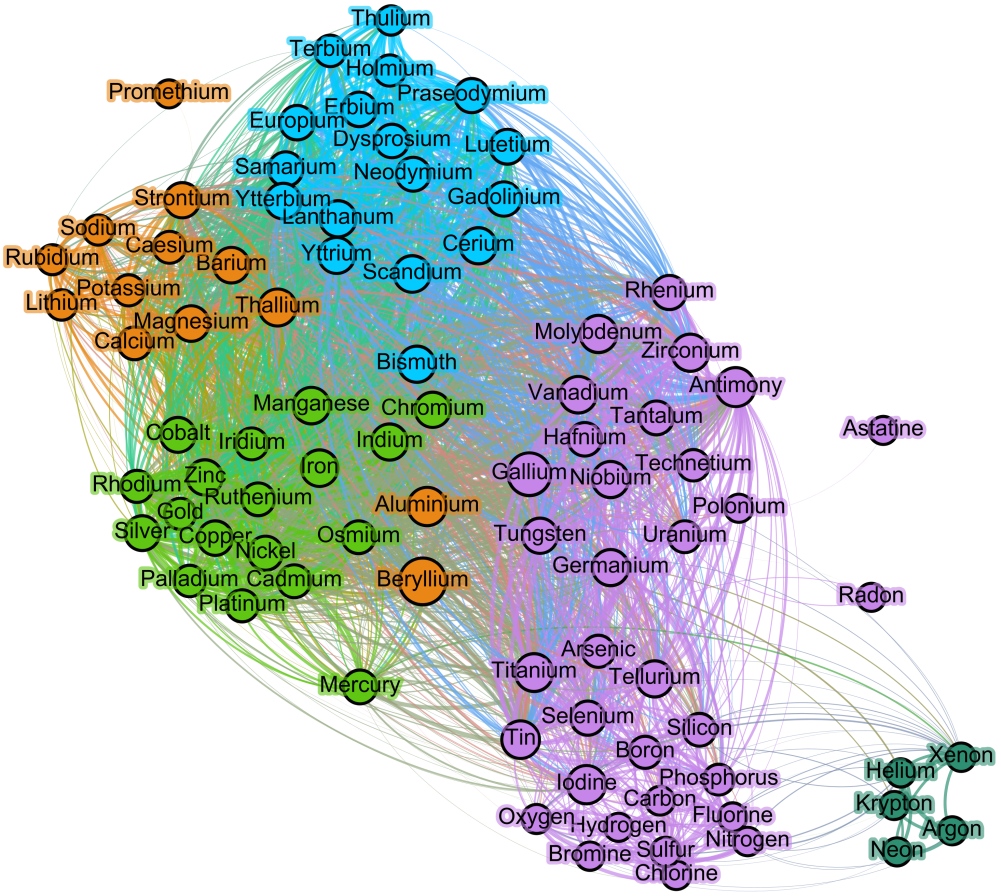
Part III Project

Automatic Chemical Analysis - Big Data and Machine Learning in Online Chemistry Literature



Patrick Lewis

Queens' College

March 2016

This dissertation is submitted in partial fulfilment of the requirements for Part III Chemistry. It describes work carried out in the Department of Chemistry in the Michaelmas Term 2015 and the Lent Term 2016. Unless otherwise indicated, the research described is my own and not the product of collaboration.

Signed:

Date:

Acknowledgements

Acknowledgements will go here

Abstract

A Large Dataset of Chemistry Literature meta-data was built up by automated scraping from freely-available online sources. A UK Chemistry Department dataset of chemical literature meta-data was built up by a similar method. Novel Natural Language Processing algorithms were used to develop powerful models to represent the chemical semantic space. These models were analysed and visualisation techniques were developed. The utility of the models was demonstrated by investigating relationships between researchers at the University of Cambridge Chemistry Department.

Contents

List of Figures

Glossary

$\Delta 1$ Database of article meta-data created from DOIs found on UK Chemistry department websites.

$\Delta 2$ Database of complete article meta-data including abstracts found on UK Chemistry department websites.

$\Delta 3$ Database of article meta-data created from DOIs found in global scraping procedure.

$\Delta 4$ Database of complete article meta-data including abstracts found in global scraping procedure.

$\Delta 5$ Database created by combining databases $\Delta 2$ and $\Delta 4$.

$\Delta 6$ Database comprised of records in database $\Delta 5$ deemed suitable for machine learning, i.e. sufficiently long titles and abstracts and predominantly ascii characters.

$\Delta 7$ Database comprised of records in Database $\Delta 6$ which had originated from DOIs found on the Cambridge University Chemistry Department.

ACS American Chemical Society. A Scientific society specialising in the chemistry domain with a large scientific academic publishing arm. The ACS also own the Chemical Abstracts Service and SciFinder®.

API Application Programmer Interface. Set of well defined input and output operations to a program or service to enable programmers to easily use the service.

Bag Of Citations Simple Document Representation model which attempts to represent a document based on presence/absence of citations.

Bag Of Words Simple Document Representation model which attempts to represent a document as a vector based on presence/absence of words.

CBOW Continuous Bag Of Words. Learning architecture used by Word2Vec algorithm. Word Vector predictions are made from sum or mean of surrounding context words.

Cluster Map A heat-map of two-dimensional data with axes arranged by a hierarchical clustering algorithm, often overlaid with dendrograms along each axes to illustrate clustering and spatial relationships [1] [2] [3].

Communities Subset of documents in a corpus identified via the Blondel-Guillaume-Lambiotte-Lefebvre modularity algorithm [4][5].

Corpus In the field of language processing, a corpus is a large body of natural language text. In the context of the project, a corpus is the combined titles and abstracts of all the records in a database.

Cosine Similarity A similarity metric for vectors derived from the angle between them. $S_{cosine} = cos\,(\theta)$ for angle θ between two vectors.

Crawling Programming technique to automatically navigate through the online landscape identifying candidate websites for scraping (See Scraping).

Crossref Crossref is an organisation promoting inter-publisher cooperation. With a mission statement to 'support ... persistent, sustainable infrastructure for scholarly communication'[6]. Crossref provides tools for researchers to access a wide range of publishers' materials.

Dendrogram Tree diagram used to illustrate relationships between clusters produced in hierarchical clustering procedures [1].

Doc2Vec Gensim implementation of Paragraph Vectors algorithm (See Paragraph Vectors).

DOI Digital Object Identifier. Unique identifier string used to index the vast majority of academic literature articles published since 2000 [7].

Euclidean Similarity A similarity metric for vectors derived from their distance in Euclidean space. $S_{euclid} = \sqrt{\sum_{i=1}^{D}(\nu_i^{\alpha} - \nu_i^{\beta})^2}$ for vectors α and β.

Gensim Gensim is an open-source library for Python programs for use in NLP applications.

Gephi An open-source network visualisation, rendering and analysis application.

Hadamard Division Element-wise division matrix operation defined as $(\mathbf{A} \oslash \mathbf{B})_{i,j} = \mathbf{A}_i / \mathbf{B}_j$ for matrices $\mathbf{A}$ and $\mathbf{B}$.

