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 tables

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.

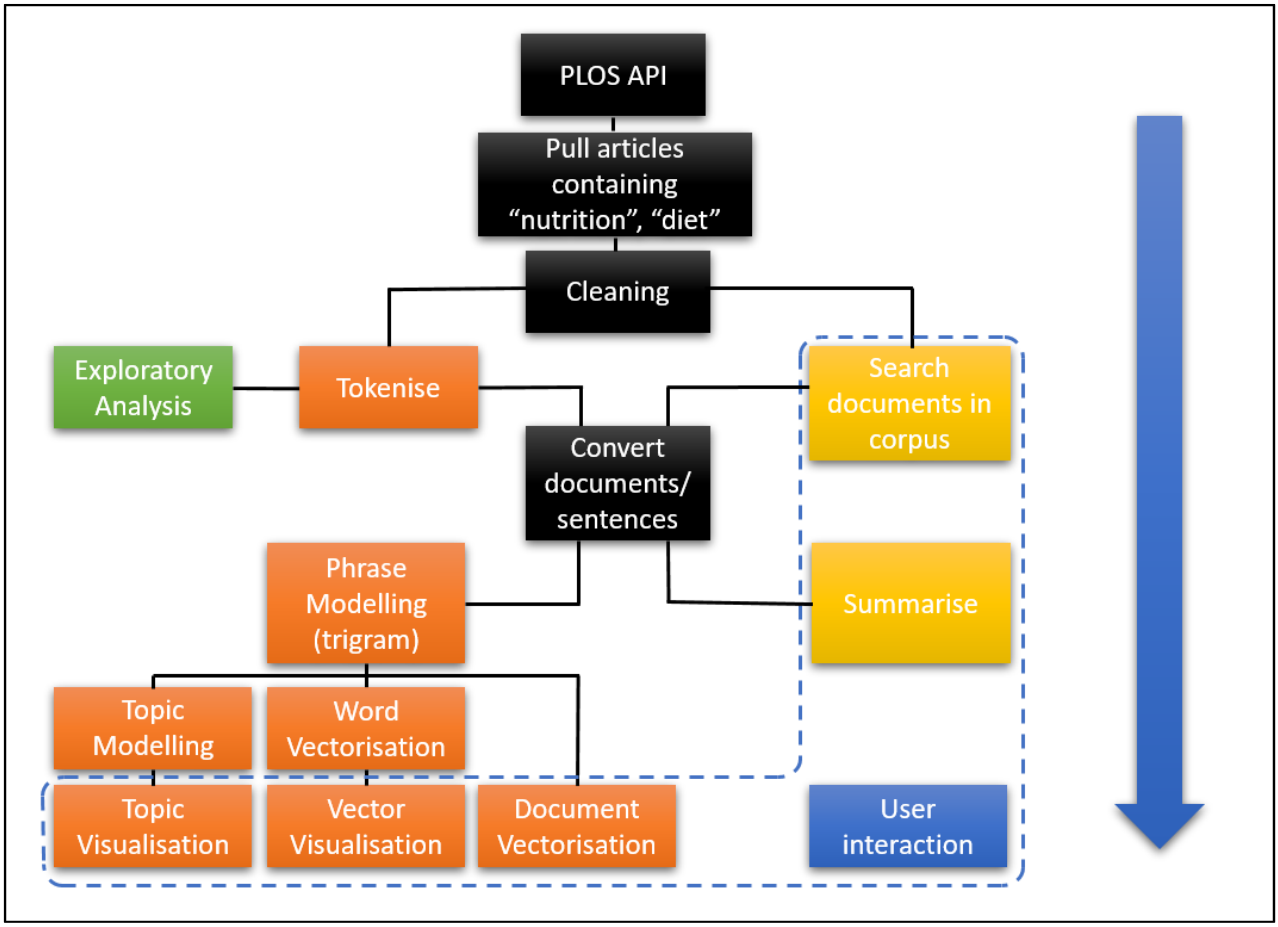

Figure 1.1: Project flow

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.

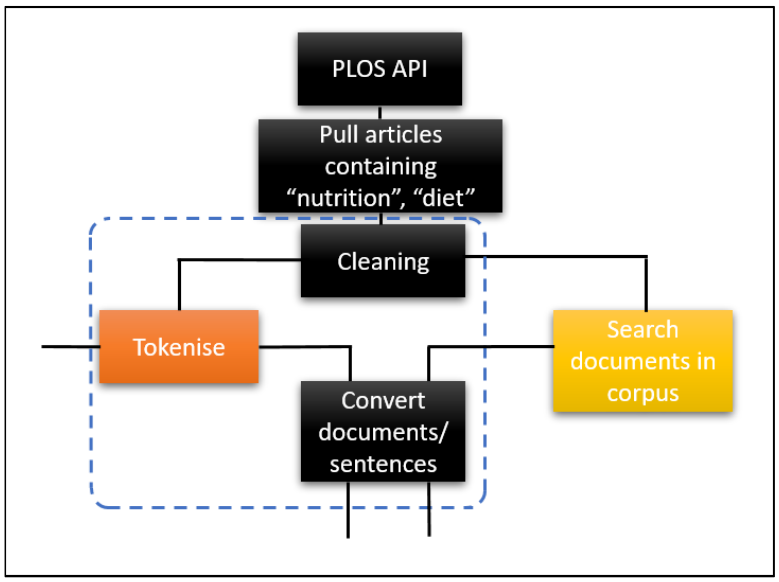

Figure 2.1: Data wrangling in the process flow chart

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.

