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 metadata 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 finding 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 [**?**] [**?**] [**?**].

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

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'[**?**]. 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 [**?**].

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 [**?**].

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 $\left(A^{\circ\frac{1}{2}}\right)_{i,j} = A_{i,j}^{/frac12}$ 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 [**?**].

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 [**?**].

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

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 [**?**].

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

Porter Early, widely used stemming algorithm [**?**].

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 [**?**] (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 [**?**].

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 [**?**].

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

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 [**?**][**?**] .

WordNet Lemmatizing stemming algorithm, based on consulting database of groups of semantically connected word concepts [**?**] [**?**] [**?**].

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 [**?**].

1. Introduction

1.1 Modern Scientific Publishing

The widespread adoption of the internet in the late 1990s and 2000s, brought fundamental changes to the academic publishing landscape. The information revolution allowed publishers‘ costs to fall, and there was a mood shift in the academic sphere away from subscription based models, towards giving open and free access to some or all of journal article contents. Simultaneously, learnèd institutions (such as university websites) began to post records of recent publications and other chemical information freely online. Publishers still protect the vast majority of journal article content and some metadata. Data is valuable and the insights within, powerful. As such, publishers are unwilling to grant free access to their data, preferring to perform in-house analysis. Article meta-data, such as authors, titles and abstracts may, however, be available, and it is this dataset which the project is focussed on.

1.2 Motivation

By collecting metadata on papers found on the internet, a large, representative dataset of chemical academic writing language can be built up. Machine Learning techniques can then be applied to find novel connections between articles, research communities, authors, institutions and fields. Machine Learning is a rapidly progressing field and data science can reveal key, non-obvious relationships to aid the scientific process. In an increasingly data-dense world, scientists require smarter tools to streamline research in order to be more productive. Several publishers provide services that perform large scale analysis and provide literature tools, such as SciFinder®and Web of Knowledge^{TM}. The techniques used and motivations behind the corporate bodies that own these services are not necessarily clear and thus there is much to be gained from independent, original analyses of the online publishing landscape.

1.3 Aims

The aims of the project are set out below:

- Collect large quantities of article meta-data from articles pertaining to chemistry as a general discipline
  - Identify websites that might contain useful chemical information
  - Write web-scraping programs that can scrape to identify and extract chemical information
  - Store information in human readable, computer readable, scalable and stable formats
- Develop novel machine learning techniques to enable meta-data to be interpreted in new ways
  - Sanitise input data effectively
  - Devise machine learning models to interpret article titles and abstracts to attempt extract their chemical meaning
  - Quantitatively represent an article's content using its collected meta-data
- Validate the model and provide evidence of their efficacy
  - Develop visualisation techniques for interpretation of algorithm output.
  - Analyse datasets using the developed model to demonstrate new and useful information
  - Provide usable code for future analyses to be performed with

This project is thus an informatics/data project, which split naturally into two sections. The first half of the project was concerned with acquiring data. This is covered in detail in §**??**. Programs were written in the python programming language, and two databases were created, one of UK Department Chemistry, and a very large database of unrestricted chemistry related material.

Once the databases were set up, focus was shifted to how to use the data to find valuable insights. §**??** and §**??** provide the background of the algorithms selected used and the process of applying them to create useful models.

Having built the models, it was now necessary to examine their outputs and develop methods to interpret results, which is covered in §**??**. Finally, when the models were shown to be performing successfully, they were used in an analytical setting; To examine the relationships between authors and research communities in the University of Cambridge Chemistry Department and eventually to recommend specific collaborations between staff that were predicted to be fruitful.