Hadamard Square Root Element-wise square root matrix operation defined as $\left(A^{\circ\frac{1}{2}}\right)_{i,j} = A_{i,j}^{\frac{1}{2}}$ for matrix A.

HTML Hypertext Markup Language. Tag-based language to encode web-pages in a hierarchical structure. Webpages are written as HTML files, which are interpreted by internet browsers to display the page's content to users.

Hyperparameters Adjustable Parameters used by a Machine Learning algorithm. Different to internal parameters automatically learnt by the algorithm.

IDF International Doi Foundation. Independent not-for-profit body which governs the use and management of the DOI system. Provide definitive service for resolving a DOI to its document [8].

IP Address Internet Protocol Address. An IP address is the identifier for any computer or device using a network that runs on Internet Protocol.

Lancaster Stemming algorithm [9].

Machine Learning Field of computer science with the aim of developing algorithms that automatically improve performance based on supplied examples [10].

Meta-data Meta-data refers to data about data. In the context of this project, it refers to data describing a chemistry article, i.e. title, abstract, DOI (See DOI), authors, affiliations, journal, publisher and date of publication.

MongoDB Schema-less 'NoSQL' database used for document storage retrieval.

Neural Net Data structure capable of developing decision pathways using supplied examples.

NLP Natural Language Processing. NLP is the field of linguistics and computer science with the aim of processing human written (natural) language with a computer.

Paragraph Vectors NLP algorithm based on Word2Vec for generating representation vectors for documents.

PCA Principle Component Analysis. Well established technique for reducing dimensionality by a series of orthogonal transformations [11].

PILA Publishers International Linking Association, Inc. . Independent not-for-profit body comprised of scientific prublishing entities. PILA operates Crossref [6].

Porter Early, widely used stemming algorithm [12].

Python Interpreted, dynamically typed programming language. Unless explicitly mentioned, Python was the language used for all development and analysis in this project.

REGEX REGular EXpression. A REGEX is a string that is used to inform a programming language of patterns to identify in a body of text.

RSC Royal Society of Chemistry. British learnèd society for chemical sciences with an academic publishing arm.

SciFinder® Bibliographic and citation search engine provided by the American Chemical Society designed for chemical research.

Scraping Programming technique to automatically extract data from online resources.

Skipgram Learning architecture used by the Word2Vec algorithm. Word Vector predictions are made from random comparison between a word and nearby context words.

Snowball Recent stemming algorithm [13] (Also known as Porter2). Snowball is also the name of the programming language developed for stemming..

Stemmer An algorithm used to relate derived words (e.g. plurals, conjugated verbs) to their roots.

Stop Words Words removed from a corpus before being processed. Stopwords are very common and/or do not encode significant information content.

Taylor & Francis Part of the Informa group. An academic publisher covering a range of scientific disciplines.

TF-IDF Term-Frequency Inverse-Document-Frequency. Method for assigning weights to words in a document for how much information the word carries.

Training Epoch A complete iteration over the training data available to a learning algorithm.

TSNE T-distributed Stochastic Network Embedding. State of the art technique for reducing dimensionality by preserving spatial clusters of vectors at high dimensions [14].

Unicode Standard for encoding characters used in worldwide communication. The Unicode character set of 120,000 characters includes mathematical symbols, punctuation, and character languages (Mandarin, Japanese etc.).

UPGMA Unweighted Pair Group Method with Arithmetic Mean. Pairwise Clustering algorithm that partitions a set into hierarchical sub-set clusters using mean distances between pairs of elements [15].

UTF-8 Universal Coded Character Set + Transformation Format 8-bit [16]. An encoding specification for the Unicode character set, where each character is encoded by 8 bits. UTF-8 is the dominant encoding used online [17].

Web Of Science ™ Bibliographic and citation search engine provided by Thomson Reuters.

Word2Vec Sophisticated distributed word vector model utilising a Neural Net to learn vector representations of component words in a corpus by training sentence by sentence [18][19] .

WordNet Lemmatizing stemming algorithm, based on consulting database of groups of semantically connected word concepts [20] [21] [22].

XML eXtensible Markup Language. Tag-based markup language, closely related to HTML. Method of encoding any type of data in a manner that is machine readable

and intelligible by humans.

XPath Query method for extracting data from XML documents. As HTML is closely related to XML, XPath strings can be used to access data in HTML documents.

Zipf's Law States that the relationship of the log of the rank of a word to the log of the frequency of that word in a large corpus of text approximates a directly proportional relationship [23].

1. Appendix

1.1 Recommendations for Further Work

As alluded to in the text, there are several recommendations for further work. The code and data will be improved and amended over time, and is freely available under MIT licence on request [1]. If attempting to carry out further work on this project, it is recommended to contact the author for in-depth explanations. This list is by no means exhaustive, and it is the author's belief that literature semantic analysis should be considered an important analytical chemical tool.

1.1.1 Greater Dimensionality and Training Improvements

The principles behind the methods discussed in the project have been shown to be sound. Models should now be improved. Computing resources should be obtained to train higher dimensional vectors [2]. The models should also be trained for longer (> 24) epochs on more data (> 460000 documents). These steps will lead to more expressive models.

1.1.2 Greater use of word vectors

This project focussed mainly on document vectors. However, word vectors may be very useful. A method for testing the quality of improved models should be developed. This could take the form of expected relationships to test the model: e.g. Fluorine is to Fluoride as Chlorine is to ... . Many hundreds of these relationships should be systematically built up to test model intuition.[3] This follows the methodologies set out in the literature [18] [19]. Furthermore, is it possible to predict chemical properties using semantic relationships found in the literature? Vec(Compound A) + Vec(Compound B) + Vec(Lab Technique) may give vec(Product C). If so, it may be possible to find

[1] A digital copy is included with this dissertation.

[2] The author recommends 400 dimensional vectors

[3] This would probably require much larger, more descriptive training sets, e.g. textbook transcipts etc.

unexpected reactions. This could be coupled with the RInChI database to form a new type of data-driven cheminformatics.

1.1.3 Time resolution in clustering

Methods have been described for clustering documents. The cluster centres represent the content of the cluster effectively. By finding early papers in the cluster, is it possible to identify influential papers or authors? By clustering on documents from particular years, is it possible to identify a path for the evolving cluster centre vector? If so, it should be possible to extrapolate to *predict* near future research directions.

1.1.4 Open Source Chemistry Vectors

With the increase in open source papers, it should be possible to build up a vast dataset of chemical language for training, using the bodies of articles published on open source platforms, and even to use supplied supporting information.

1.1.5 Structure stemming

Chemical names could be smartly preprocessed to classes of chemicals, for example by identifying a compound from its name and mapping to InChI key, then to a chemical class. This would allow better association of chemical fragments in training.