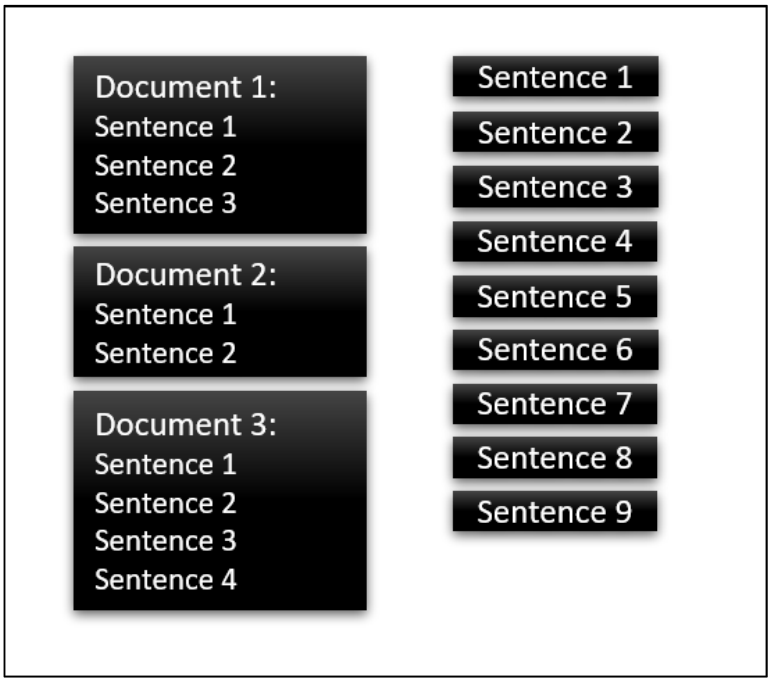

Figure 2.2: Documents versus Sentences

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. There are altogether 1,530 articles in the PLOS API containing the keywords “nutrition” and “diet”. Together these articles have 12,466 sentences and 192,273 words.

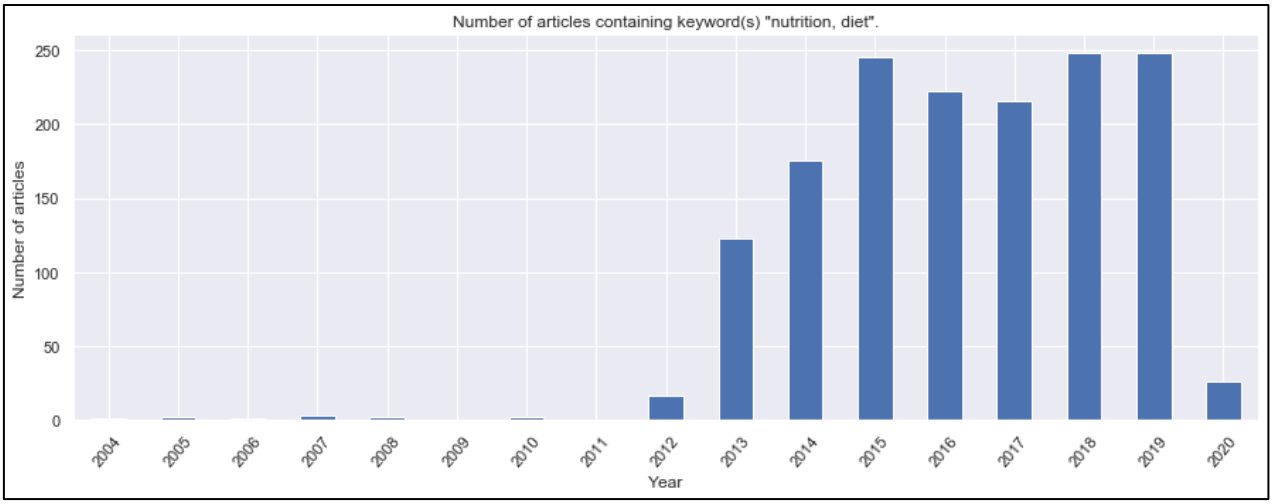

Figure 3.1: Number of Articles Published for “nutrition” and “diet”

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” is arguably less relevant and could offer interesting insights. Figure 3.3 presents the classic graphic representation of the BOW for the article Titles.

