A Research Tool for Insights in Academic Topics

A Proof of Concept Application in the field of “Nutrition”

Springboard – Capstone Project 2

Jacques Poolman

February 2020

Table of contents

List of figures

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

This is the second project of two capstone projects that forms part of the Data Science Career Track course offered by Springboard. This project aims to showcase the skills learnt during the course by answering interesting real-life data science questions, especially in the Data Science field of Natural Language Processing (NLP) – the chosen specialisation for this course.

Inspiration for this project came from a recent debate between James Wilks and Chris Kresser on the Joe Rogan Experience podcast. The guests could not agree on the current academic consensus in many areas regarding plant-based diets.

In a different field, for a while, many also argued divided academic consensus on climate change, and some still do. This inspired the idea of a literature review research tool to quickly gain insight into any academic field and discover, for yourself, the academic consensus for any topic in that field.

1.2 PROBLEM STATEMENT

While it seems that there is no official approach to determine academic consensus, at least in the medical field, no algorithms or guidelines exist to this end. Being able to capture and present key opposing views within an academic field, as well as how they measure up to each other, will benefit from an automated meta-analysis. This could provide common ground from where to debate relevant issues and avoid wasting time on semantics.

This vision, in its end state, will benefit any researcher, business or academic, with access to the most current academic consensus in a field, validated by statistical models, summarised by topic analysis models and backed up with sources and citations. An automated visual Wikipedia, if you will, that requires no human contributors.

While this problem might not have clear business impact, yet, it could provide interesting information and insights on the degree of consensus in certain academic fields. An automated tool could analyse any keyword and provide a quick reference for researchers, news reporters interested in specific academic or scientific fields, as well as new technologies. A rapid meta-analysis could have further-reaching applications.

As proof of concept, this project lays the groundwork towards that goal.

1.3 APPROACH

Keeping in mind the vision of a self-service research tool, the scope of this project is larger than what could be accomplished in a single capstone project, however, the proof of concept approach allows for flexibility in testing a single case, with the idea of scaling to any field in the academic domain. To this end, and since the inspiration for this project resulted from a debate regarding “nutrition”, this project investigates the academic field of “nutrition” and “diet”. Although any academic or scientific field would suffice.

This project consists of four parts:

1.3.1 Data Collection

The goal of this project is to automate and scale the academic meta-analysis for any keyword representing a field in the academic domain. In this case, “nutrition”. This requires data collected from an API to satisfy automation. This project uses the Public Library of Science (PLOS) API.

1.3.2 Exploratory Analysis

After data collection, a quick overview of the corpus presents the number of articles related to “nutrition” and how the number of articles changed over the years. In addition, after trigram phrase modelling, an interactive graph displays the top 50 words in a Bag of Words model per publication year and highlights potentially emerging topics.

1.3.3 Data Modelling

To gain insight into underlying themes, topic modelling attempts to identify and explore the relations of subfields within the larger academic field and transforming this output into visualisations. Word vectorisation aids in identifying constructs and finding similar keywords, and document vectorisation offers several applications into finding similar concepts and related articles in the corpus. Lastly, the text summarisation module summarises articles and sentences to explain concepts.

1.3.4 User Interaction

With the data modelling tools in place, the user can input any concept, in the form of words or statements, and obtain related and summarised concepts or articles from the corpus of articles. This aids the researcher in gaining further insight into the selected academic field.

1.4 PROCESSS

Figure 1.1 shows the outline of the project and explains the process from data collection, through modelling, to output and user interaction.