1.1.6 Multiply labelled Documents

In Training Doc2Vec, by specifying document with more than just their unique identifiers allows more vectors to be associated. By identifying and labelling all documents with a particular concept, e.g. 'palladium-catalysed', and then training Doc2Vec, one defines an 'palladium-catalysed' vector, specifically trained for the concept. These concept vectors would be robust and information-rich[4]

[4]e.g. which documents are close to the indium-catalysed vector but do not contain the word indium...

1.2 Technical Details

In the interest of future work, this section details the technical details of artefacts provided with this project.

1.2.1 Code Artefacts

The python code used in this project was written in a largely self-documenting style. The time limits did not permit for professional doc-strings to be produced, or for anaconda packages to be provided, but the code is well commented. There is also a comprehensive set of `Jupyter Notebooks` as tutorial guides for using the code[37]. The core code has been presented in a 'package' style. The module was named `fruitbowl` with five submodules,

- `Cherry` for operations concerning scraping and data collection.
- `Orange` for operations concerning NLP corpus creation and big data memory-friendly streaming
- `Strawberry` for operations concerning Word2Vec and Doc2Vec model Training
- `Apple` for operations concerning analysis of trained models (visualisation, export management etc.)
- `Pomegranate` for operations interfacing with Gephi and community generation.

There are approximately 30 python source files included in the module. If using the code it is recommended to read and adapt the jupyter notebooks `Fruitbowl Example 1.ipynb` to `Fruitbowl Example 3.ipynb`. It is recommended to write code in a directory that contains the `fruitbowl` module. The Module is free to distribute and adapt under the MIT licence, which must be included in any copy. The list of dependencies required for fully functional behaviour for the `fruitbowl` suite is as follows:

- `Python 2.7` Developed on Python 2.7.11 (recommended version)
- Python 2 external modules required:
  - `matplotlib 1.5.1` Plotting modules [38]
  - `Seaborn 0.7.0` Extension to plotting modules and data analysis [2]
  - `numpy 1.10.4` Computational Library [39]
  - `Scikit-Learn 0.17` Machine learning library [35]
  - `Scrapy 1.0.3` Scraping framework
  - `Gensim 0.12.2` Natural Langauge Processing library [29]
  - `nltk 3.1` Natural Language ToolKit library [31]
  - `pandas 0.17.1` Data analysis and management library [40]
  - `pymongo 3.0.3` Python driver for MongoDB database
  - `requests 2.9.1` Web scraping library
  - `scipy 0.17.0` Scientific computing library [3]

  - `jupyter 1.0.0` Jupyter notebooks will be required to use the tutorial notebooks.
- `JDK` Java Development Kit - for Gephi graph analysis via gephi api
- `apache-maven-3.3.9` Java dependency manager - for Gephi graph analysis via gephi api
- `C Compiler` for use in BHTSNE reductions[34].
- `mongoDB` The program was built around use of MongoDB. Not strictly necessary but strongly recommended. Recommended versions >3.2.

1.2.2 Data Artefacts

Data used in the project was dumped from their mongo databases is also supplied in .json format. The data provided is as follows:

- `Delta1.json` : These are the DOIs found in the UK scrape
- `Delta2.json` : These are the complete meta-data results found in the UK scrape
- `Delta3.json` : These are the DOIs found in the global scrape
- `Delta4.json` : These are the complete meta-data results found in the global scrape
- `Delta6.json` : This is the data used for training and analytical purposes in the project
- `Delta7.json` : The subset of $\Delta 6$ from Cambridge used in §**??**
- `cbow_model` : Gensim binary saved model for final cbow Word2Vec model used in the project
- `sg_model` : Gensim binary saved model for final skipgram Word2Vec model used in the project
- `FULL_DOC2VEC` : Gensim binary saved model for final Doc2Vec model used in the project

Note $\Delta 5$ is not provided to save disk space (It is simply $\Delta 2$ combined with $\Delta 4$).

1.3 Word Vecotr Analysis of Chemical Elements

As an investigation into the utility of word vectors trained, it was decided to briefly investigate the word similarities between chemical elements. This analysis is included as an appendix as there was not sufficient space for it to be included in the main body, and due to its self-contained nature. It was hoped that this short investigation would provided evidence that methods sketched in §1.1.2 could work.

The similarity matrix was produced for chemical elements mentioned in the text corpus (115 out of 118 known chemical elements). This required mapping both chemical names and symbols together (e.g. `florin`[5] and `F` for fluorine) to represent the concept vector for the elements in question.

A modified data sanitation pipeline was created to substitute the chemical symbol for the chemical name. This was only done for chemical symbols longer than 1 letter to dissuade conflating different concepts to the same word vector (`S` could represent Sulfur or a stereochemical label.)

A CBOW model was trained using this modified input data with the same presets as the main CBOW model, detailed in table **??**. The Cosine Similarity matrix was produced for the 115 elements found in the corpus. UPGMA clustering was performed[35], as well as graph visualisation with modularity clustering [4],[5]. The dendrogram of the UPGMA clustering is shown in figure 1.1. The process identified 5 main branches:

- The gold region includes a sub-branch of noble gases, the other branch mainly actinoids.
- The magenta region contains non-metals mostly associated with organic compounds
- The cyan region contains mainly metalloids, actinoids and alkali metals
- the red region contains mainly transition metals
- The green region contains almost exclusively lanthanoid metals

[5] Fluorine is stemmed to fluorin by the stemming process (§**??**)

