *Qtinfotoken* | None | None | None | 0.753 | N/A | N/A | 0.01898 |
| A9 | A, cited.by < 100 | Negative Binomial | *Qtinfotoken* | None | None | None | 0.0279 | N/A | N/A | 0.11446 |
| B1 | B | Linear | *Qtinfotoken* | None | None | $x^{1/10}$ | 0.00343 | 0.005835 | 8.595 | 0.07513 |
| B2 | B | Linear | *Qtinfokw* | None | None | $x^{1/10}$ | 0.000109 | 0.01075 | 15.06 | 0.06624 |
| B3 | B | Linear | *Keyword.count* | None | None | $x^{1/10}$ | 0.00554 | 0.005166 | 7.719 | 0.011827 |
| B4 | B | Linear | *Qtinfokw* | *Keyword.count* | None | $x^{1/10}$ | 0.000104 | 0.01258 | 9.243 | 0.057782<br>0.008093 |
| B5 | B | Linear | *Qtinfokw* | *Qtinfotoken* | None | $x^{1/10}$ | 0.0004428 | 0.01035 | 6.263 | 0.05697<br>0.02252 |
| B6 | B | Linear | *Qtinfokw* | *Qtinfotoken* | *Keyword.count* | $x^{1/10}$ | 0.0003214 | 0.01205 | 6.263 | 0.050497<br>0.018163<br>0.007910 |

*Table 20- Model Summary*

# 5 DISCUSSION

## 5.1 GENERAL REMARKS

Interpreting these results is not straightforward- for one thing, the bare fact of the matter is that even in the strongest of the models detailed above, the effect measured accounts only for a very small proportion (c. 1%) of the total variance in citation rates- this is not out of line with what we might expect, but needs to be borne in mind at all times.

Most important is to avoid the danger of over-interpretation. Although the *QtInfo* measure correlates well with human evaluation of metadata quality in one, relatively small study (and the enduring effectiveness and popularity of the tf-idf measure in document retrieval and ranking bears this out), what certainly *cannot* be deduced from these results is the notion of any statistical link between metadata quality and citation rates.

It is also clear that, at least with the current data set, the effect is rather weak- the more strongly-performing of the models accounts for only approximately 1% of variance. However, this is not out of line with expectations- it would be very surprising, given the heterogeneity of discovery mechanisms for academic papers, if the keyword quality accounted for anything more than a relatively small proportion of the variance in citation rates.

## 5.2 STATISTICAL SIGNIFICANCE AND RELATIVE PERFORMANCE OF MEASURES

The $Qtinfo_{token}$ P-value (0.004438) for Data Set A is small enough to indicate that the result, while very small, *is* highly significant.

The p-value for the same measure calculated for Data Set A is 0.00343, which is in line with its performance not The p-value for the keyword field-based measure ($Qtinfo_{kw}$) used in Data Set B performs over an order of magnitude better at 0.000109, indicating that this measure (for this data set at least), or related measures, is the one worth pursuing in any further research.

The $R^2$ value for $Qtinfo_{kw}$ also outperforms $Qtinfo_{token}$ by a factor of approximately 20, indicating that (once again, for this data set!) the former measure is not only substantially more significant but also much more predictive than the latter.

| Model | Measure | $R^2$ | p |
|---|---|---|---|
| B1 | $Qtinfo_{token}$ | 0.005835 | 0.00343 |
| B2 | $Qtinfo_{kw}$ | 0.01075 | 0.000109 |

*Table 21: Key values for models B1 and B2*

Interestingly, the multivariate model including both measures (B5) performs *worse* than the bivariate $Qtinfo_{kw}$ model, indicating that any semantic 'noise' added by calculating values for constituent parts of keyword phrases outweighs any benefit that improved string-matching sensitivity might bring.

## 5.3 PROPORTION OF VARIANCE ACCOUNTED FOR

As discussed earlier, with the large variety of factors impacting citation rates, the fact that only a small proportion of the total variance is accounted for by the measures under investigation is entirely expected but with an $R^2$ level of just 0.002254 it is clear that the predictive utility of this measure at present is limited- the reasons for this will be discussed later.

The total variance we expect to have *even with a perfect capturing of keyword quality* therefore is very likely to be a relatively small proportion of total variance even under the most optimistic possible interpretation of the influence of keyword-based database search mechanisms.

The extent to which even the best-performing measure in this study, $Qtinfo_{kw}$, captures keyword quality is perhaps not one which can ever be fully quantified. However, there are at least three important respects in which the measure at present fails to do this.

### 5.3.1 *$Qtinfo_{kw}$ does not assess all keywords attached to articles.*

Firstly, only a subset of the keywords attached to the articles (Author Keywords) have been calculated in this study. For most papers, the number of Author Keywords is comparatively tiny compared to the total of Indexed Keywords which clearly also assist retrieval. For example, the article "*Left ventricular assist device implantation in high risk destination therapy patients: An alternative surgical approach*" (Samuels *et al.*, 2012), the article with the lowest $Qtinfo_{token}$ score in Data Set A, has just one Author Keyword ("Clinical Review") but a total of 16 distinct Index Keywords from MeSH and EMTREE. This level of disparity is not atypical, and it is not unreasonable to assume that were these Index Keywords analysed, a far greater proportion of total citation rate variance would be accounted for.

### 5.3.2 *$Qtinfo_{kw}$ does not accurately match all high-quality keywords.*

As a relatively crude string-matching measure, only some keywords will be accurately matched under the current analysis. Although terms were stemmed in order to reduce the effect of pluralization, no attempt was made to account for e.g. the effect of variant US and UK spellings ('colour' would not match to 'color' for example). The stemming procedure may also have produced a number of false positive matches.

### 5.3.3 *$Qtinfo_{kw}$ does not directly measure keyword quality*

Even with perfect matching and perfect keyword capture, it is important to remember that (intuitively) the measures under investigation here are only correlated with, not identical to, a *true* measure of keyword quality. Such a 'true' measure would require access to data which in all likelihood will never be available- however, theoretically, if information on which search terms are input into Scopus, along with information on whether a particular article was downloaded as the result of a particular search term input, were available and could be matched up with information on citations, a much improved quantification of keyword quality could perhaps be made.