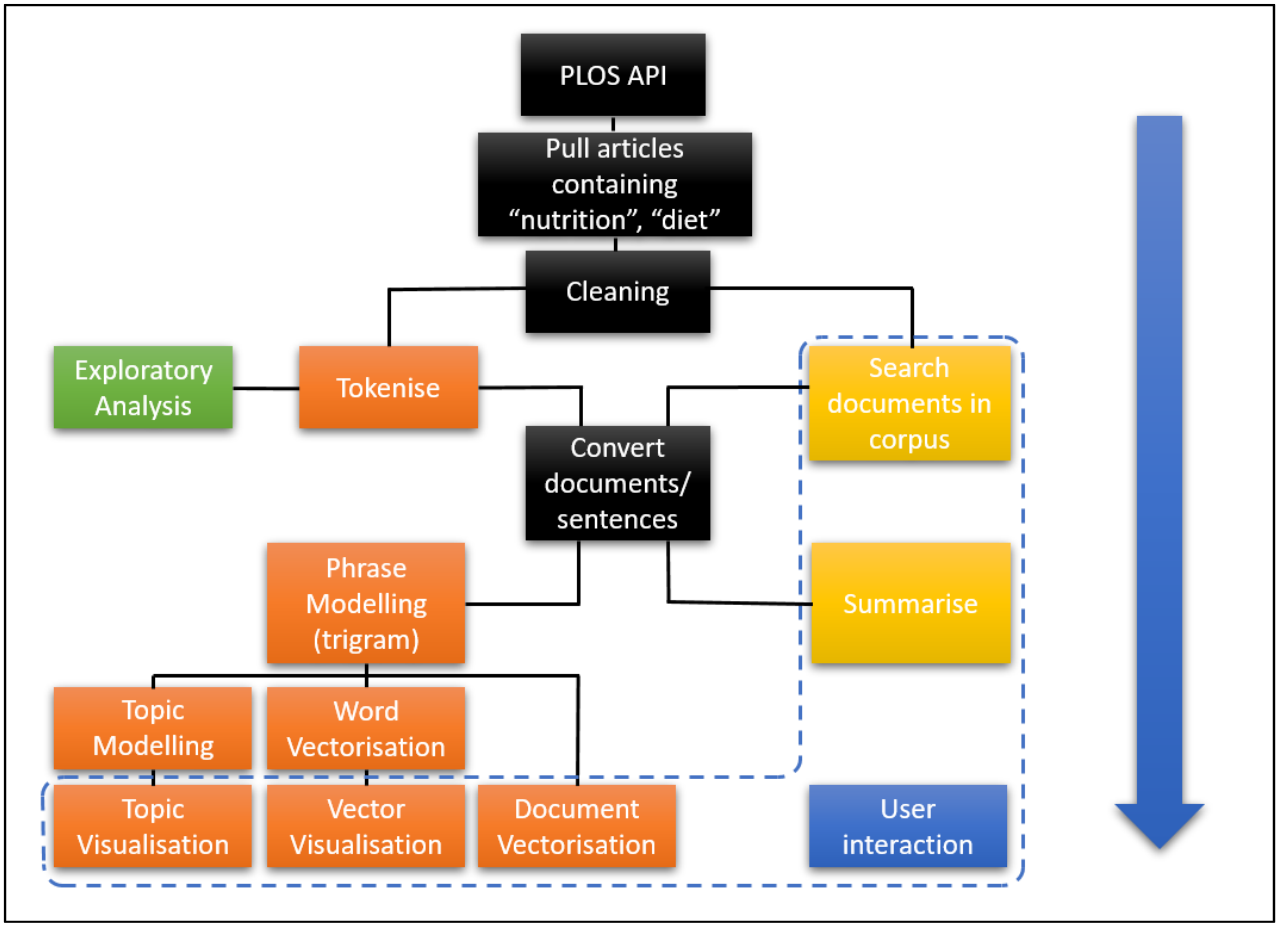

Figure 1.1: Project flow

1.5 MEDIUM

The project code is currently contained in four Jupyter Notebooks; however, the completed code will be productionised for publishing online and through an API.

CHAPTER 2
DATA

2.1 DATA COLLECTION

Many scholarly APIs containing a collection of academic and scientific articles can be found online. The scholarly publishing section in the MIT Libraries contains a list of some academic publishing APIs. On reviewing some of these APIs, the PLOS (Public Library of Science) API offers the most favourable features. The PLOS API requires no API Key and allows researchers to search the Conclusions section of articles, as most APIs only allow searching of the Title section. It is clearly vital to collect relevant Conclusions sections of articles, as this is arguably where the most informative and condensed insights are summarised.

The data collections notebook (1_api_get.ipynb) contains the code for collecting the relevant articles. It retrieves all articles in the PLOS database containing the keywords “nutrition” or “diet” in the Title and Conclusions sections, converts it from JSON to a data frame and saves it in the file called corpus_raw.csv. Depending on the size of the corpus, 1,531 articles in this case, retrieving all the articles could be time consuming in order to satisfy the rate limit imposed by the API.