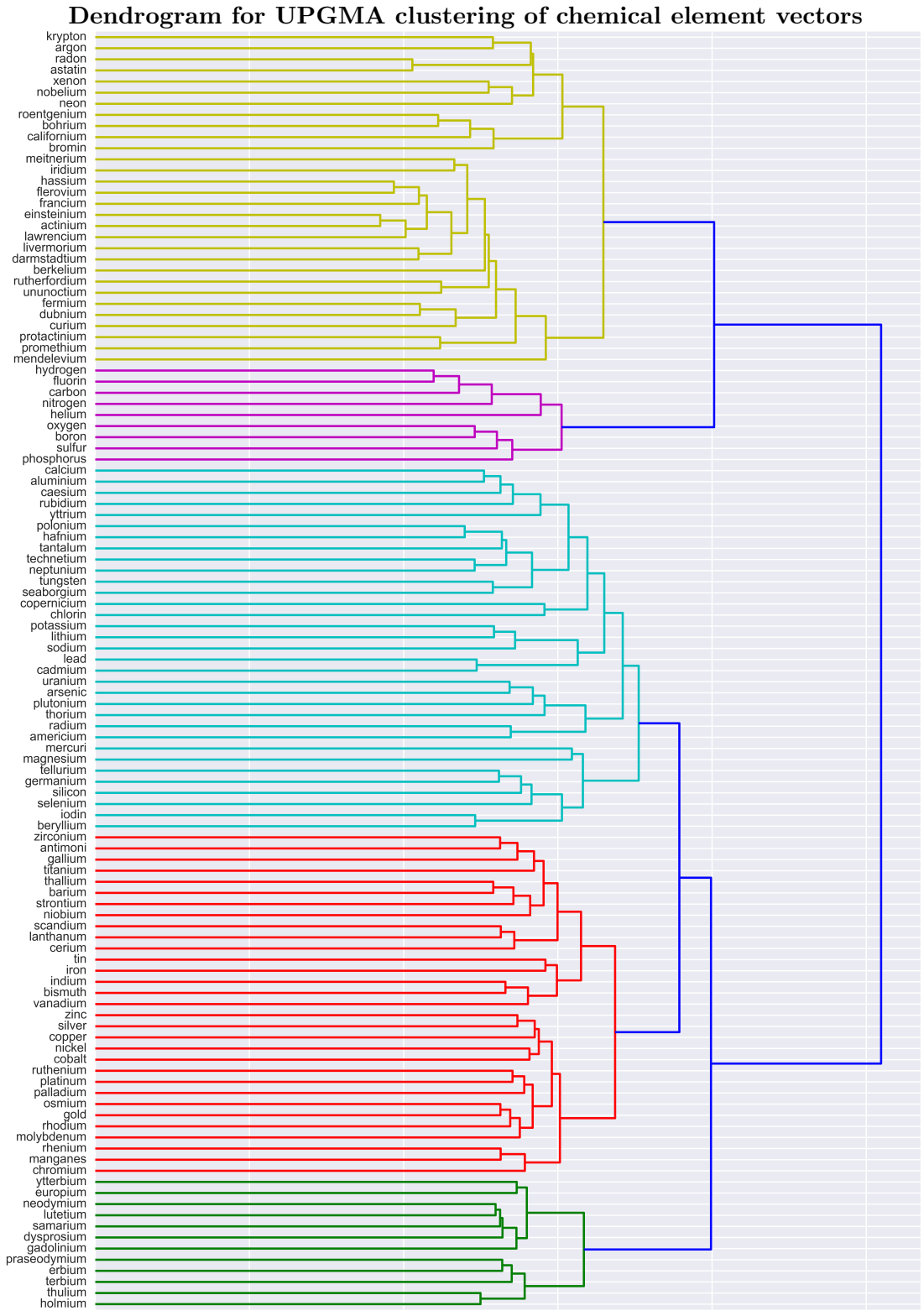


Figure 1.1: Dendrogram for UPGMA clustering of chemical element vectors. Colours indicate distinct branches

The dendrogram shows that the classifications broadly fall into intelligible categories within the periodic table. There are, however, some surprises, especially the halogens, with bromine in the actinoid subbranch, and chlorine associating with copernicium. This may be because the symbols Cl and Cn occur together often in the literature due to mentions of carbon and nitrogen, not copernicium. Similar reasoning can be used for bromine (`cf br` could refer to a CFC rather than californium and bromine) This exposes a flaw in the symbol/name association process that could be tackled in further work.

The graph visualisation is shown in figure 1.2 (Also the front cover of this dissertation). Period 7 was removed from this graph as there were too few mentions in the corpus for reliable vectors[6], and to remove cluttering nodes.

[6]Uranium was kept as it had non-negligible corpus mentions

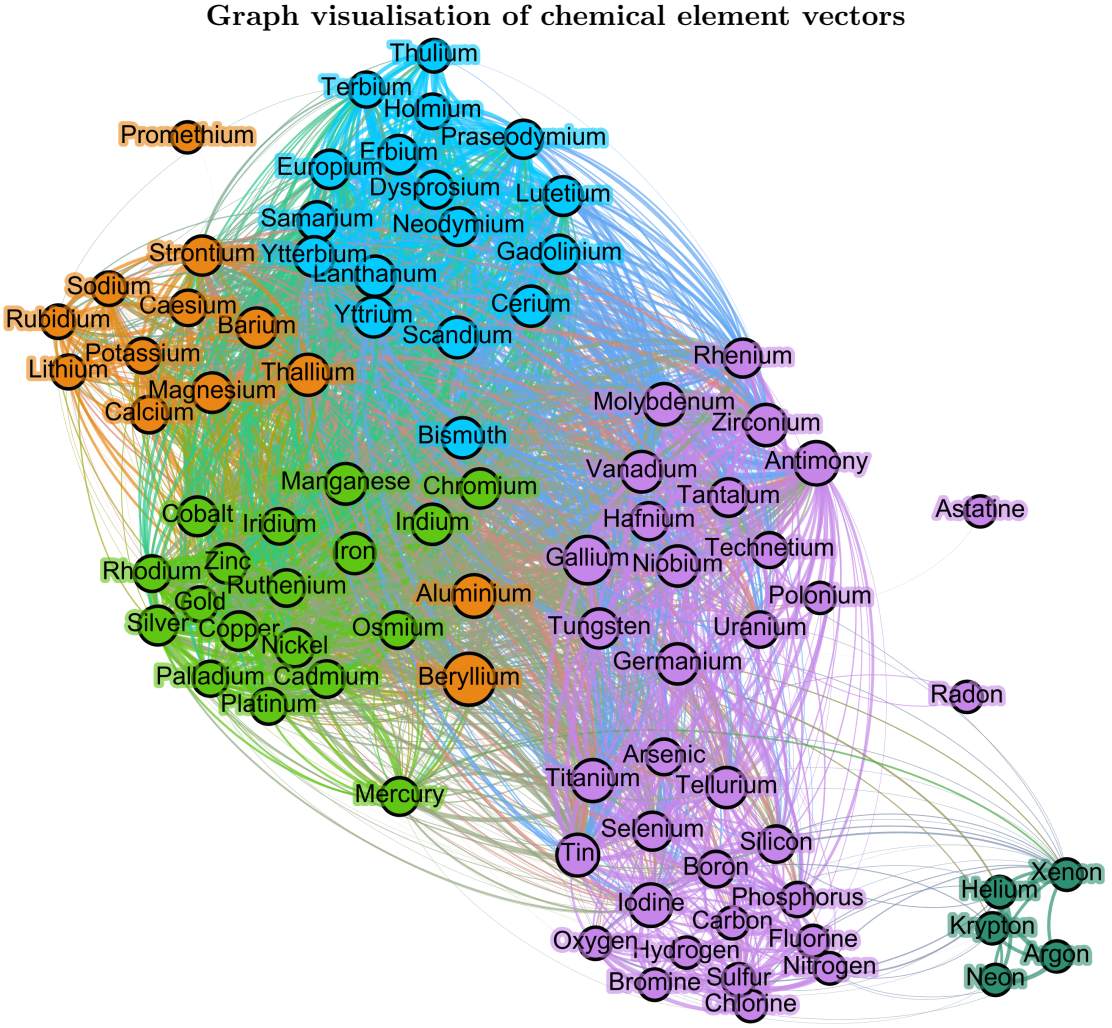


Figure 1.2: Nodes are coloured by their communities, and are spatially arranged by modelling edge weights as springs. Node sizes are proportional to their connectivity.