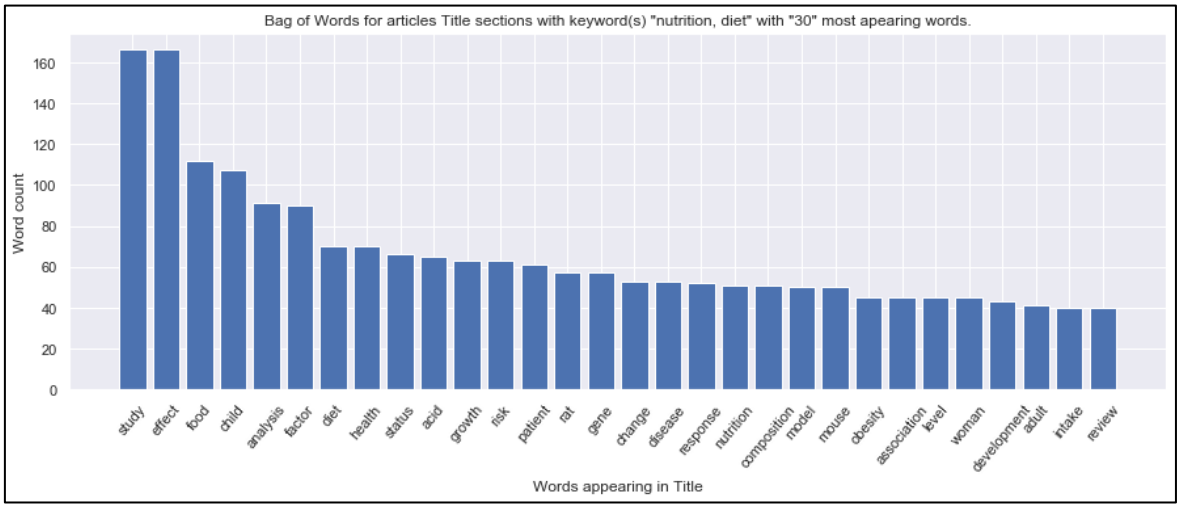

Figure 3.2: Bar Chart for BOW – Titles

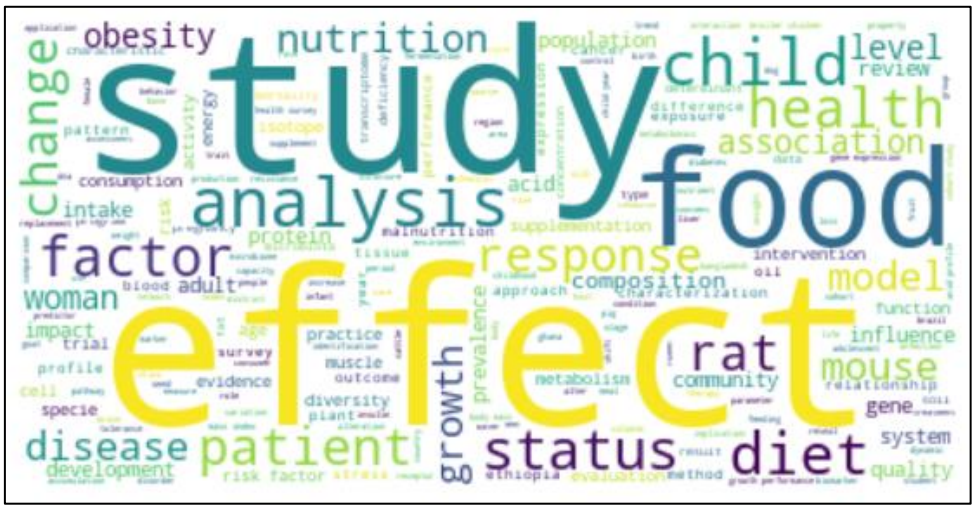

Figure 3.3: BOW – Titles

3.2.2 “Conclusions”

The BOW model for the Conclusion section of articles does not reveal 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.