The current data frame contains three features, the Publication Date, Title and Conclusions sections for each article. More features can be included in future to expand functionality, for example the article ID, authors, affiliates and disclosures.

2.2 DATA WRANGLING

After data collection from the API, the data required cleaning, tokenisation and storing into a list of documents (articles) and sentences. The data of interest now is the feature ‘Conclusions’. All further data cleaning and modelling applies to this feature. Figure 2.1 show these steps in the blue broken-line box of the process flow chart.

2.2.1 Cleaning and Tokenisation

As articles are published using different software applications, it results in different text formats. This requires general cleaning to normalise text to the same format. Text in some articles contain new-line escape characters, and others contain punctuation errors between sentence breaks, which makes sentence splitting difficult.

Once each article, from now on referred to as a document, are cleaned and separated into sentences, each sentence is parsed and tokenised for further standardisation using the spaCy NLP library. Once tokenised, punctuation and stop words are removed and remaining words lemmatised.

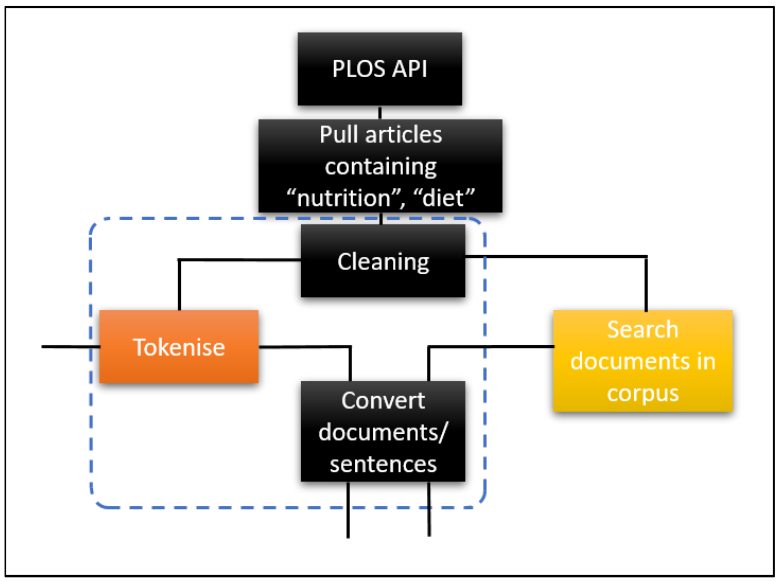

Figure 2.1: Data wrangling in the process flow chart

2.2.2 Documents and sentences

Depending on the requirements for the application, there are two ways to present the data for further processing, as illustrated in Figure 2.2. The group on the left is a list of sentences per document and the group on the right combines all the sentences from all the documents into one list of sentences. The reason for the two approaches is determined by the context of the model and the information your wish to learn from it. This concept will become clear in later sections.