5 distinct communities were identified:

- The Orange community contained mainly alkali metals and alkali earth metals
- The Blue community contained mainly lanthanoids
- The Light Green community contained mainly transition metals
- The Purple community separated into two spatially distinct regions. The northern region generally contained transition metals and metalloid, and the southern region contained organic non-metals.

- The Dark Green community contained noble gases.

The community finding process reflects a similar situation to the UPGMA process, but is perhaps more successful. The removal of the actinoids appears to have improved community finding. The community finding process was repeated with period 7 included, resulting in broadly the same communities, but with bromine and chlorine leaving the purple community to join a a loose community of actinoids, however they remained strongly associated with each other. The degree of connectivity between nodes is similar for most nodes, but larger for some nodes, e.g. beryllium, which is difficult to interpret.

Attention was turned to a practical example. Palladium is used widely in catalysis but is rare and expensive, and alternatives would be economically and environmentally beneficial[**?**]. With this in mind, the cosine similarities of all the elements to palladium vector were computed, and a selection of metallic elements with high similarity is shown in figure 1.3.

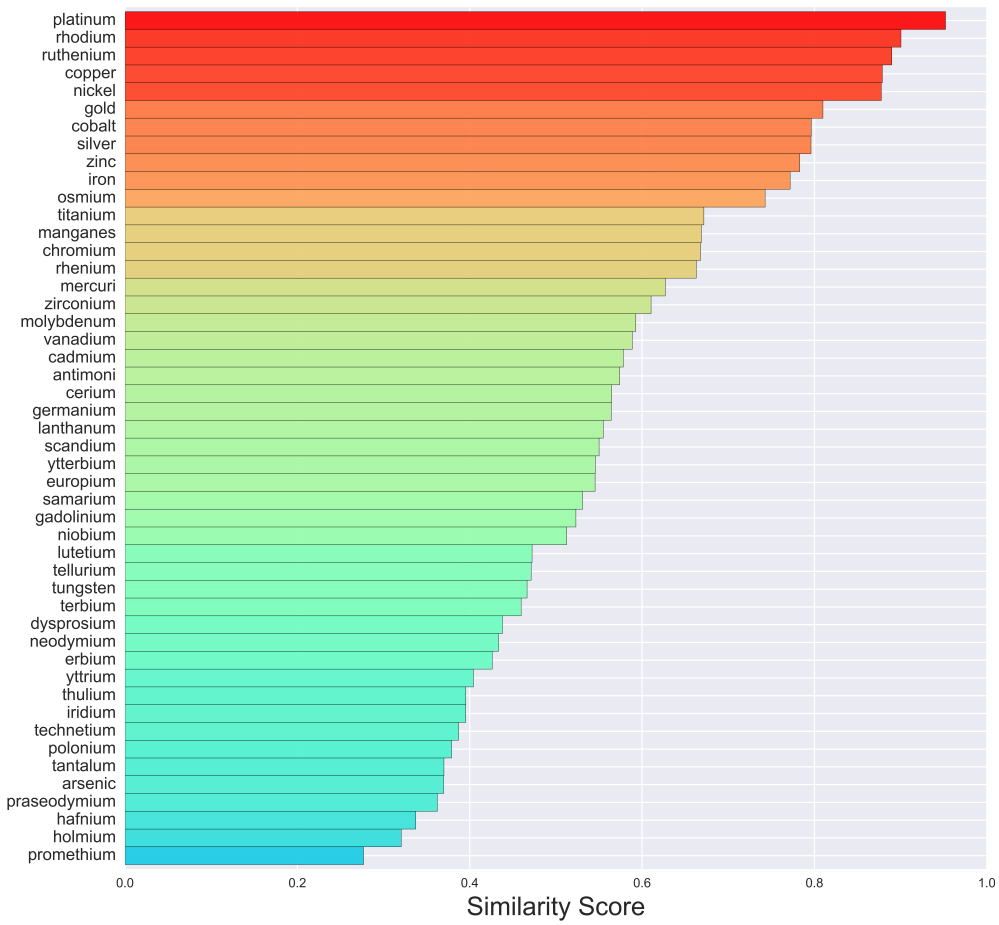


Figure 1.3: The metals are ordered to to bottom from most similar (long bars, red) to low similarity (short bars, blue). The colours are a guide to the eye.

Platinum, rhodium, ruthenium, copper and nickel all had very high scores. The models could be interpreted as suggesting that these metals have similar properties to palladium. This is very much the case for platinum and rhodium (pd, pt and rd are all platinium group metals)[**?**]. Nickel and copper are predicted to be similar to palladium, and there is evidence that nickel could be used for some palladium-catalysed reactions[**?**], whereas copper is often combined with palladium to form more effective catalysts[**?**]. Thus it

could be argued the models suggest that more attention should be focussed to nickel catalysis.

This analysis, whilst brief, is promising. This lends weight that more in-depth considerations of word vectors and concept vectors would be fruitful.

1.4 Data Acquisition Supplementary Information

1.4.1 Publisher Denial of Service

As mentioned in §**??**, Taylor & Francis and ACS banned the scraping computer's IP address during the second stage of global scraping. This section explores why this occurred.

Taylor & Francis banned the IP address after it detected over 100 requests were made within five minutes. This corresponds to a request every three seconds. This was a modest server load compared to other publishers, and was not foreseen to cause problems.

The ACS banning occurred because of a bug in the randomisation of requests. The program was instructed to take a DOI from a random publisher every time it made a request, rather than just a random DOI. Since the largest publisher was ACS, the program eventually exhausted DOIs from the other publishers, until there were only ACS DOIs to 'randomly' draw requests from. This meant the request frequency to the ACS server went up dramatically. This increase broke the threshold of allowed requests at the ACS server which then banned the IP (approximately 10 requests a second).