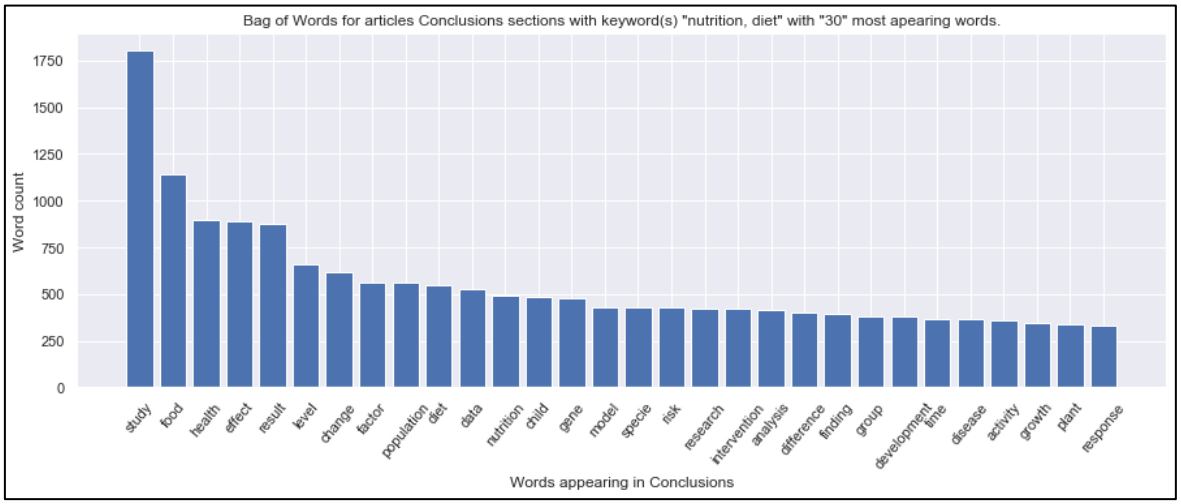

Figure 3.4: Bar Chart for BOW - Conclusions

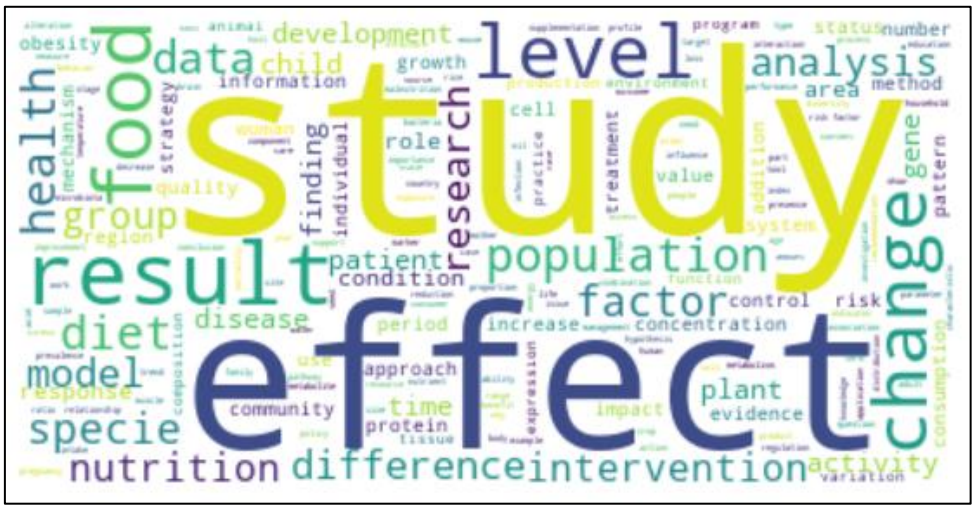

Figure 3.5: BOW – Conclusions

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 do the heavy lifting to identify topics and gain insight into the corpus of academic articles to aid research. Phrase modelling, topic modelling, word and document vectorisation build on the limited keywords identified in the BOW.

CHAPTER 4
DATA MODELLING

4.1 ANALYSIS

The data modelling approach included various models and methods to gain insight into the academic articles. Figure 4.1 explains the process of further data preparation through phrase modelling for the benefit of topic modelling and word vectorisation. Document vectorisation and summarisation offers additional insight by the user when interacting with the data.

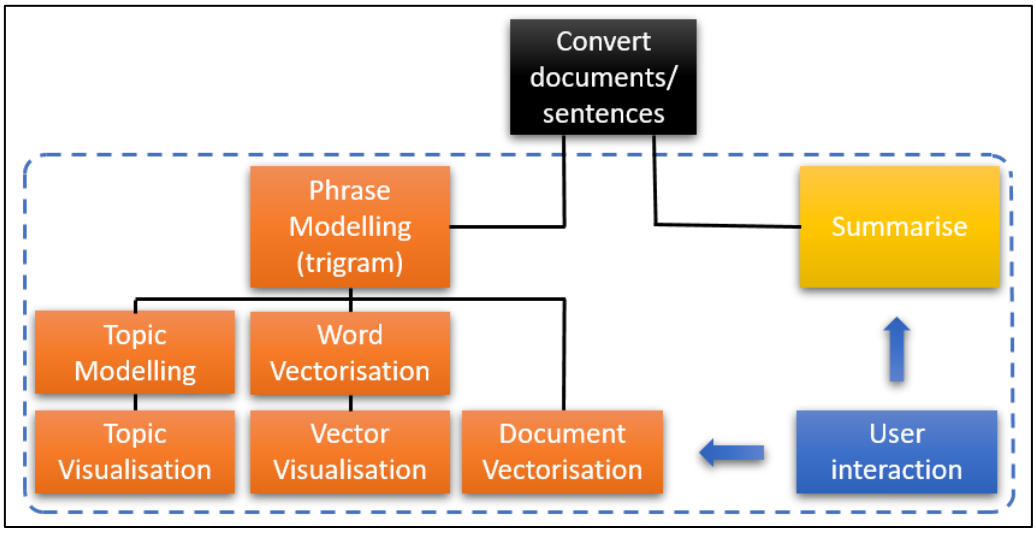

Figure 4.1: NLP modelling in the process flow chart

The aim of this approach is to demonstrate the potential that exist for researchers to gain insight into any academic field. Keep in mind the user's unfamiliarity with the academic field of interest. In this case, the author's interest into the field of nutrition with a complete lack of subject matter expertise. The challenge therefore is not to prove the successful classification of topics, but to reveal patterns that will aid understanding in this field.

In addition, where classical supervised machine learning models can measure accuracy, the best model can be found deterministically. However, with unsupervised learning such as topic modelling, the correct number and fit of topics are mostly a subjective measurement by the user. Measuring how much insight is gained from analysing the corpus of articles is challenging.

4.2 PROCESS AND METHODS

This section briefly describes each model with its associated results. Refer again to Figure 4.1 for the process flow and where each model fits into the data process.

4.2.1 Phrase Modelling

This project uses Gensim, an open-source library for topic modelling and NLP processing using statistical machine learning. Phrase Modelling is necessary to prepare the data for topic modelling, word vectorisation and document vectorisation.

After each iteration of Phrase Modelling, the total number of words in the corpus decreased, as can be expected. The first model resulted in bigrams that reduced the overall word count from 192,273 to 179,960, or 12,313 less words. Running Phrase Modelling the second time resulted in trigrams and reducing the total word count to 178,246, or 1,714 less words. Running Phrase Modelling a third time does not reduce the total word count by a significant enough margin.

Phrase Modelling resulted in some of the following bigrams and trigrams:

- life_cycle
- public_health
- evidence_base
- nutritional_supplement
- consumption_fruit_vegetable
- plant_base_diet
- dietary_diversity_score
- low_birth_weight

Clearly the bigram “life_cycle” infers more meaning than the separate words of “life” and “cycle”. This naturally improves topic modelling performance.