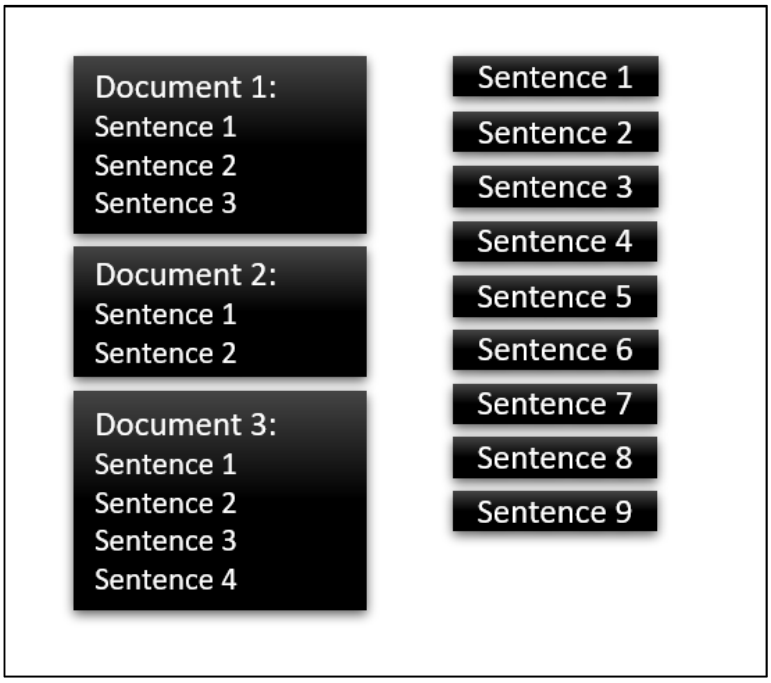

Figure 2.2: Documents versus Sentences

CHAPTER 3
EXPLORATORY DATA ANALYSIS

3.1 SCOPE OF AVAILABLE ARTICLES

In the first step to gain insight into any academic field, it is necessary to understand the context of where the data originate from. In the case of the PLOS API, it launched its first open access journal in 2003. Unfortunately, it does not include historic academic articles. Figure 3.1 shows the number of articles published with the keywords “nutrition” and “diet” in its Conclusion section. The number of articles increased drastically from 2012 as the PLOS API gained popularity.

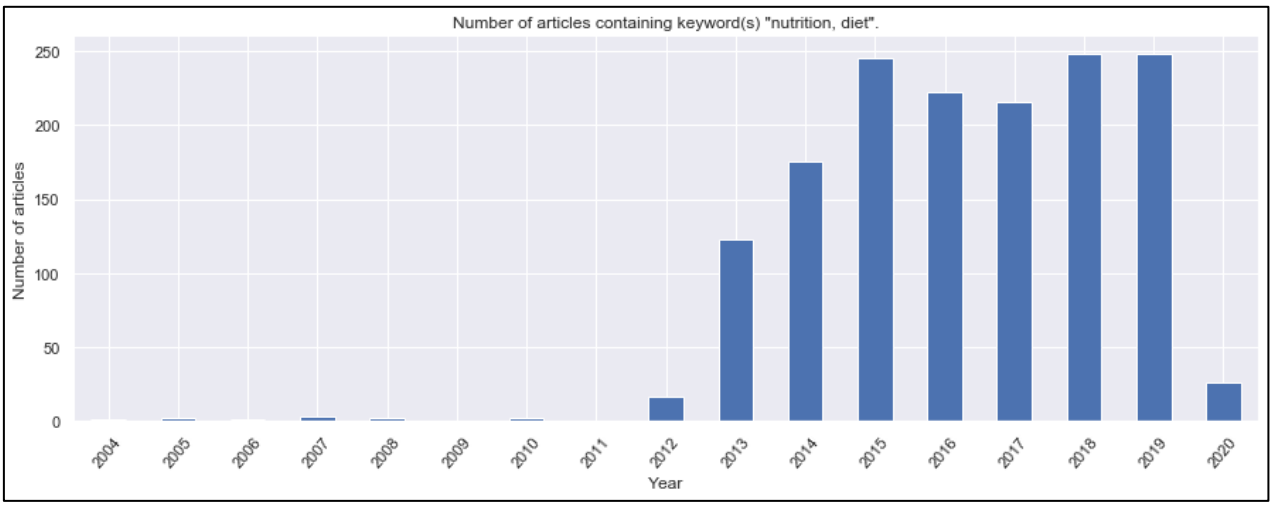

Figure 3.1: Number of Articles Published for “nutrition” and “diet”

Keeping in mind the proof-of-concept approach to this project, one API, even with limited number of articles, is sufficient in the first development iteration. Future builds will capture articles from all scholarly APIs and APIs that include historic articles. This will offer the context required to present a comprehensive scope of any academic field.

3.2 BAG OF WORDS

As a brief overview of the topics we could expect, a simple Bag of Words (BOW) model gives the word counts for the top 30 most occurring words. Figure 3.2 shows a bar plot for the most occurring words in the Titles of all the articles combined.

3.2.1 “Titles”

It is not surprising that the words “study” and “effect” have the highest count, as such words are to be expected in the Title of most articles. The most descriptive words with the highest counts are “food”, “child”, “diet” and “health”, which are intuitively related to the academic field of nutrition. The next word, “acid”, could is arguably less relevant and could offer interesting insights. Figure 3.3 presents the classic graphic representation of the BOW for the article Titles.