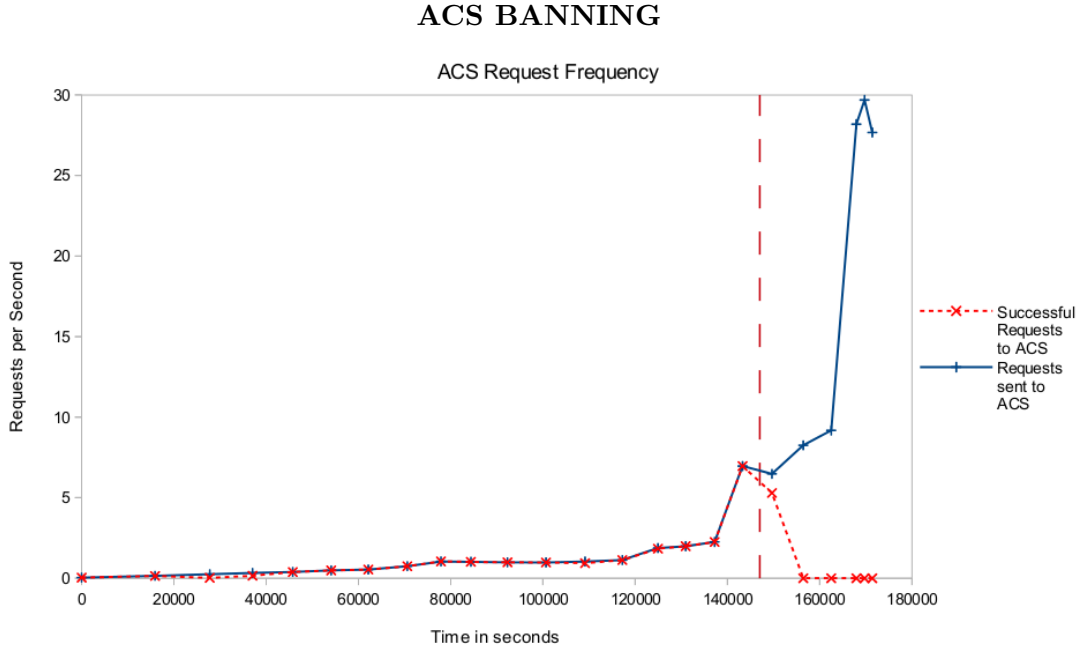


Figure 1.4: The request frequency is plotted in blue, the received pages frequency in red. The vertical dashed line shows where the server detected the scape and banned the IP.

The program was capable of making a total number of approximately 30 requests per

second. As can be seen in figure 1.4, the program began to run out of requests to other publishers after approximately 140,000 seconds, resulting in an increase in the proportion of total requests per second to ACS. The ban occurred after approximately 150,000 seconds, after which there were no more responses received.

1.4.2 Some Observations on $\Delta 1$ through $\Delta 6$

There is much to be learnt by examination and simple statistics of the collected data. This section details some of this analysis which was used in development of the scraping program and to inform upon algorithm and processing design choices.

When deciding how many XPaths were required, it was necessary to examine publication profiles. The publisher 'market share' can be approximated from examining $\Delta 3$.

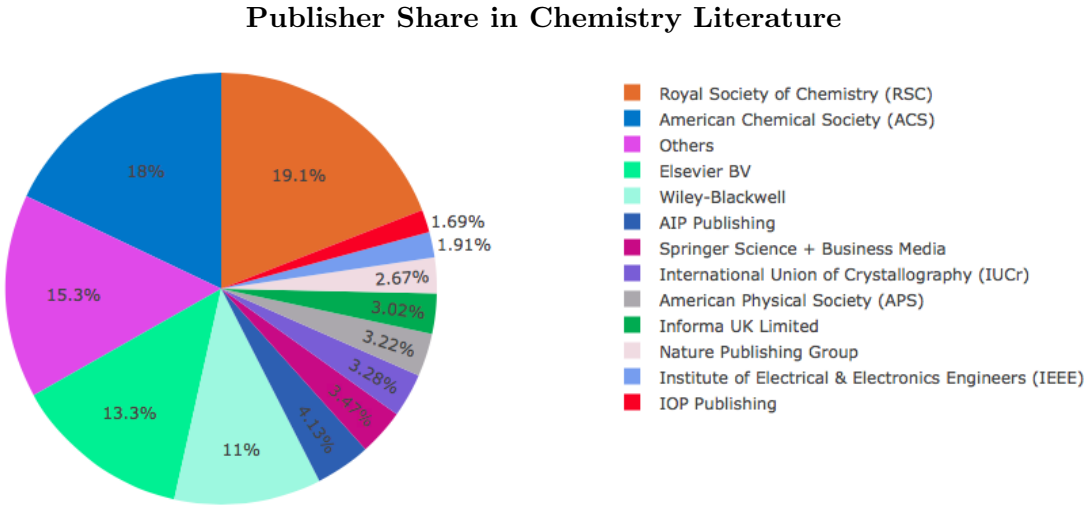


Figure 1.5: Articles grouped by publisher in $\Delta 3$. Only the top 12 publishers are shown.

As shown in shown in figure 1.5, it can be seen that 90% of all the chemistry literature collected was published by just 12 publishers, the majority from ACS, Wiley-Blackwell, Springer and Elsevier BV. Looking at the UK scraping DOI dataset (Figure 1.6), the same large publishers are represented, but the Royal Society of Chemistry has a much larger share. This is to be expected, as the RSC is a UK based body. In the UK, there is a more even distribution between the large publishers.

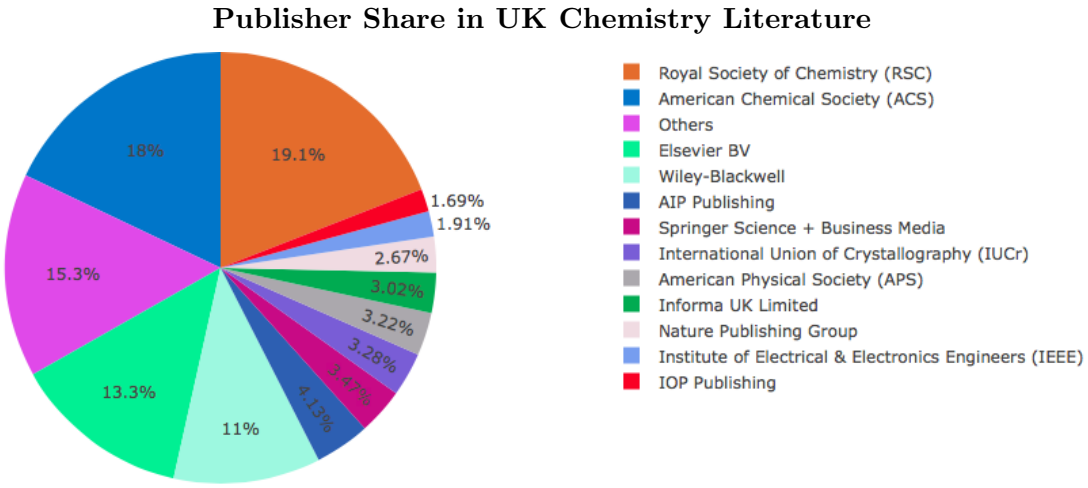


Figure 1.6: Articles grouped by publisher in $\Delta 1$. *Only the top 12 publishers are shown.*

The corpus of combined titles and abstracts in $\Delta 6$ was then examined. An understanding of word distributions would inform data sanitiation practices. It is is included here for interes and completeness. The word frequencies across all the data were found to be approximately Zipfian, with a gradient of -1.11[7] See figure 1.7

[7] A Zipfian distribution is a subset of the Pareto distribution, stating that the frequency of a word is proportional to its ranking in the word frequencies table. Ideally, the gradient of a log(frequency) vs log(rank) should be -1.0 [23]

Approximate Zipfian Distribution of Collected Corpus

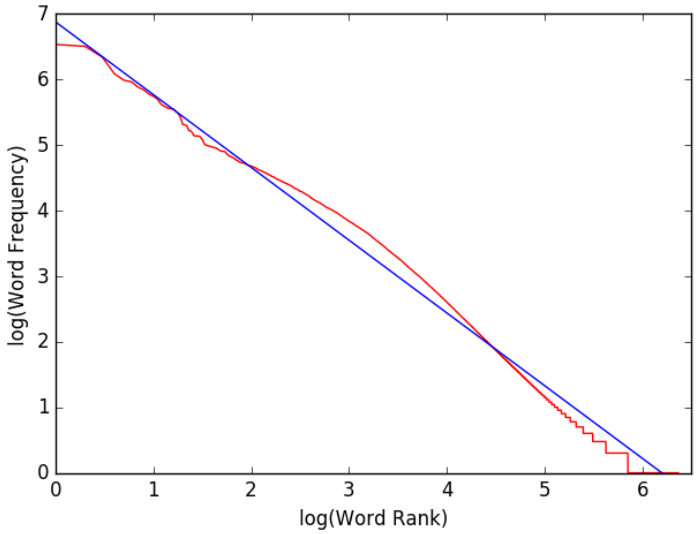


Figure 1.7: The log Frequency of words vs the log of their position in the rank in the word frequency table in blue. Best fit line in red, gradient = -1.11, intercept 6.3.