## 5.4 IS THE OBSERVED EFFECT REAL?

With an effect of this size question however it is particularly important to explore whether or not the effect measured is real, or whether it is an artefact generated by the inappropriate application of a linear model regression model to data where the dependent variable is not normally distributed, or whether some quirk of the data preparation model results in the apparent correlation.

The initial approach taken to scaling of the dependent variable (logarithmic, Model A1) yielded empirically effective results but necessitated the addition of a small constant to all values in order to prevent citation rates of 0 from resulting in invalid values. When 0 count data was removed, the effect disappears, leading to concerns that this procedure may have been in some way responsible for the observed effect.

The disappearance of predictive significance upon removal of a relatively small number of 0- citation count data points from the log-scaled model (model A2), as well as the fact that a negative binomial model (Model A8) yields significant (although not incredibly so) results.

Other models and scaling approaches were therefore tried:

Fractional exponent scaling, from cube root scaling upward, were tried, with $x^{1/10}$ scaling empirically determined to produce results comparable to those of log-scaling:

| Model | Scaling | p | $R^2$ |
|---|---|---|---|
| A1 | $Log(x + 10^{-4})$ | 0.00444 | 0.001976 |
| A3 | $x^{1/10}$ | 0.00522 | 0.001894 |

*Table 22:Key values for models with dependent value scaling changes*

In order to further test the hypothesis that some property of the distribution scaling was responsible, a model was also tested which kept the output scaling (and therefore distribution of the transformed output) identical, but randomised the association between the independent and dependent variables.

| Model | Scaling | Association | P | $R^2$ |
|---|---|---|---|---|
| A3 | $x^{1/10}$ | Original | 0.00522 | 0.001894 |
| A4 | $x^{1/10}$ | Randomised | 0.939 | -0.0002825 |

*Table 23: Effect of variable association randomisation*

This randomisation demonstrates that the observed effect must be due to the association between the independent and dependent variable, as this is the only change between the two models.

## 5.5 CITATION RATE DISTRIBUTION AND MEANS OF DISCOVERY FOR ACADEMIC PAPERS

The most appropriate model for the distribution of citation counts has been a matter of considerable debate in the bibliometric community, with power-law, lognormal, negative binomial, and other distributions have been suggested. Log and fractional exponent scaling have been used with some success in bibliometric studies. (Katchanov, 2015, Brzezinski, 2015, Thelwall, 2016).

Evidently, there are many other mechanisms than keyword search by which academic papers can be discovered, and which affect citations rates- as noted in the literature review, the quality and noteworthiness of the research itself is a major factor, as is the timeliness and relevance of the research to other strands of research, the various properties of the title, and various properties of the textual makeup of the papers as a whole (for example the proportion of the paper which is taken up by mathematical equations) (Wesel,Wyatt and Haaf, 2014)which all will exert an influence on the likelihood that a paper will be cited. (Letchford,Moat and Preis, 2015, Nair and Gibbert, 2016) There are also other important discovery mechanisms for papers, such as a paper being cited in another influential paper or published in a high-profile journal (although interestingly Journal Impact Factor itself is not predictive of citation rates for individual papers due to the high right skew of citation rate distributions (Prathap,Mini and Nishy, 2016)), which will ensure that it is exposed to a large potential readership without search terms having to be involved.

It should be clear here that the matter of the appropriate distribution of citation rate data is of absolutely key importance in this analysis: of particular interest is the suggestion that a single distribution may not be appropriate for modelling citation rates since there is some evidence that the causative mechanism behind early citations (soon after publication) is different from that which takes place a longer period after discovery- if the 'first wave' of citation data is due primarily to discovery mechanisms other than keyword-based database search, then isolating and analysing the distribution of 'second wave' citations (perhaps by conducting an analysis of citations in papers 10+ years old and subtracting citations received in the first couple of years after publication) may reveal the appropriate distribution which better fits the analysis. (Low,Wilson and Thelwall, 2016)

The failure of quasi-Poisson models to account for this relationship is not necessarily surprising when one considers a primary assumption of the Poisson distribution- event independence. There is evidence that citation rates are a phenomenon where something of a feedback loop applies- that accruing a large number of citations may itself lead to a paper's accruing further citations, and

therefore the consideration of citations for papers as statistically independent phenomena may simply be false, in a way which does not apply to many other count-based data types, where event occurrences or frequencies may be properly considered as independent from one another.

| Model | Data Set | Type | p |
|---|---|---|---|
| A7 | A | Quasi-Poisson | 0.853 |
| A8 | A | Negative Binomial | 0.753 |
| A9 | A, cited.by < 100 | Negative Binomial | 0.0279 |