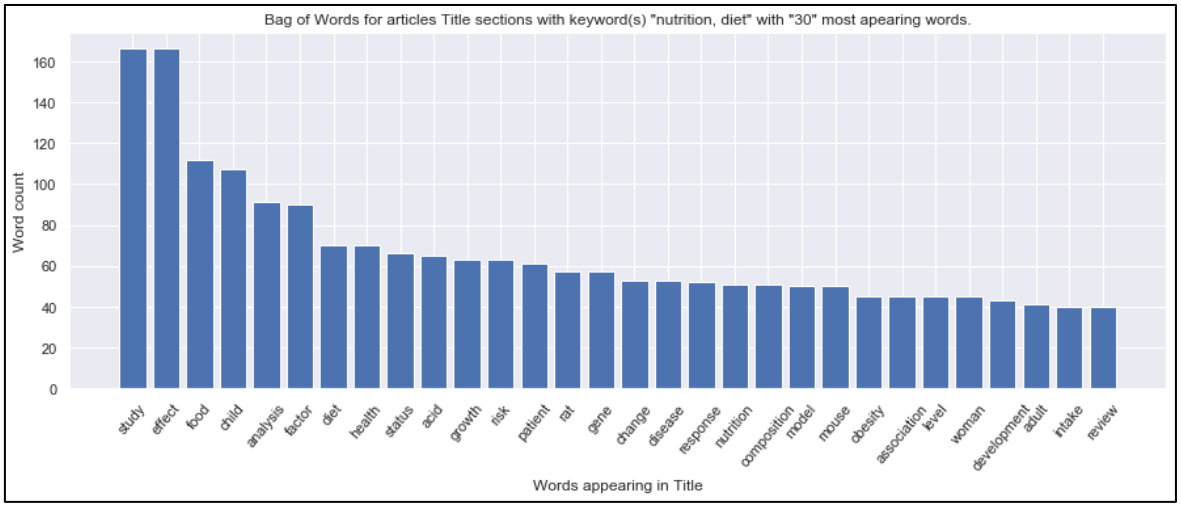

Figure 3.2: Bar Chart for BOW – Titles

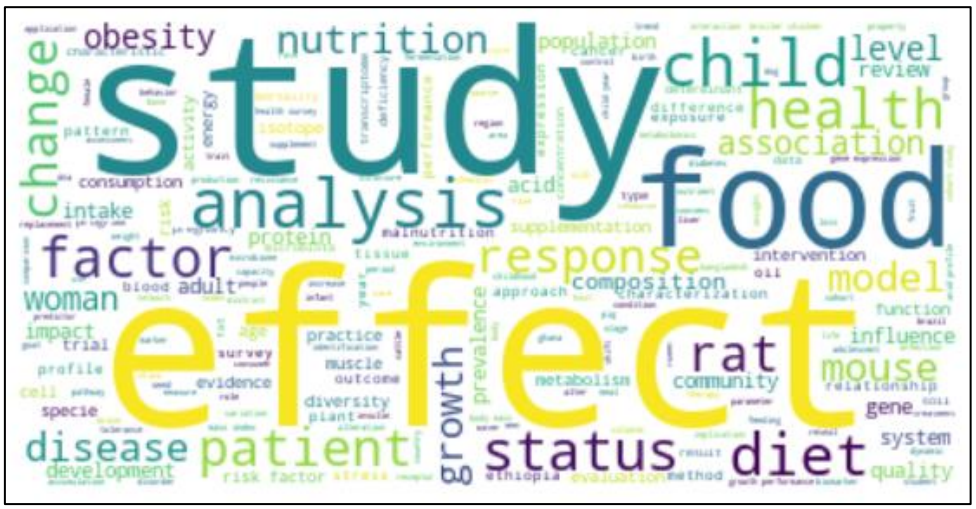

Figure 3.3: BOW – Titles

3.2.2 “Conclusions”

The BOW model for the Conclusion section of articles does not reveal much new insight. In fact, the descriptive words that were evident in the Titles BOW model, now gets diluted by words typical in academic research paper nomenclature. Figures 3.3 and 3.4 nonetheless displays the BOW model for the Conclusion section. Perhaps a named entity recognition vocabulary that filters out common academic research words as stop words would improve the model.

