THE UNIVERSITY OF
MELBOURNE

Assignment 1: Analysis Report

Elements of Data Processing

Claire Tosolini

April 20, 2023

Contents

1 Introduction 2

2 Results 3

2.1 Task 4 3

2.2 Task 5a 3

2.3 Task 5b 4

3 Analysis 5

3.1 Task 4 5

3.2 Task 5 5

4 Discussion 6

5 Conclusion 6

1 Introduction

This assignment combined and applied lecture concepts, forming a data processing pipeline to extract information from the given source material.

- ***Crawling***: For two 'seed' (starting) URLs: links present on the seed page were added to a corresponding list. Any unique links found by visiting these retrieved links were also added.

  References for the purposes of this document:
  Seed URL 1: http://115.146.93.142/fullwiki/Gerard_Maley
  Seed URL 2: http://115.146.93.142/fullwiki/A12_scale

- ***Scraping and Pre-Processing***: Extracted main text from every retrieved link's page using BeautifulSoup. Tokenised through the following steps: case folding, non-alphabetic character removal, whitespace conversion, explicit token separation, stopword removal, removal of tokens less than two characters, Porter stemming.

- ***Bag of Words (BoW)***: A BoW representation, i.e. a collection of all words present, was produced for each page.

- ***Word Frequencies***: Extracted the top 10 most frequent tokens (words) from the BoW corresponding to each Seed URL.

- ***Dimensionality Reduction - Principal Component Analysis (PCA)***: Produced, Vectorized and Normalised the BoW representation of all tokens present across all retrieved pages. Applied PCA to reduce the dimension of this BoW, enabling analysis of the relationship between each page (i.e., article), with regards to its topic and its similarity to other pages.