4.2.2 Topic Modelling and Visualisation

Topic Modelling is a statistical method for identifying abstract themes in a collection of documents, or articles in this case. This project uses Gensim’s Latent Dirichlet Allocation (LDA) model for text classification to a topic.

Since the LDA model requires user input for the number of topics it must identify, an arbitrary number of 15 topics was chosen. See the Suggestions section of this report on how to determine the optimum number of topics. Table 4.1 shows the results for the top 10 relevant words for each topic.

Table 4.1: Top 10 Words per Topic

Topic	Top 10 relevant words									
No.	0	1	2	3	4	5	6	7	8	9
topic0	food	result	nutritional	change	increase	health	datum	low	include	provide
topic1	result	level	population	change	increase	find	low	suggest	patient	research
topic2	high	food	increase	result	use	important	low	protein	change	level
topic3	increase	low	result	population	level	high	health	child	food	different
topic4	change	different	high	find	result	group	specific	protein	use	population
topic5	child	high	result	feed	protein	research	food	datum	need	female
topic6	dietary	level	increase	result	high	datum	model	provide	change	individual
topic7	food	different	change	woman	result	increase	dietary	high	group	population
topic8	result	different	low	food	intervention	control	increase	change	reduce	high
topic9	increase	high	change	nutritional	obesity	population	level	improve	base	result
topic10	increase	intervention	change	high	base	health	result	different	cell	suggest
topic11	plant	change	high	food	result	child	specie	time	low	increase
topic12	high	result	increase	level	factor	different	suggest	improve	find	population
topic13	population	level	community	increase	model	result	low	health	child	associate
topic14	increase	high	result	patient	specie	low	reduce	protein	find	gene

Unfortunately, attempting to extract meaning here does not offer much insight into the underlying topics associated with nutrition. Even running the model for a range of topics from 5, 15, 30, 60 and 120 does not reveal any identifiable topics. Eliminating academic word nomenclature, as suggested before, could perhaps display more relevant words per topic. Also, revisiting LDA after scaling data collection to source articles from all APIs, with historic articles, could drastically improve this model's performance.

To visualise the topics interactively, the pyLDAvis library is used to reduce the high dimensionality of the data to two principle components, in order to view it 2-dimentionally. Figure 4.2 displays the 15 topics visually. For the same reasons as mentioned above, it is not possible to extract much meaning at this stage.

4.2.3 Word Vectorisation

Word Vectorisation is the method of producing word embeddings through shallow neural networks that are trained to produce meaningful context of words. This project uses Gensim's Word2Vec model with the following parameters: Convert each word to 100 dimensions, window size of 5 words, ignoring words with frequency less than 20, using the skip-gram training algorithm and for 100 iterations. This model results in a Word2Vec dictionary size of 1,681 words.

The Word2Vec model produces interesting results. For example, for any given word of interest, the model finds other similar words in the vector space close to the selected word. In the same way opposite words (difference) are obtained by finding words far removed in the vector space. It is also possible to add and subtract words from each other and discover contextual meaning in this way.