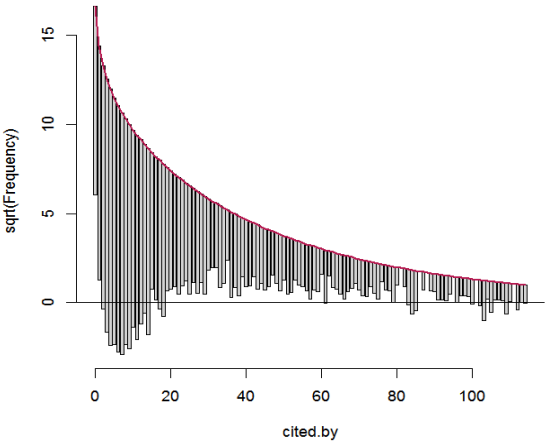


*Figure 15: Rootogram, Model A8*

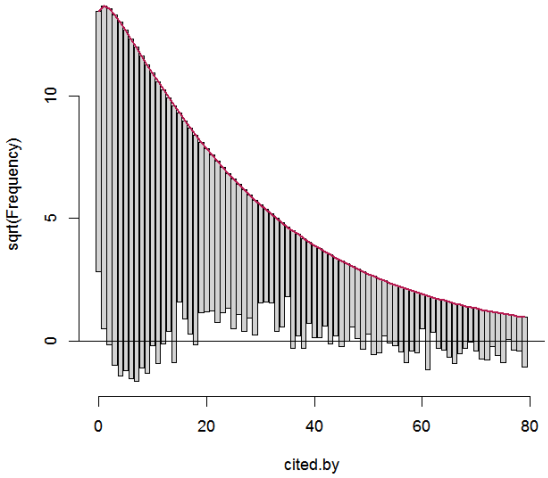


*Figure 16: Rootogram, Model A9*

Figures 15 and 16 provide a means of visualising the extent to which the observed distribution of citation rates diverge from that required by the negative binomial mode- in both, citation rates up to approximately 15 are overcounted, while higher citation rates are undercounted compared to the modelled distribution. Removing higher citation rates as in model A9 produces a somewhat better fit visually, with roughly equal levels of undercounting and overcounting, but it is nonetheless clear that the binomial distribution assumptions are not a good fit for the observed distribution of citation rates, which would account for the poor model fit.

Each citation for a particular paper provides one more avenue by which that paper may be discovered by a reader searching for information, thereby raising the probability of a further citation. Attempts to account for the highly skewed nature of the citation counts in this analysis have some justification when one considers the mechanisms by which very highly-cited papers will be discovered by those who cite them- accrue large numbers of citations- placement in visible journals, author eminence, eminence of author-associated institutions can all contribute.

We would therefore expect keyword-based database search to account for a higher proportion of citations in low-citation rate papers with high quality keywords. It may well be, ironically, that the underlying rate of *keyword search* based citation rates for papers, based as it is on a presumably independent set of researchers keying in terms and assessing results, may be more properly modelled as a process satisfying the requirement of statistical independence- it may be that this underlying process, which is ultimately what is under investigation here, in a way which a citation

discovered by other means, such as one which is the result of a secondary citation in another paper, is not. Sadly, however, isolating only those citations which are the result of database searches is probably impossible. This may, however, go some way to explaining the apparent sensitivity of the models to the exclusion of zero citation counts, and the increased effectiveness of the models which exclude higher citation counts- it seems plausible that the mechanisms by which papers accrue extremely high citation counts orders of magnitude higher than their peers are likely to be those governed by 'success breeds success' mechanisms.

## 5.6 SAMPLING ISSUES, HIGH- AND LOW- PERFORMING KEYWORD TYPES

| Rank | keyword | id | n | tf | idf | tf_idf |
|---|---|---|---|---|---|---|
| **1** | tr4 | 1221 | 221 | 0.0354337021 | 6.0738122885 | 0.2152176552 |
| **2** | plin5 | 1192 | 200 | 0.0238635008 | 7.1724245771 | 0.1711591595 |
| **3** | legionella | 624 | 71 | 0.0256225189 | 5.5629866647 | 0.1425377312 |
| **4** | goishi tea | 588 | 86 | 0.0185945946 | 7.1724245771 | 0.1333683273 |
| **5** | mif | 804 | 176 | 0.0215264188 | 6.0738122885 | 0.130747427 |
| **6** | hornerin | 615 | 146 | 0.0176477699 | 7.1724245771 | 0.1265772982 |
| **7** | ppargc1a | 475 | 102 | 0.0175529169 | 7.1724245771 | 0.1258969724 |
| **8** | tendon | 312 | 163 | 0.0271802568 | 4.6074752197 | 0.1252323596 |
| **9** | batf | 194 | 88 | 0.0172616712 | 7.1724245771 | 0.1238080351 |
| **10** | desmopressin | 348 | 96 | 0.0171092497 | 7.1724245771 | 0.122714803 |
| **11** | honey | 859 | 311 | 0.0269801336 | 4.464374376 | 0.1204494171 |
| **12** | migrain | 648 | 149 | 0.0292558414 | 4.0813821238 | 0.1194042679 |
| **13** | endostatin | 505 | 84 | 0.0169834209 | 6.4792773966 | 0.1100402955 |
| **14** | osteocalcin | 259 | 94 | 0.0220398593 | 4.9751999998 | 0.1096527081 |
| **15** | vur | 787 | 117 | 0.0168272688 | 6.4792773966 | 0.1090285424 |
| **16** | middl manag | 1263 | 141 | 0.0149064383 | 7.1724245771 | 0.1069153045 |
| **17** | paclitaxel | 530 | 123 | 0.0241365777 | 4.3998358549 | 0.10619698 |
| **18** | cd34 | 232 | 198 | 0.0233931947 | 4.464374376 | 0.104435979 |
| **19** | pyuria | 253 | 130 | 0.0187671431 | 5.5629866647 | 0.1044013666 |
| **20** | mrsa | 858 | 121 | 0.0230300723 | 4.464374376 | 0.1028148648 |
| **21** | oxaliplatin | 914 | 114 | 0.0190731136 | 5.3806651079 | 0.1026260369 |
| **22** | leprosi | 560 | 96 | 0.0190362879 | 5.3806651079 | 0.1024278902 |
| **23** | cari | 50 | 183 | 0.025061627 | 4.0369303612 | 0.1011720427 |
| **24** | iron | 1070 | 237 | 0.0355962752 | 2.8157157504 | 0.1002289926 |
| **25** | hepcidin | 864 | 84 | 0.0171533592 | 5.786130216 | 0.09925157 |