A summary of the corpus statistics are shown below:

Table 1.1: Titles and Abstracts in Training Database

Total Word Count	61,296,410
Total Unique Words	2,326,725
Total Document Count	464,712
Mode Words per Title	11
Mean Words per Title	12.2
Mode Words per Abstract	156
Mean Words per Abstract	119.7
Mode Sentences per Abstract	4
Mean Sentences per Abstract	5.4

1.4.3 UK Departments

University Chemistry departments with suitable websites were considered when building the input list for the scraping program. Table 1.2 details all the departments that were included. A crawler program was written to navigate through these websites and store urls which had DOIs in them for the main program to scrape.

Table 1.2: UK Chemistry Departments considered in Scraping

Department	URL
Aberdeen	http://www.abdn.ac.uk/chemistry/
Aston	http://www.aston.ac.uk/eas/about-eas/academic-groups/ceac/
Bangor	http://www.bangor.ac.uk/chemistry/index.php
Bath	http://www.bath.ac.uk/chemistry/
Belfast (Queen's)	http://www.qub.ac.uk/schools/SchoolofChemistryandChemicalEngineering/
Birmingham	http://www.birmingham.ac.uk/schools/chemistry/index.aspx
Bradford	http://www.brad.ac.uk/acad/chemistry/
Brighton	http://about.brighton.ac.uk/pharmacy/
Bristol	http://www.bris.ac.uk/Depts/Chemistry/Bristol_Chemistry.html
Cambridge	http://www.ch.cam.ac.uk/
Cardiff	http://www.cardiff.ac.uk/chemistry
Dundee	http://www.lifesci.dundee.ac.uk
Durham	http://www.dur.ac.uk/chemistry/
Edinburgh	http://www.chem.ed.ac.uk/
Essex	http://www.essex.ac.uk/bs/
Glasgow	http://www.chem.gla.ac.uk/
Greenwich	http://www.gre.ac.uk/engsci/study/pharchemenv
Heriot-Watt	http://www.eps.hw.ac.uk/institutes/chemical-sciences.htm
Hertfordshire	http://www.herts.ac.uk/research/hhsri/research-areas-hhsri/pharmacy-and-pharmacology/pharmaceutical-chemistry
Huddersfield	http://www.hud.ac.uk/sas/chemistry/
Hull	http://www2.hull.ac.uk/science/chemistry.aspx
Keele	http://www.keele.ac.uk/chemistry/
Kent at Canterbury	http://www.kent.ac.uk/bio/
Kingston	http://sec.kingston.ac.uk/research/research-centres/
Lancaster	http://www.lancaster.ac.uk/chemistry/
Leeds	http://www.chem.leeds.ac.uk/
Leicester	http://www.le.ac.uk/chemistry/

Department	URL
Lincoln	https://www.lincoln.ac.uk/home/chemistry/
Liverpool	http://www.liv.ac.uk/chemistry/
Liverpool John Moores	https://www.ljmu.ac.uk/about-us/faculties/faculty-of-science/school-of-pharmacy-and-biomolecular-sciences
London Met.	http://www.londonmet.ac.uk/faculties/faculty-of-life-sciences-and-computing/school-of-human-sciences/
Loughborough	http://www.lboro.ac.uk/departments/chemistry
Manchester	http://www.manchester.ac.uk/chemistry/
Manchester Met.	http://www.sste.mmu.ac.uk
Newcastle	http://www.ncl.ac.uk/chemistry/
Northumbria	https://www.northumbria.ac.uk/about-us/academic-departments/applied-sciences/
Nottingham	http://www.nottingham.ac.uk/chemistry/
Nottingham Trent	http://www.ntu.ac.uk/sat/about/academic_teams/chemistry.html
Open University	http://www.open.ac.uk/science/chemistry/
Oxford	http://www.chem.ox.ac.uk/
Univ. West Scotland	http://www.uws.ac.uk/schools/school-of-science/departments/chemistry-and-chemical-engineering/
Plymouth	https://www.plymouth.ac.uk/schools/school-of-geography-earth-and-environmental-sciences/chemistry
Reading	http://www.reading.ac.uk/chemistry/
Robert Gordon	http://www.rgu.ac.uk/about/faculties-schools-and-departments/faculty-of-health-and-social-care/school-of-pharmacy-and-life-sciences1
St Andrews	http://ch-www.st-and.ac.uk/
Salford	http://www.salford.ac.uk/environment-life-sciences/research/biomedical
Sheffield	http://www.sheffield.ac.uk/chemistry
Sheffield Hallam	http://www.shu.ac.uk/schools/sci/chem/
South Wales	http://www.southwales.ac.uk/chemistry/
Southampton	http://www.soton.ac.uk/chemistry/
Strathclyde	http://www.strath.ac.uk/chemistry/
Sunderland	http://www.sunderland.ac.uk/ug/subjectareas/pharmacychemistrybiomedicalsciences/
Surrey	http://www.surrey.ac.uk/chemistry/
Sussex	http://www.sussex.ac.uk/chemistry/
Teesside	http://www.tees.ac.uk/schools/sst/
UEA	http://www.uea.ac.uk/chemistry
Warwick	http://www2.warwick.ac.uk/fac/sci/chemistry/
York	http://www.york.ac.uk/depts/chem/

Department	URL
Bradford Ploymer IRC	http://www.brad.ac.uk/acad/irc/
Cardiff Pharmacy	http://www.cardiff.ac.uk/pharmacy-pharmaceutical-sciences
Burbeck Chemistry	http://www.bbk.ac.uk/bcs/
Burbeck Crystallography	http://www.cryst.bbk.ac.uk/
Imperial College London	http://www.imperial.ac.uk/chemistry/
King's College London	http://www.kcl.ac.uk/nms/depts/chemistry/index.aspx
Queen Mary London	http://www.sbcs.qmul.ac.uk/
UCL School of Pharmacy	http://www.ucl.ac.uk/pharmacy
University College London	http://www.ucl.ac.uk/chemistry/
Sheffield Comput. Chem.	http://www.sheffield.ac.uk/is/research/groups/chemoinformatics

1.4.4 Publishers Considered in UK scraping

The UK scraping run found articles published by 36 different publishers. These are detailed below in table 1.3.