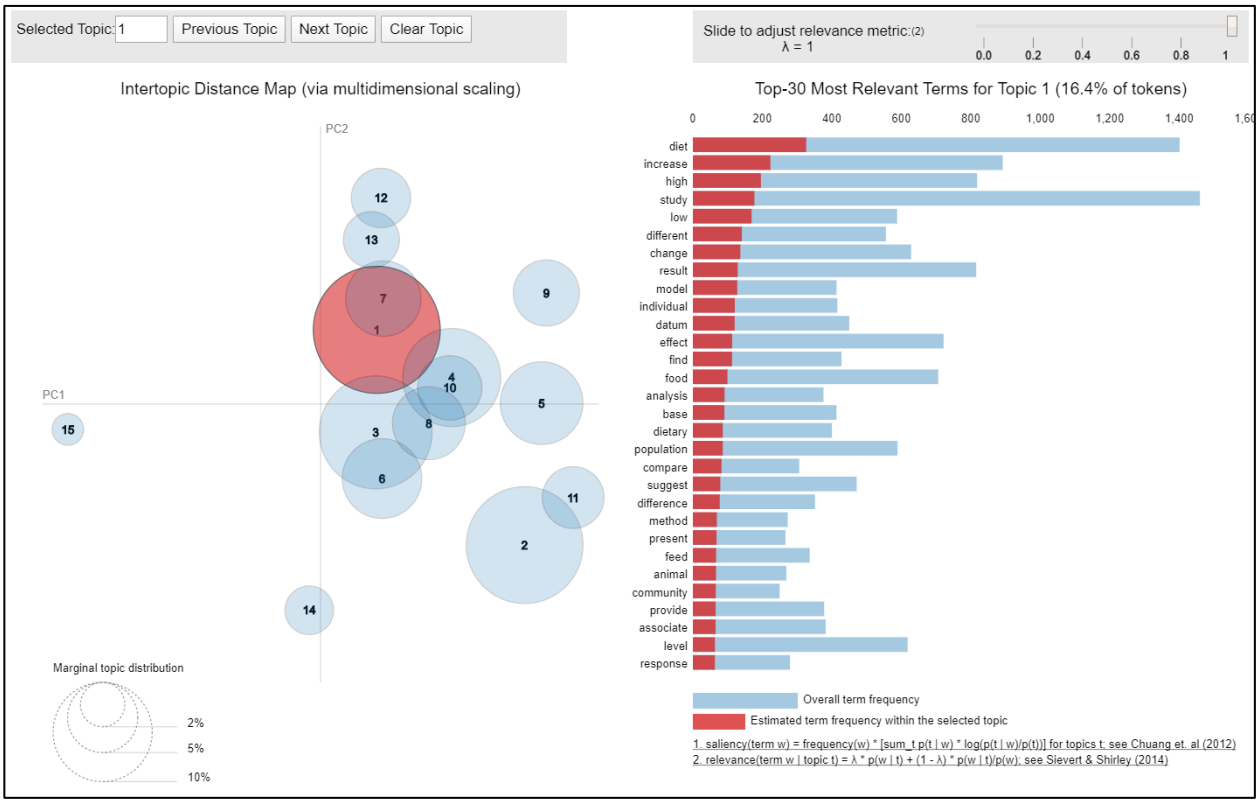

Figure 4.2: LDA Visualisation with pyLDAvis for 15 Topics

i. Similarity

Table 4.2 Here are a few examples of words and its associated relevance score between 0 and 1.0 for similarity. A higher score infers higher similarity.

Table 4.2: Similarity

Similar Words							
carbohydrate		**fat**		**protein**		**death**	
synthesis	0.4663	cholesterol	0.4715	plasma	0.5040	clinical	0.3634
energy	0.3804	supplement	0.4500	acid	0.4908	mortality	0.3618
insulin	0.3765	fatty_acid	0.4492	fatty_acid	0.4818	life	0.3578
intake	0.3724	liver	0.4358	expression	0.4619	cardiovascular	0.3465
metabolism	0.3644	body_weight	0.4187	enzyme	0.4599	neonatal	0.3399
fat	0.3584	dietary	0.4169	amino_acid	0.4148	peptide	0.3289
transport	0.3567	total	0.4115	source	0.4142	perform	0.3208
ingest	0.3562	DM	0.4010	metabolism	0.4005	screening	0.3174
reveal	0.3558	weight_gain	0.3916	degrade	0.3971	newborn	0.3135
high_fat	0.3532	dairy	0.3898	induce	0.3935	birth	0.3126

Similar word meanings for carbohydrates, fat, protein and even death, result in relevant meanings. Although, some familiarity in the field of nutrition would certainly aid in identifying further meaningful insights.

ii. Similarity with Difference

Since word vectorisation results in vectorising each to word to 100 dimensions, it is possible to apply some word algebra to the word vectors. For example, subtracting one word vector from the another, or even adding two or more word vectors together, result in interesting interpretations. Table 4.3 shows results of this approach.

The results of these examples seem intuitive and demonstrate a level of understanding by the model. For example, “child” minus “food” equals “undernutrition”, or the opposite of life, is death (life + negative = death). Although this model does not aid in topic identification, it could contribute as a research tool in identifying related concepts for further analysis.

Table 4.3: Similarity with Difference

Word Algebra				
life + negative	**child - food**	**nutrition - food**	**growth - food**	**health - nutrition**
death	undernutrition	breastfeed	height	colony
year	stunting	organ	organ	nutritional_intervention
young	stunt	clearly	damage	overweight
10	height	nutrition_intervention	deposition	behavioral
live	nutritional_intervention	underlying	inflammation	indicator

iii. Visualisation

An effective to way to visualise word embeddings, or any high dimensional data such as word vectors, is through the sklearn library called t-SNE, or t-distributed Stochastic Neighbour Embedding. The 100-dimensional vector space is reduced to two principle components for viewing the data 2-dimensionally. The plot in Figure 4.3 in an interactive plot with a hovertool to show a point on the figure with a cluster of similar words.

With clearly defined topics and presented with different colours in the plot, would improve the readability of the data significantly.

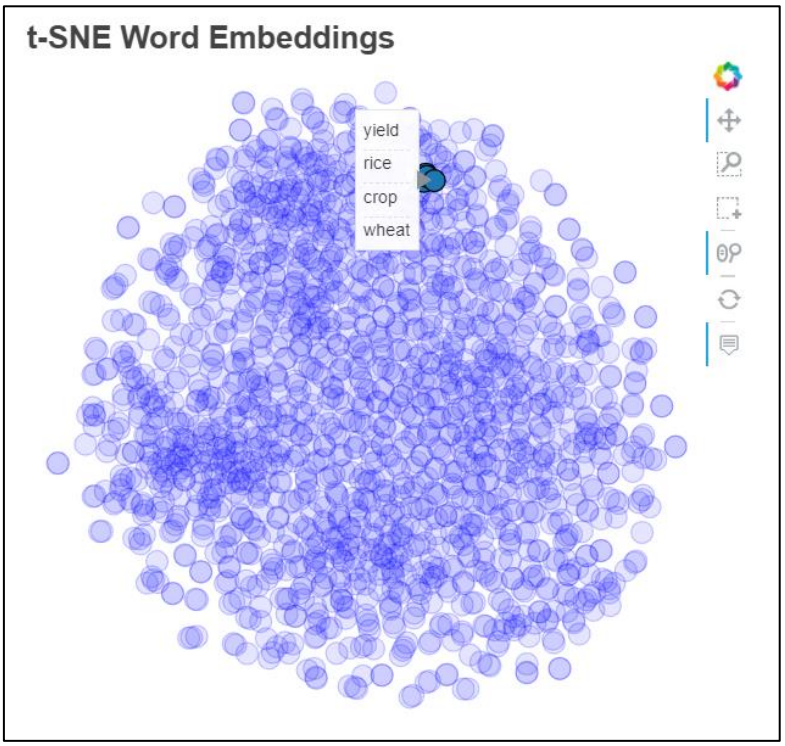

Figure 4.3: t-SNE Plot for all Words

4.2.4 Document Vectorisation

Like word vectorisation, Document Vectorisation is an unsupervised learning method to generate vectors from sentences or paragraphs. This project uses Gensim's Doc2Vec model with the following parameters: Convert documents to 100 dimensions, using the distributed memory training algorithm, with initial learning rate of 0.025 and 100 iterations.