2 Results

2.1 Task 4

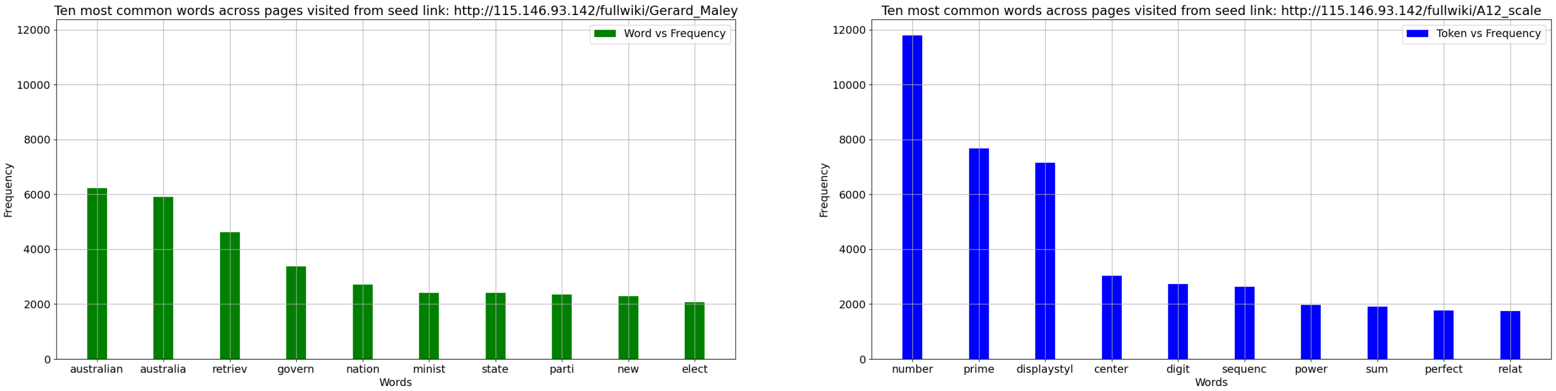

Figure 1: Ten Most Common Words corresponding to each Seed URL

2.2 Task 5a

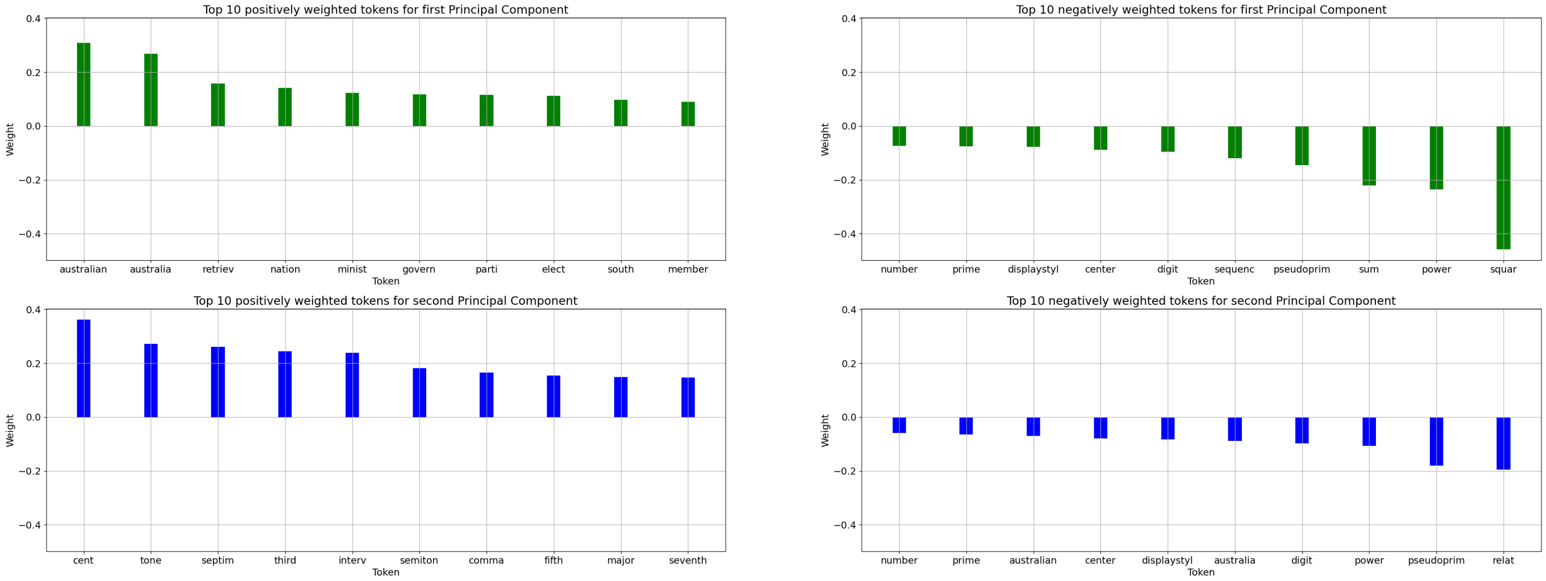

Figure 2: Ten Most Positively- and Negatively-Weighted Tokens for each Principal Component

2.3 Task 5b

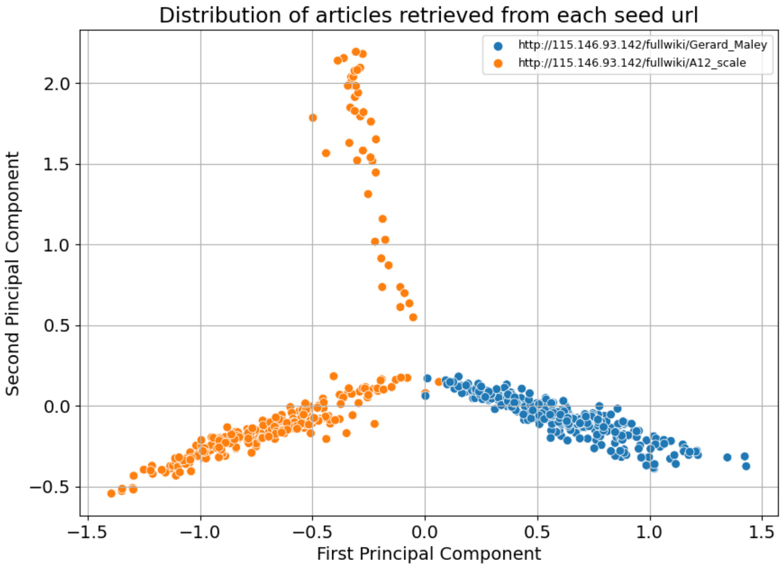

Figure 3: Principal Component Analysis (PCA) Article Distribution

3 Analysis

3.1 Task 4

Compare the ten most common words in each seed URL in Figure 1. Why might the differences be present?

There are no common words between the most frequent words for the two Seed URLs. The most likely explanation for this is that the **topic**, or contents, of the seed pages are significantly different.

3.2 Task 5

Interpret what words might not be surprising to find in articles for each Seed URL, based on the information in Figure 2.

Articles corresponding to Seed URL 1 appear to be related to Australian History and Politics; words such as *Aboriginal*, *senate*, *house*, *represent* etc., would not be surprising to observe. Similarly, since articles corresponding to Seed URL 2 appear to relate to Music and Numerical Values, finding words like *fourth*, *aural*, and *major* would be consistent.

Interpret the distribution of URLs shown in Figure 3. Do you think you could determine which Seed URL a new unseen link originated from when plotted in the 2D space after applying PCA?

PCA finds a new set of features, i.e. *principal components*, that retain correlations between datapoints in a dimension that is more observable. In this case, the original features of the articles were individual words present. By applying PCA and reducing the amount of features from total number of words (3606), to just two, what we observe in Figure 3 is how similar in topic each article (datapoint).

Dense clusters of datapoints represent articles with high similarity. Since the first Principal Component explains the most data variation, decreased distance along this axis reflects a greater article similarity than along the other axis.

Figure 3 reflects a very strong similarity between all articles originating from Seed URL 1, suggesting that the topic of the articles remains consistent. Articles corresponding to Seed URL 2 demonstrate a different correlation; two quite distinct clusters suggests the presence of two related but separate topics.

The consistent clustering of datapoints suggests it would be possible to accurately determine the Seed URL of an unseen link after applying PCA.

4 Discussion

What are the limitations of this dataset? What are the limitations of the processing techniques? What could be done in the future to provide further insights?

A significant limitation of the dataset is the fact that the Crawling undertaken was only partial. Retrieving more links for each Seed URL would increase the accuracy of their BoW representations, hence increasing the validity of comparisons.

Limitations are also present in the processing techniques. Casefolding, stopword removal, and loss of word order from Bag of Words representation can cause loss of contextual information. The Porter stemming algorithm may not always produce correct results, and the steps related to PCA also result in some information loss.

Along with further crawling of the Seed URLs, sources for further insight include:

- Lemmatisation instead of Stemming
- TF-IDF instead of Bag of Words
- K-Means instead of PCA

5 Conclusion

The data processing applied in this assignment produced informative results that could be used to accurately determine similarity of unseen articles to the assignment dataset. The process could also be useful in applying to new datasets, or extending through addressing the limitations.

References

List of Figures

1 Ten Most Common Words corresponding to each Seed URL 3
2 Ten Most Positively- and Negatively-Weighted Tokens for each Principal Component 3
3 Principal Component Analysis (PCA) Article Distribution 4