*Table 24: 25 highest-ranking tf-idf values, Data Set B*

| | keyword | id | n | tf | idf | tf_idf |
|---|---|---|---|---|---|---|
| 1 | in | 741 | 125 | 0.0194159677 | 0 | 0 |
| 2 | diabet | 582 | 2 | 0.0001895735 | 0.1318881869 | 2.50024998879411e-05 |
| 3 | mechan of gene regul | 997 | 1 | 5.46134731438246e-06 | 7.1724245771 | 3.91711017018915e-05 |
| 4 | review | 1200 | 4 | 0.0003935071 | 0.1110902102 | 4.3714790049939e-05 |
| 5 | diabet | 84 | 3 | 0.0003638128 | 0.1318881869 | 4.79826049874689e-05 |
| 6 | human | 831 | 1 | 0.0001186521 | 0.4050814519 | 4.80637697982265e-05 |
| 7 | review | 59 | 12 | 0.0004888382 | 0.1110902102 | 5.43051377942354e-05 |
| 8 | has | 1016 | 37 | 0.0035950253 | 0.0185907755 | 6.68343077343657e-05 |
| 9 | therapi | 258 | 1 | 9.15164271986822e-05 | 0.828544143 | 7.5825399743618e-05 |
| 10 | record | 237 | 6 | 0.000100447 | 0.8163169164 | 8.19965764079106e-05 |
| 11 | diabet | 617 | 3 | 0.0006410256 | 0.1318881869 | 8.45437095569805e-05 |
| 12 | diabet | 853 | 5 | 0.0006607638 | 0.1318881869 | 8.71469452285513e-05 |
| 13 | common | 1313 | 16 | 0.0025835621 | 0.0343515431 | 8.87493443061467e-05 |
| 14 | consent | 191 | 1 | 0.000116225 | 0.8250353675 | 9.58897451730399e-05 |
| 15 | determin | 242 | 6 | 0.0005811701 | 0.1693591183 | 9.84264538967744e-05 |
| 16 | diabet | 128 | 9 | 0.0007515658 | 0.1318881869 | 9.91226456935287e-05 |
| 17 | effect | 1047 | 6 | 0.0013178124 | 0.0823477413 | 0.0001085189 |
| 18 | morbid | 518 | 1 | 9.48946669197191e-05 | 1.1489769842 | 0.0001090318 |
| 19 | marker | 191 | 1 | 0.000116225 | 0.9538244574 | 0.0001108583 |
| 20 | public health | 1141 | 1 | 0.0001097333 | 1.034697523 | 0.0001135408 |
| 21 | softwar | 14 | 1 | 0.000120308 | 1.028238943 | 0.0001237054 |
| 22 | assess | 1179 | 10 | 0.0007556865 | 0.1748285941 | 0.0001321156 |
| 23 | epidemiolog | 1255 | 1 | 0.000152045 | 0.9128431131 | 0.0001387932 |
| 24 | predict | 317 | 3 | 0.0002813203 | 0.4940824625 | 0.0001389954 |
| 25 | access | 114 | 18 | 0.0020033389 | 0.0715726682 | 0.0001433843 |

*Table 25: 25 lowest tf-idf value keywords, Data Set B*

Examination of Tables 16 and 17 is instructive when considering one evident shortcoming (or at least feature) of the current research. Among the 25 worst-performing keywords, the string 'diabet' (the stemmed version of 'diabetes', the search term used to select the papers, accounts for five of them- not surprising when one considers that this is the one term which was used to select the dataset- what is however surprising is that the idf of this term is not 0!

However, this does raise the question what an appropriate sample selection procedure should be in order to fairly assess the discriminatory capacity of *all* keywords assigned to a paper. In some respects, restricting papers to those in some way related to diabetes seems appropriate- but when considered against the entire 'real' corpus under consideration (ie the entire body of literature indexed by Scopus), it is clear that this method of selection is not likely to provide a level playing field. However, this is not necessarily a problem for the present study, where the aim is to demonstrate the *existence* rather than the precise properties of a keyword quality effect on citation rates.

## 5.7 RELATIONSHIP OF QTINFO TO 'REAL' METADATA QUALITY

Other badly-performing keywords are of the types (general, common terms such as 'software' which appear in many articles) that one might expect to generally speaking perform poorly in this type of

test- however, it is not necessarily the case that this makes them poor keyword choices- for example the term 'review' appears twice in the bottom 25, perhaps unsurprisingly given that the vast majority of papers in the sample will contain the term 'review' as part of the 'literature review' heading- indeed, the idf value for 'review' is even lower than that for 'diabet' – but this does not necessarily make the term a bad descriptive keyword attached to a paper (a literature review for example). These types of features highlight the large gap between the statistical measure under consideration and the *real* property of keyword quality.

High-performing keywords tend to be highly specific nouns or phrases, which appear with great regularity in papers. This type of keyword seems a plausible candidate for one type of keyword which can efficiently discriminate and aid retrieval of a paper if attached to the metadata- however, other equally effective if more general and semantically rich terms may not be picked up by this measure at all- although it may also be that many keywords which may theoretically be considered to provide a good description of article content may fail as keywords for the exact reason that they are not picked up by the *Qtinfo* measure- that is, that the retrieval and ranking algorithms in Scopus and similar databases may simply not effectively match them at a high enough rate, or users may not think to input them.

## 5.8 MECHANISMS OF ACTION

That said, the range of possible mechanism by which a higher *Qtinfo* value may account significantly for any proportion of the variance in citation rates needs consideration- the hypothesis (that a *Qtinfo* score for keywords aids the retrieval of documents considered to be relevant and useful by researchers, and that that increased rate of retrieval is reflected in higher citation counts) is one possible interpretation.

Other mechanism by which citation rates and *Qtinfo* values may be related are discussed below as confounding factors.

## 5.9 POTENTIAL CONFOUNDING FACTORS INCLUDED IN MULTIVARIATE ANALYSIS

The potential for confounding factors is considerable in this analysis- with the effect as small and delicate as it is, the possibility that the correlation is due to some third factor must be considered sensitivity to any potential confounding factor is considerable.

The possibility that the observed correlation is due to some unnoticed quirk of the calculation procedure or the tools used to calculate the result is certainly there, although none of the procedures used to calculate *Qtinfo* values are particularly esoteric. The possibility that the effect was due to e.g. rounding errors in the calculation of values cannot be completely eliminated in the current analysis, although no values were explicitly rounded and *Qtinfo* values are given to 11 decimal places.