Also, like with word vectorisation where the model can find the most similar words, here a sentence or statement produces the most similar documents. For example, observe the most similar document extract for the following statement: "a child requires healthy nutrition to avoid stunting growth". This yields an article in which the following sentence appears with a relevance score of 52.99%: "We observed an association between early nutrition or growth, and depression at 30 years, this association seems to be cumulative, because the risk of depression was higher only among subjects who were SGA and were also stunted at age two or four years."

This could offer useful insight as a research tool in finding relevant articles in the corpus. In addition, relevant article's Title and ID number could direct the researcher towards finding applicable content quicker. Lastly, once an effective topic model exists, displaying the positions of most applicable documents on the topic model visualisation, could further indicate some consensus where visual clusters of documents exist.

4.2.5 Summarisation and Search

Again, using the Gensim library, documents are summarised into the most important sentences only. The summarise model works on the ranks of text sentences using the Gensim TextRank algorithm. With this approach it is possible to summarise each article into one sentence. In fact, what the model does is select only the most relevant sentence from the document.

The real value of the summarising module becomes apparent when searching for certain keywords. The search function allows for input of any number of keywords which searches the corpus for all articles containing all the key words, inclusive. The model then summarises these articles or sentences for review.

There are two methods of searching and reviewing the corpus. One is by combining all the sentences in the corpus together, independent of the articles it originates from, and only selecting these sentences containing the key words. The second method is by searching each article and including whole articles that include the desired key words. When referring to articles, the model is still only interested in the Conclusion sections of the article. This is after all the only text retrieved from the API.

4.3 SUMMARY

With the cleaned data, Phrase Modelling converts words frequently used together into bigrams and trigrams as required. This aids Topic Modelling, however the LDA topic model cannot identify and display topics in a way that uncovers distinct topics or themes. In addition, manually choosing the number of topics is not ideal.

Word and document vectorisation achieve interesting results in displaying similarly relevant words and documents. Documents Vectorisation presents immediate application value in finding similar articles and ideas, whereas a word vectorisation application is not immediately obvious, but it does have interesting visuals potential through means of the t-SNE model.

Lastly, along with document vectorisation, summarisation displays definite potential, especially though user interaction when searching for selected key words. An added benefit is the option of searching per article, or the combined sentences corpus.

4.4 FINDINGS

At the outset the project attempts to search for and find academic consensus. To achieve this, topic modelling must be able to effectively distinguish between topics and display it visually. This approach does not seem to achieve its aim at the moment and requires more research.

Word and document vectorisation along with summarisation deliver interesting insights, especially with user interaction. A user can search for key words and find related words or articles and summarise content of interest.

As it stands, the project currently leans itself more towards a research tool, rather than a tool to measure academic consensus. However, there are still more open-source models available to add value to what the project attempts to achieve.

CHAPTER 5
CONCLUSIONS

5.1 SUMMARY

This project set out in laying the groundwork, as proof of concept, a method that measures academic consensus and a research tool to gain insight into any academic field of interest. To achieve this the process required automation, from data collection, user interaction and results output.

For data collection, this project used only one API as minimum viable product, with the intention to scale to more APIs. The PLOS API allowed easy access to the Conclusions section of research papers, although rate limits imposed by the API slowed the data retrieval process.

As a test-case, the academic field of nutrition was chosen which contained over 1,500 articles in the PLOS API. Data exploration revealed that most of these articles were published after 2012, as the API gained popularity and confirmed the need for more APIs that included historic research papers.

Except for some minor punctuational errors, the data did not require much cleaning and further data preparation was straightforward. The project applied NLP models such as phrase modelling for topic classification, word and document vectorisation as well as summarisation. Topic modelling did not sufficiently contribute towards identifying common themes in the academic field of nutrition, which was necessary to identify academic consensus. However, word and document vectorisation, as well as summarisation, showed promise as a research tool in finding similar concepts and recommending similar research papers of relevance.

In conclusion, in the pursuit of creating a means to measure academic consensus, only a research tool presents immediate application. However, this is merely the first iteration towards the ultimate vision of measuring academic consensus. The next section offers recommendations for further development towards this goal.

5.2 FURTHER DEVELPOMENT

This project highlighted many opportunities for further development and improvement. Many of these suggestions were part of the original project proposal but due to time constraints and the increasing scope of the undertaking, could not be achieved in time.