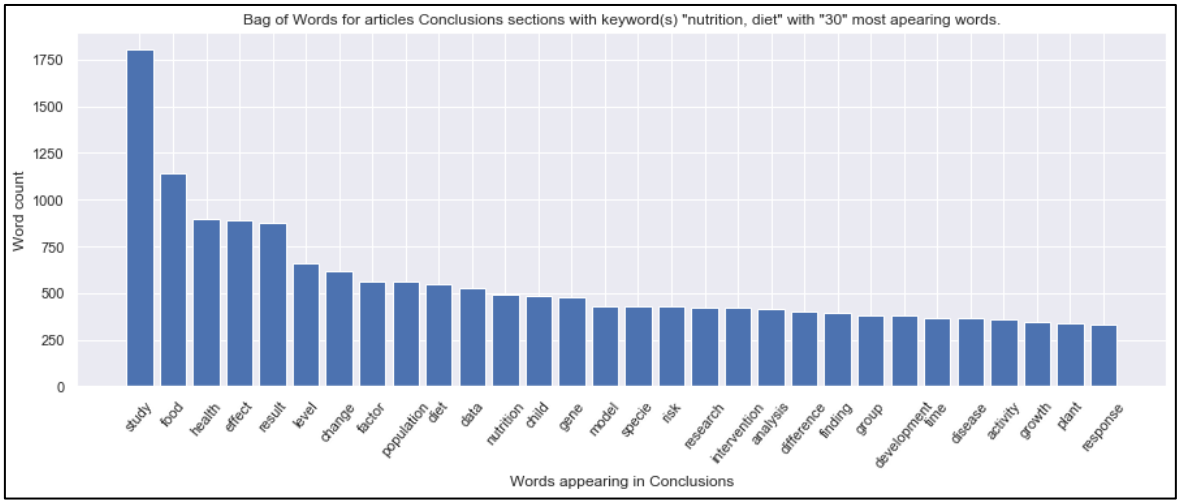

Figure 3.4: Bar Chart for BOW - Conclusions

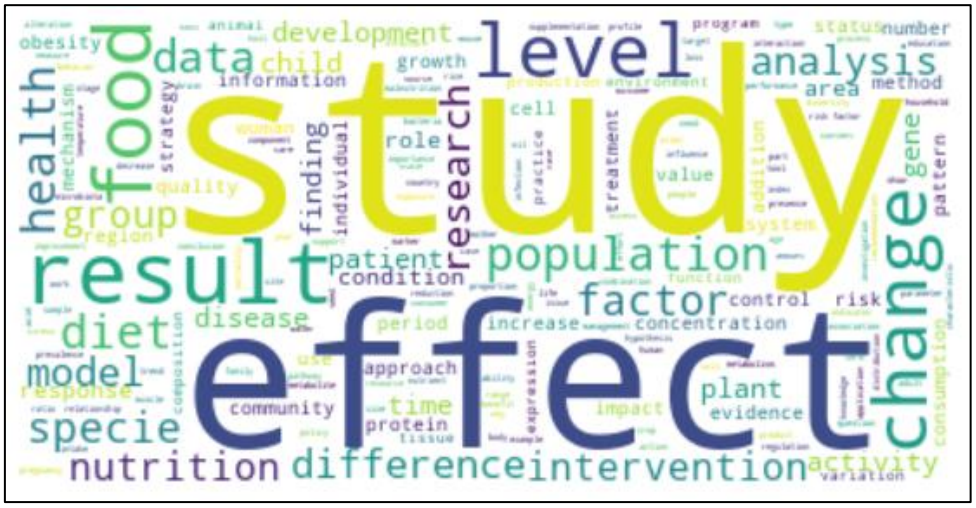

Figure 3.5: BOW – Conclusions

3.3 SUMMARY

Due to the limited variation of data features required for this NLP project, statistical inference does not reveal much interesting insights, other than the growth of published articles since 2012. A quick data exploration otherwise revealed the limitations of using only one API, however only one API is required for proof of concept. Scaling towards more data access by including more APIs, especially access to historic articles, should be simple.

The BOW model on article Titles hinted at some interesting topics, although the same approach on article Conclusions offered little insight.

In the next step, some powerful NLP models will do the heavy lifting to identify topics and gain insight into the corpus of academic articles. Phrase modelling, topic modelling, word and document vectorisation will build on the limited keywords identified in the BOW.