The parameter with the most potential to produce a false positive result is the simple count of keywords (or keyword tokens)- this is a result of the fact that the *qtInfo* parameters are calculated on the basis of the *sum* of the tf-idf values for each token. It is to be expected that these values (included in the models as *token.count* and *Keyword.count)* will then be positively correlated with

$Qtinfo_{token}$ and $Qtinfo_{kw}$ respectively, and this is indeed the case (the Pearson Correlation Coefficient between $Qtinfo_{token}$ and *token.count* is 0265835, while between $Qtinfo_{kw}$ and *Keyword.count* it is 0.2598211).

Multivariate linear models were therefore run to test for this possibility. The *token.count* measure was found to perform poorly and not to yield significant results when used on its own (Model B5)- however, the effect of the *Keyword.count* measure did show a signifcan positive correlation with citation rates, in confirmation of other studies which have found a significant effect of keyword number on citation rates (Uddin and Khan, 2016)- although the results of the two studies cannot be directly numerically compared due to differences in data preparation and analysis techniques.

In both cases where a multivariate model was run to compare the effects (Models A6 and B4) the *Qtinfo* measure outperformed the keyword count measures by a large margin, indicating that although the information content measures are correlated with the keyword count measures, their predictive potential is not due entirely, or even predominantly, to the relationship between these measures and the count-based measures.

## 5.10 Potential confounding factors not included in multivariate analysis

Other potential confounding factors are more complicated to control for, and future research may be directed fruitfully in this direction.

Citation rate data is not 100% accurate. Scopus data overall has found to it may also be that there is a systematic relationship between different types of metadata error- poor keyword quality (by this measure) may be correlated with other metadata problems- further investigation of variability of *Qtinfo* values between and within subjects would certainly be of benefit. For example, one possible explanation consistent with the current results is that higher *mean* outgoing citation counts (ie number of citations made by each are higher simply commonplace in subject areas where a proliferation of highly specific terminology means that *Qtinfo* scores tend to be higher – although there is some evidence that this is not, by and large, the case and that outgoing citation rates are relatively constant across disciplines. (Marx and Bornmann, 2015).

Equally, it is possible that there is a systematic relationship between the subject-based differences in terminology potentially resulting in *Qtinfo* score variation and the degree to which citations are efficiently picked up in Scopus- STEM citations are picked up much more reliably and efficiently than humanities citations for example, and it may also be the case that the controlled, specific vocabulary of STEM means that highly discriminatory keywords are relatively easy to attach to papers- it is notable for example that . Investigation into the variations in *Qtinfo* scores between disciplines and further more focused analysis concentrating on a selection of papers more narrowly selected than those in the current study (perhaps from a selection of specific journals, which would also have the effect of making testing for correlation effects with Journal Impact Factor more straightforward).

Further multivariate analysis would help to control for these other bibliometric measures, although even if (for example in the case of Journal Impact Factor) the results are found to be significantly correlated with *Qtinfo* this may simply be indicative of e.g. higher standards of metadata applied to more prominent journals.

Conducting a more comprehensive multivariate analysis, feeding in other parameters known to affect citation rates abstract length, author eminence, number of keywords etc) would provide considerable insight into the place of *Qtinfo* among these other bibliometric measures.

Finally, there is one potential and very important confounding factor for which it may, happily, be relatively easy to control due to the structure of the keyword fields. It may be that there is no causal relationship at all between higher Author Keyword quality and citation rates due to differences in retrieval effectiveness, but that both may simply be the common result of more assiduous researchers- that is, researchers who produce higher-quality research also tend to invest more time and effort into their keyword selection! This potential confounding factor however does have the advantage of being relatively easy to investigate with the help of index keyword data- if a positive correlation is found between Author Keyword quality and citation rates, but no such relationship (or a weaker one) is found for the more numerous Index Keywords, then we might conclude that the results found above are simply the result of more assiduous authors!

# 6 CONCLUSION

It is clear that the results presented here can represent no more than the barest start toward the investigation of the utility of statistical metadata quality measures on citation rates. However, as imperfect as the results are, it does however at least represent a first step toward a statistical quantification of the positive effect that the quality of metadata has on academic discourse, as measured by citation rates- quantifying the full extent to which this occurs is far beyond the scope of this exercise, due to the highly imperfect nature of the investigation carried out, but as discussed in the 'Future Directions' section following this, there are numerous avenues by which this analysis could be considerably improved and perhaps begin to start answering quantitatively, as well as providing some degree of objective measure pointing toward a means by which policy-makers can justify investing resources into improving metadata quality, which is often under-resourced in part, I believe, precisely because it is so difficult to quantify.

*Hypothesis 1:* Higher-quality (more descriptive) keywords enable articles to be found more efficiently, both by enabling articles to be retrieved, and by boosting the position of retrieved results in search rankings when many papers are retrieved.

*Hypothesis 2:* Higher-quality keywords results in more efficient discovery of articles of interest, and therefore result in higher citation rates for those articles.

***Hypothesis 3:*** If higher quality keywords result in more citations, then a positive correlation will be found between *Qtinfo* and citation rates.

It is clear that, although encouraging, a great deal more work is necessary to demonstrate a *causal* relationship between keyword quality and citation rates- the current statistical measures of

metadata, as discussed in section 7.7, do not have straightforward relationship to 'real' keyword quality. Furthermore, it would be premature to consider a causal relationship on the basis of an initial, exploratory study- although these results are very much in line with what one would expect to find if higher keyword quality were to have a positive effect on citation rates. However, the results obtained here are certainly not inconsistent with any of these hypotheses and while they do not definitively answer any of the hypotheses, may be considered  a positive first step.