5.2.1 Data Collection and Scope

The case for the requirement of more APIs have been made, especially APIs that include historic research papers. The scholarly publishing section in the MIT Libraries contains a list of 27 academic publishing APIs. A quick search on the Journal Storage website (JSTOR) alone, reveals a collections of over 20,000 articles related to nutrition and diet, of which over 15,000 articles were published before 2012. Unfortunately, the API does not allow searching of the Conclusions sections. Perhaps a similar model using data from the Abstract section of articles could produce significant results, since most scholarly APIs provide access to the Abstract.

Another challenge with using a scientific field, such as nutrition, is the lack of subject matter expertise that is essential in identifying topical themes. An academic field such as History could produce more intuitive insights that are easier to interpret.

5.2.2 Data Exploration

As a research tool there are some low hanging fruit which will be relatively easy to implement and improve functionality. To give more context to the data, some features could include a geographical map indicating where most articles originate from, or a breakdown of which journals published the most articles and by which authors. Maybe some authors could skew results by their bias from publishing many articles with the same theme. Not least, who funded the research and flag any conflict of interest.

A technical suggestion is to do exploratory analysis after completing phrase modelling, to better capture key conceptual words.

5.2.3 Data Modelling

i. Hierarchical Vectorisation Topic Labelling

Although not included in the report, the Jupyter Notebook (3_topic_modelling_visualisation.ipynb) contains a section on Hierarchical Clustering and a dendrogram containing all the words in the corpus dictionary. As expected, this graph is too large to visually present meaningful clusters of words. However, as each word is vectorised, given a certain hierarchical level, the average vector of a cluster of words could be calculated and be represented by the most relevant ‘similar’ word(s). In this way it will be possible to label clusters of words, or rather different topics or themes.

But the problem persists, how to effectively break down the optimal number of topics?

ii. Optimum Number of Topics

An article by Sooraj Subrahmannian published in Medium called "Learn to find topics in a text corpus", explains how to estimate the number of topics in a corpus. By evaluating the average coherence score per topic, for a range of models with different topic numbers, the optimum number of topics can be found. Figure 5.1 displays this graphically. In this case the optimum number of topics is between 12 and 14.

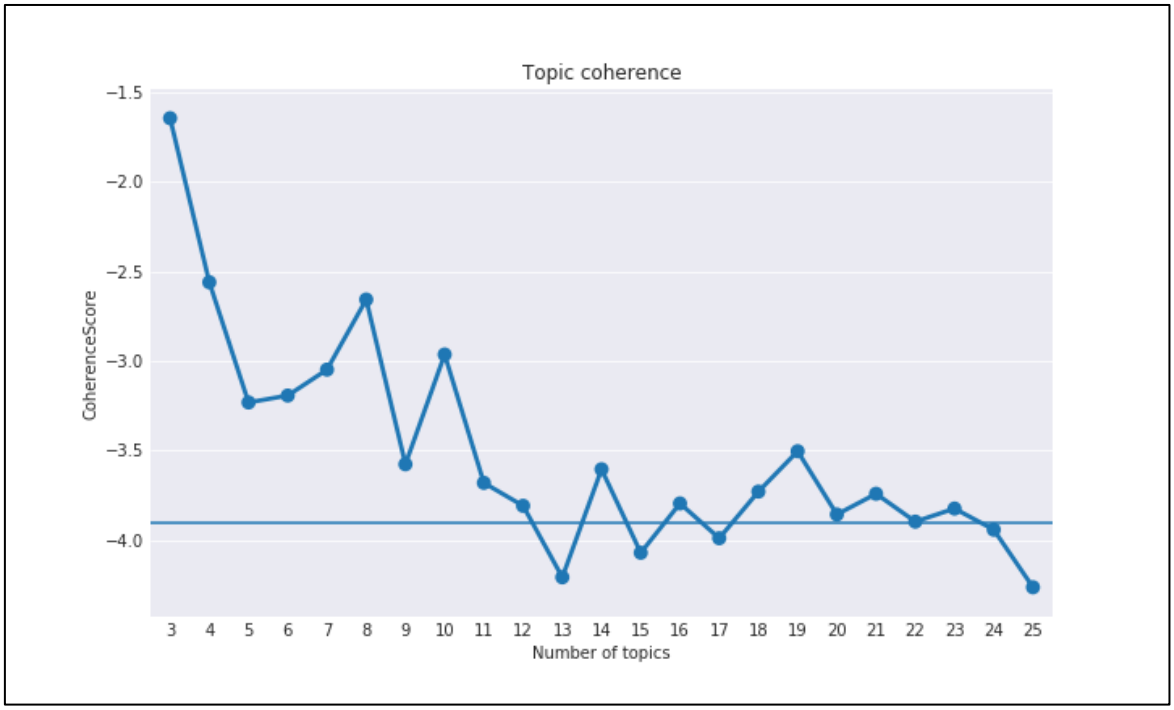

Figure 5.1: Optimum Number of topics

Source: Medium, 2018; Sooraj Subrahmannian.

In conclusion, the ultimate quest is to avoid subjective measuring of topic insight and somehow attempt to score the concept of understanding. Thus, with the optimum number of topics and a means to label it with hierarchical vectorisation, the distance between topics in the vector space can be measured for significance. This significance can be determined with hypothesis testing of one cluster versus many from the distribution of word vectors in the clusters. This is a significant leap towards determining academic consensus.

5.2.4 Project Integration and Production

Lastly, the original intent was to productionise and publish this project to a web page. It is still the intention of the author to see this through till the end and develop this project into a meaningful research tool that can finally quantify academic consensus in any scientific or academic field.