Table 1.3: All publishers found in the UK scraping run

IBM
Pleiades Publishing Ltd
Informa Healthcare
Informa UK Limited
Royal Society of Chemistry (RSC)
Vilnius Gediminas Technical University
Technical Association of Photopolymers, Japan
Springer US
Trans Tech Publications
Thieme Publishing Group
Nature Publishing Group
American Physical Society (APS)
IOP Publishing
Institute of Electrical & Electronics Engineers (IEEE)
American Chemical Society (ACS)
Walter de Gruyter GmbH
Pharmaceutical Society of Japan
American Association of Physics Teachers (AAPT)
AIP Publishing
Japan Society of Applied Physics
American Vacuum Society
Wiley-Blackwell
Springer Berlin Heidelberg
Springer New York
Royal Society of Chemistry
Public Library of Science (PLoS)
Surface Science Society Japan
Springer Science + Business Media
The Royal Society
Society of Rheology
Acoustical Society of America (ASA)
Springer International Publishing
Proceedings of the National Academy of Sciences
Japan Society for Analytical Chemistry
International Union of Crystallography (IUCr)
Chemical Society of Japan
EDP Sciences

Bibliography

[1] Brian S. Everitt. *Cambridge Dictionary of Statistics*. Cambridge University Press, 1998. ISBN 0521593468.

[2] Michael Waskom et al. seaborn: v0.7.0 (january 2016). 2016. doi: 10.5281/zenodo.45133. URL `http://dx.doi.org/10.5281/zenodo.45133`.

[3] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL `http://www.scipy.org/`. [Online; accessed 2016-03-30].

[4] Vincent D Blondel, Jean-Loup Guillaume, and Etienne Lefebvre Renaud Lambiotte. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Thory and Experiment*, 10:1000, 2008.

[5] R. Lambiotte, J.-C. Delvenne, and M. Barahona. Laplacian dynamics and multiscale modular structure in networks. *IEEE Transactions on Network Science and Engineering*, 1:76–90, 2015.

[6] Crossref Foundation. *The Formation of CrossRef: A Short History*. 2009. URL `http://www.crossref.org/08downloads/CrossRef10Years.pdf`. [Online; accessed 2016-03-10].

[7] Norman Paskin. Digital object identifier (doi®) system. *Encyclopedia of Libary and Information Sciences*, pages 1586–1592, 2010.

[8] The doi handbook - international doi foundation (2016). 2014. URL `http://www.doi.org/doi_handbook/7_IDF.html#7.2.1`. [online; Accessed 2016-03-25].

[9] Chris D. Paice. Another stemmer. *SIGIR Forum*, 24(3):56–61, November 1990. ISSN 0163-5840. doi: 10.1145/101306.101310. URL `http://doi.acm.org/10.1145/101306.101310`.

[10] "machine learning". encyclopdia britannica. encyclopdia britannica online. *Encyclopdia Britannica Inc.*, 2016. URL `http://www.britannica.com/technology/machine-learning`. [Online; Accessed 2016-03-25].

[11] Karl Pearson. LIII. on lines and planes of closest fit to systems of points in

space. *Philosophical Magazine Series 6*, 2(11):559–572, nov 1901. doi: 10.1080/14786440109462720. URL http://dx.doi.org/10.1080/14786440109462720.

[12] M. F. Porter. *An Algorithm for Suffix Stripping*, volume 14. 1980.

[13] M. F. Porter. The Porter2 stemming algorithm, 2002.

[14] G.E. van der Maaten, L.J.P.; Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[15] Sokal R and Michener C. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.

[16] The Unicode Consortium. *Unicode® 6.0.0*. 2010. ISBN 978-1-936213-01-6. URL http://www.unicode.org/versions/Unicode6.0.0/. [online; Accessed 2016-03-25].

[17] Mark Davis. Unicode nearing 50% of the web. *Official Google Blog*, 2010. [online; Accessed 2016-03-25].

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL http://arxiv.org/abs/1301.3781.

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013. URL http://arxiv.org/abs/1310.4546.

[20] Ingo Feinerer and Kurt Hornik. *wordnet: WordNet Interface*, 2016. URL https://CRAN.R-project.org/package=wordnet. R package version 0.1-11.

[21] Mike Wallace. *Jawbone Java WordNet API*, 2007. URL http://mfwallace.googlepages.com/jawbone.

[22] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[23] A. Ullah and D.E.A. Giles. *Handbook of Empirical Economics and Finance*. Statistics: A Series of Textbooks and Monographs. CRC Press, 2010. ISBN 9781420070361. URL https://books.google.co.uk/books?id=QAUv9R6bJzwC.

[24] The doi handbook - numbering (2014). 2016. URL https://www.doi.org/doi_handbook/2_Numbering.html#2.2.2. [online; Accessed 2016-03-25].

[25] The copyright and rights in performances (research, education, libraries and archives) regulations, no. 1372 regulation 3. 2014. URL http://www.legislation.gov.uk/uksi/2014/1372/regulation/3/made. [Online;accessed 2016-03-22].

[26] Responsible content mining. 2015. URL http://contentmine.org/wp-content/uploads/2015/06/responsible-content-mining-1.pdf. [Online; accessed 2016-03-22].

[27] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, May 2013. URL http://research.microsoft.com/apps/pubs/default.aspx?id=189726.

[28] David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group, editors. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, Cambridge, MA, USA, 1986. ISBN 0-262-68053-X.

[29] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[30] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014. URL http://arxiv.org/abs/1405.4053.

[31] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009. ISBN 0596516495, 9780596516499.

[32] Surasak Seesukphronrarak and Toshikazu Takata. Novel fluorene-based biphenolic monomer: 9, 9-bis(4-hydroxyphenyl)-9-silafluorene. *Chem. Lett.*, 36(9):1138–1139, 2007. doi: 10.1246/cl.2007.1138. URL http://dx.doi.org/10.1246/cl.2007.1138.

[33] Yu. B. Tsaplev. Photochemical transformations of anthraquinone in polymeric alcohols. *Russian Journal of Physical Chemistry A*, 86(12):1909–1914, oct 2012. doi: 10.1134/s0036024412120266. URL http://dx.doi.org/10.1134/s0036024412120266.

[34] Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014. URL http://jmlr.org/papers/v15/vandermaaten14a.html.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009. URL http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[37] Fernando Pérez and Brian E. Granger. IPython: a system for interactive scientific

computing. *Computing in Science and Engineering*, 9(3):21–29, May 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.53. URL `http://ipython.org`.

[38] John D. Hunter. Matplotlib: A 2d graphics environment, 2007.

[39] S. Chris Colbert Stfan van der Walt and Gal Varoquaux. The numpy array: A structure for efficient numerical computation, 2011.

[40] Wes McKinney. Data structures of statistical computing in python, 2010.