# 7 FUTURE DIRECTIONS

The analysis presented here is by necessity crude, limited and cursory- that there is a link between There are a great number of potential directions for improvement of this research- the $Qtinfo_{kw}$ parameter is based on a relatively crude sum of disparate N-gram forms of *tf-idf* weightings and it is by no means apparent that such an *ad hoc* weighting scheme is the best possible means of deriving a measure of keyword quality. Given the *de facto* effectiveness of this crude measure it is not clear that its inelegance (and inherited lack of theoretical foundation) necessarily counts against it for the current purposes0, but it does imply that better measures have yet to be discovered.

## 7.1 LARGER DATASETS

The most obvious initial avenue for improvement is simply to run the analysis again with much larger datasets obtained by the same means- the API return limits of 20,000 records per week applied by the Scopus API, combined with the relatively high attrition rare imposed by the fact that keywords were not present for many of the papers reviewed means that only a relatively small number of records were analysed. An analysis of greater breadth would be time-consuming in terms of API download limits, but not any more onerous than the one undertaken here, although RAM limits may have to be taken into account. A larger sample set would also allow

## 7.2 ADOPTION OF MORE SUITABLE TOOLS FOR DATA STORAGE AND ANALYSIS

One very significant limitation of the current study was the limitation of both the sample size and number of N-grams which could be accommodated within the 8GB of RAM available on the computer used for analysis. A system based on a relational database, allowing storage and manipulation of much larger file sizes, would be a considerable improvement, at least for the data preparation portions of the analysis involving Tidy text tables (which are perfectly suited to this type of analysis)- once prepared, the final datasets are on the order of hundreds of megabytes which is perfectly tractable, so final analysis could easily be performed if the final data sets were to be reimported to R.

## 7.3 INCLUSION OF OTHER KEYWORD TYPES

Many other questions which present themselves depend on the question being asked- if the

question is whether quality of *Author* keywords have a large effect on citation rates, the answer appears to be 'no'- the more interesting question however is perhaps 'What is the total contribution of keyword quality to citation rates?'

In order to answer this question in a semi-convincing manner we need to try to comprehensively retrieve and  It is clear that the number of keywords applied to papers significantly outstrips those which were harvested under the current analysis- the Scopus abstract metadata supplies keywords in a number of different places, and only the most obvious of these were used for the current analysis- however, from manual inspection of search results retrieved from the Scopus web interface it is apparent that the keywords harvested in the current work form at best minority of the total attached to an article even the current, rather crude mode of analysis would presumably be considerably improved by harvesting a greater proportion of keywords, combining results from multiple citation indices and further investigating the output from API results to determine if any further index keywords can be extracted from the data returned,  to say nothing of other related metadata which may also provide information content- for example subject classification, as well as controlled vocabulary systems which fall somewhere in between, such as MeSH terms.
Assuming that this and related measures can be developed into a meaningful and reliable indicator of metadata quality, another possible line of inquiry is the relative contribution to successful retrival (and hence citation rates) made by various different forms of structured metadata versus unstructured keywords.

## 7.4 IMPROVED CAPTURE OF MEANINGFUL KEYWORDS

The basis for the calculation of the measure bears examination and there are in this regard several potential avenues for developing and testing other measures of metadata quality in this regard- indeed, it could even be argued that a measure of metadata quality which explained a larger proportion of the variance in citation rates has some right to be named a superior measure of metadata quality! Other measures of information quality (Shirakawa,Hara and Nishio, 2017)which apply across different N-gram lengths have been proposed, and adapting the research in order to calculate and use these values instead, and to develop a measure of keyword quality based on these measures instead, may provide fruitful avenues by which a larger proportion of the (presumably) much larger proportion of citation rate variance accounted for by keyword quality can be properly captured.

Finding measure of citation rates which better approximate the 'natural' distribution of keyword-search retrieved articles is another potentially fruitful avenue of investigations. There have been attempts to produce normalised citation rates enabling comparison of citation rates across years (Uddin *et al.*, 2012), but the measure of interest when investigating the effect of keywords may be more suited to a measure which attempts to strip out the effects of high 'social' visibility of papers and approximate better the 'background' rate of keyword-based retrieval, whether it is possible to estimate this rate statistically using the information currently publicly available from citation indexes or whether this type of measure would require more detailed research into researchers' information-seeking behaviours.

Conversely, research into what types of keywords are most associated with higher citation rates is an avenue which is potentially very fruitful- given the high level of institutional interest in citation rates, the potential for this type of research to help develop automatic measures for validating and quantifying metadata quality in, for example, Institutional Repositories could potentially result in

very useful applications- if keyword quality, as appears to be the case, accounts for several per cent overall in citation rate variance, then the overall results when considered from the point of view of an institution's entire body of research output over the long term could be considerable.

## 7.5 IMPROVED SAMPLING METHODOLOGY

Similarly, the restrictions imposed by using a single word search to restrict the corpus is rather arbitrary and more than likely does not accurately reflect the real 'corpus' from which documents are selected- it is difficult to specify exactly what the appropriate corpus is in a situation such as this, indeed- is the entire body of work covered by the citation index to be considered, or simply a more narrow subsection of documents- particular journals for example, or articles covering particular subject matter?

The danger of course in restricting things too narrowly in this manner is that because one is analysing metadata, to be *too* reliant on metadata in selecting a sample runs a risk of skewing a sample in a particular way- it is not clear what the effect of this might be. Future research into the effects of both restricting the corpus to a more tightly controlled set of documents, and of broadening it out to more closely approximate the effect of a 'in the wild' keyword search of a citation index such as Scopus (ie to determine the discriminatory

## 7.6 OTHER MEASURES OF INFORMATION QUALITY

Intuitively, there should be a very clear and substantial value in providing keywords which *do not* appear as tokens in the text- for example semantically related or synonymous terms applied as keywords may provide valuable avenues for search-based discovery of documents but would simply not be picked up at all by the type of analysis applied here. Finding means of characterising and measuring the efficacy of these keywords, especially in the case of academic terms of art where the same token may have entirely different meanings in different academic fields- non-manually is a particularly challenging task- whether semantic ontologies potentially provide a way forward in this respect is perhaps a future direction rather far down the road, but one which is worth considering nonetheless.

Equally it is quite apparent that maximising the Q*tinfo* statistic and maximising the metadata quality to which it seems to be correlated are two separate items. Further, more distant research directions might involve using citation index and database controlled vocabularies to construct lists of synonymous or related terms, using an ontology or manual approach to substituting for synonymous terms may be of use in this situation.

The approach taken here also fails to capture the information content of any keywords attached to the document which do not appear in exactly the same string format in the document. Adopting a query-term proximity-based analysis such as graph-of-word approaches (Rousseau and Vazirgiannis, 2013) would allow a greater number of high-quality search terms to be picked up and once again potentially allow the model to explain a larger proportion of search term variance.

# 8 REFERENCES


Anderson, D.L., Smart, W. and Tressler, J. (2013) 'Evaluating Research - Peer Review Team Assessment and Journal-based Bibliographic Measures: New Zealand PBRF Research Output Scores in 2006.'. *New Zealand Economic Papers,* 47 (2), pp. 140.

Australian Research Council (2017) *Excellence for Reseach in Australia.* Available at: http://www.arc.gov.au/excellence-research-australia (Accessed: 05 August).

Bornmann, L. and Mutz, R. (2015) 'Growth Rates of Modern Science: a Bibliometric Analysis Based on the Number of Publications and Cited References'. *Journal of the Association for Information Science and Technology,* 66 (11), pp. 2215-2222.

Brzezinski, M. (2015) 'Power Laws in Citation Distributions: Evidence from Scopus'. *Scientometrics,* 103 (1), pp. 213-228.

De Groote, S.L. and Raszewski, R. (2012) 'Coverage of Google Scholar, Scopus, and Web of Science: a Case Study of the h-Index in Nursing'. *Nursing Outlook,* 60 (6), pp. 391-400.

Duffy, R.D. *et al.* (2008) 'Measuring Individual Research Productivity: A Review and Development of the Integrated Research Productivity Index'. *Journal of Counseling Psychology,* 55 (4), pp. 518-527.

Egghe, L. (2006) 'Theory and Practise of the g-Index'. *Scientometrics,* 69 (1), pp. 131-152.

Ellis, D., Cox, D. and Hall, K. (1993) 'A Comparison of the Information Seeking Patterns of Researchers in the Physical and Social Sciences'. *Journal of Documentation,* 49 (4), pp. 356-369.

Elsevier B.V. (2016) *Scopus Content Coverage Guide.* Available at: https://www.elsevier.com/__data/assets/pdf_file/0007/69451/scopus_content_coverage_guide.pdf (Accessed: 08 August).

European Commission (2014) *Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report From the Expert Group and the Need for a Science-Friendly EU Copyright Reform.* Available at: http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf.

Falagas, M.E. *et al.* (2013) 'The Impact of Article Length on the Number of Future Citations: a Bibliometric Analysis of General Medicine Journals'. *PLoS ONE,* 8 (2), pp. e49476.

Garfield, E. (1964) 'Science Citation Index-A New Dimension in Indexing'. *Science,* 144 (3619), pp. 649-654.

Garfield, E. (2006) 'The History and Meaning of the Journal Impact Factor'. *JAMA,* 295 (1), pp. 90-93.

Gavrilis, D. *et al.* (2015) '*Measuring Quality in Metadata Repositories*'.

Gazni, A. (2011) 'Are the Abstracts of High Impact Articles More Readable? Investigating the Evidence from Top Research Institutions in the World'. *Journal of Information Science,* 37 (3), pp. 273-281.

Gilbert, G.N. (1977) 'Referencing as Persuasion'. *Social Studies of Science,* 7 (1), pp. 113-122.

Gil-Leiva, I. and Alonso-Arroyo, A. (2007) 'Keywords Given by Authors of Scientific Articles in Database Descriptors'. *Journal of the American Society for Information Science and Technology,* 58 (8), pp. 1175-1187.

Handel, M.J.P. (2014) '*Article-Level Metrics—It's Not Just About Citations*'.

Harzing, A.-W. and Alakangas, S. (2016) 'Google Scholar, Scopus and the Web of Science: a Longitudinal and Cross-Disciplinary Comparison'. *Scientometrics,* 106 (2), pp. 787-804.
Haslam, N. *et al.* (2008) 'What Makes an Article Influential? Predicting Impact in Social and Personality Psychology'. *Scientometrics,* 76 (1), pp. 169-185.
Hemminger, B.M. *et al.* (2007) 'Information Seeking Behavior of Academic Scientists'. *Journal of the American Society for Information Science and Technology,* 58 (14), pp. 2205-2225.
Hillmann, D.I., Westbrooks, E.L. and American Library, A. (2004) *Metadata in Practice.* Chicago: Chicago : American Library Association.
Hirsch, J.E. (2005) 'An Index to Quantify an Individual's Scientific Research Output'. *Proceedings of the National Academy of Sciences of the United States of America,* 102 (46), pp. 16569-16572.
HM Government (2014) *Research Excellence Framework- Citation Data.* Available at: http://www.ref.ac.uk/about/guidance/citationdata/ (Accessed: 05/08).
Hughes, B. (2004) 'Metadata Quality Evaluation: Experience from the Open Language Archives Community'. *Digital Libraries: International Collaboration And Cross-Fertilization,* 3334 320-329.
Inácio B., Ferreira J.D., Couto F.M. (2017) 'Metadata Analyser: Measuring Metadata Quality'. *11th International Conference on Practical Applications of Computational Biology & Bioinformatics,* 616 197-204.
Katchanov, Y. (2015) 'Towards a simple mathematical theory of citation distributions'. *SpringerPlus,* 4 (1), pp. 1-14.
Knoth, P. and Pontika, N. (2016) 'Aggregating Research Papers from Publishers' Systems to Support Text and Data Mining: Deliberate Lack of Interoperability or Not?'. *INTEROP2016.* Available at: http://oro.open.ac.uk/46870/.
Letchford, A., Moat, H.S. and Preis, T. (2015) 'The Advantage of Short Paper Titles'. *Royal Society Open Science,* 2 (8), pp. 150266.
Letchford, A., Preis, T. and Moat, H.S. (2016) 'The Advantage of Simple Paper Abstracts'. *Journal of Informetrics,* 10 (1), pp. 1-8.
Low, W., Wilson, P. and Thelwall, M. (2016) 'Stopped sum models and proposed variants for citation data'. *Scientometrics,* 107 (2), pp. 369-384.
Lu, K. and Kipp, M.E.I. (2014) 'Understanding the Retrieval Effectiveness of Collaborative Tags and Author Keywords in Different Retrieval Environments: an Experimental Study on Medical Collections'. *Journal of the Association for Information Science and Technology,* 65 (3), pp. 483-500.
Margaritopoulos, M. *et al.* (2012) 'Quantifying and Measuring Metadata Completeness'. *Journal of the American Society for Information Science and Technology,* 63 (4), pp. 724-737.
Martin-Martin, A. *et al.* (2017) 'Can We Use Google Scholar to Identify Highly-Cited Documents?'. *Journal of Informetrics,* 11 (1), pp. 152-163.
Marx, W. and Bornmann, L. (2015) 'On the Causes of Subject-Specific Citation Rates in Web of Science'. *Scientometrics,* 102 (2), pp. 1823-1827.
Medoff, M.H. (2006) 'Evidence of a Harvard and Chicago Matthew Effect'. *Journal of Economic Methodology,* 13 (4), pp. 485-506.
Moed, H.F. (2005) *Citation Analysis in Research Evaluation.* Dordrecht ; [Great Britain]: Dordrecht ; Great Britain : Springer.
Mongeon, P. and Paul-Hus, A. (2016) 'The Journal Coverage of Web of Science and Scopus: a Comparative Analysis'. *Scientometrics,* 106 (1), pp. 213-228.
Mryglod, O. *et al.* (2013) 'Comparison of a Citation- Based Indicator and Peer Review for Absolute and Specific Measures of Research- Group Excellence'. *Scientometrics,* 97 (3), pp. 767-777.
Nair, L.B. and Gibbert, M. (2016) 'What Makes a 'Good' Title and (How) Does It Matter for Citations? a Review and General Model of Article Title Attributes in Management Science'. *Scientometrics,* 107 (3), pp. 1331-1359.

Ochoa, X. and Duval, E. (2009) 'Automatic Evaluation of Metadata Quality in Digital Repositories'. *Int J Digit Libr,* 10 (2), pp. 67-91.
Prathap, G., Mini, S. and Nishy, P. (2016) 'Does High Impact Factor Successfully Predict Future Citations? an Analysis Using Peirce's Measure'. *Scientometrics,* 108 (3), pp. 1043-1047.
Rousseau, F. and Vazirgiannis, M. (2013) 'Graph-of-Word and TW-IDF: New Approaches to ad-hoc IR'. *Proceedings of ACM Conference on Information and Knowledge Management (CIKM).*
Samuels, L.E. *et al.* (2012) 'Left Ventricular Assist Device Implantation in High Risk Destination Therapy Patients: an Alternative Surgical Approach'. *Journal of Cardiothoracic Surgery,* 7 (1), pp.
Shirakawa, M., Hara, T. and Nishio, S. (2017) 'IDF for Word N-grams'. *ACM Transactions on Information Systems (TOIS),* 36 (1), pp. 1-38.
Small, H.G. (1978) 'Cited Documents as Concept Symbols'. *Social Studies of Science,* 8 (3), pp. 327-340.
Sohrabi, B. and Iraj, H. (2017) 'The Effect of Keyword Repetition in Abstract and Keyword Frequency per Journal in Predicting Citation Counts'. *Scientometrics,* 110 (1), pp. 243-251.
Sparck-Jones, K. (1972) 'A Statistical Interpretation of Term Specificity and Its Application in Retrieval'. *Journal of Documentation,* 28 (1), pp. 11-21.
Stephen, R. (2004) 'Understanding Inverse Document Frequency: on Theoretical Arguments for Idf'. *Journal of Documentation,* 60 (5), pp. 503-520.
Taylor, J. (2011) 'The Assessment of Research Quality in UK Universities: Peer Review or Metrics?(Report)'. *British Journal of Management,* 22 (2), pp. 202.
Thelwall, M. (2016) 'The Discretised Lognormal and Hooked Power Law Distributions for Complete Citation Data: Best Options for Modelling and Regression'. *Journal of Informetrics,* 10 (2), pp. 336-346.
Tsiflidou, E. and Manouselis, N. (2013) 'Tools and Techniques for Assessing Metadata Quality*'*. In*:* Garoufallou, E. and Greenberg, J. (eds.) *Metadata and Semantics Research: 7th Research Conference, MTSR 2013, Thessaloniki, Greece, November 19-22, 2013. Proceedings.* Cham: Springer International Publishing, pp. 99-110.
Uddin, S. *et al.* (2012) 'Trend and Efficiency Analysis of Co- Authorship Network'. *Scientometrics,* 90 (2), pp. 687-699.
Uddin, S. and Khan, A. (2016) 'The Impact of Author-Selected Keywords on Citation Counts'. *Journal of Informetrics,* 10 (4), pp. 1166-1177.
van Raan, A. (2004) 'Sleeping Beauties in Science'. *Scientometrics,* 59 (3), pp. 467-472.
Ward, J. (2003) '*A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage Within Data Providers Registered with the Open Archives Initiative'*. USA.
Wesel, M., Wyatt, S. and Haaf, J. (2014) 'What a Difference a Colon Makes: How Superficial Factors Influence Subsequent Citation'. *Scientometrics,* 98 (3), pp. 1601-1615.
Windnagel, A. (2014) 'The Usage of Simple Dublin Core Metadata in Digital Math and Science Repositories'. *Journal of Library Metadata,* 14 (2), pp. 77-102.