COMP2008: Elements of data processing
Assignment 1

Task 6: Analysis Report

The principal task in the assignment one was to apply some of the general knowledge taught during the first part of the semester in a practical project. Thus, it has been proposed to apply Natural Language Processing (NLP) over information crawled and scraped from two specific seed URLs looking for some relations among the original pages and the NLP results. The seed pages are shown and briefly defined below:

A12_scale: "A12 is a non-octave-repeating scale or musical tuning featuring twelve steps to the tritave".
Gerard_Maley: "Australian politician who serves as MP for the Country Liberal Party in the Northern Territory".

The results of the first approach of text processing, where each child URL page text was gathered, pre-processed, tokenized and aggregated in relation with the seed page is shown below:

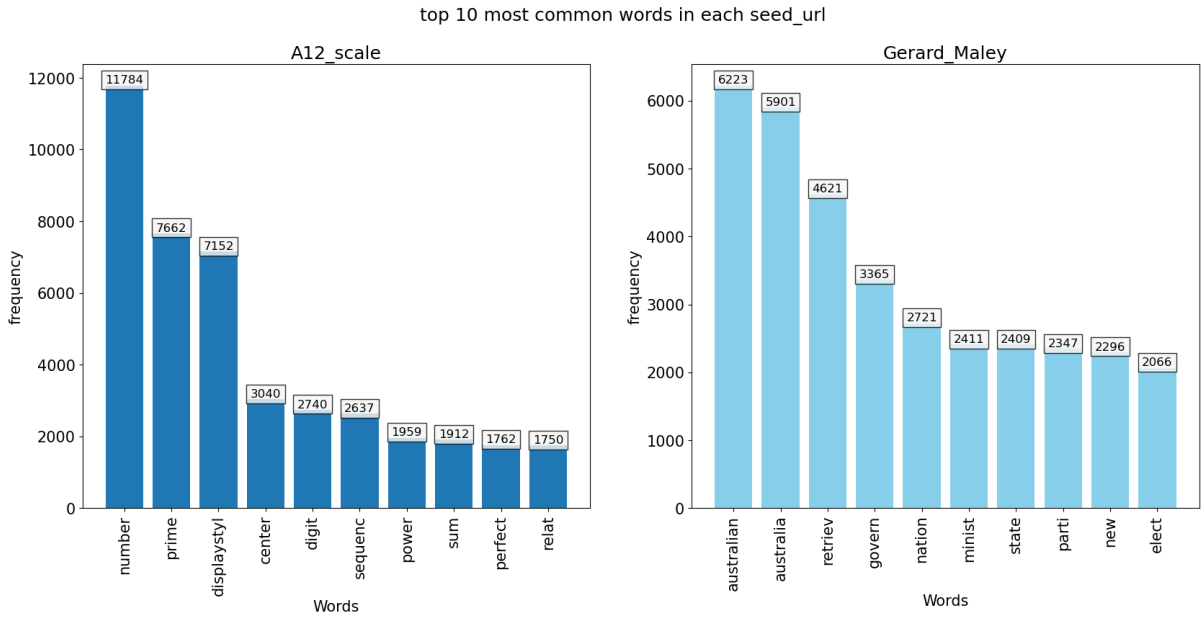

***Figure 1**: Top 10 most common features in each seed URL*

The information shown in the figure, highlights the top 10 more frequent words in the children URLs in relation with the seed URLs. Here we can find that for the **A12_scale** seed page the most frequent words are related to mathematical terms such as "number", "prime" and "digit". In the other hand, the most repeated words in the **Gerard_Maley** seed page are widely related with politics and Australia, with words such as "australian", "govern", "minist", "parti". These differences are driven by the seed URLs topics, where on each of the web pages we can see the most repeated words relative to them.

As a first approach the tokenization with words steaming shows to be functional at gathering relevant information from the web pages, however, it is not yet useful to know the specific topic of the different web pages contained within the seed pages. For this purpose, we apply a vectorization method over the corpus of each of the URL's followed by a dimensionality reduction algorithm, in this case, we have been asked to apply a BOW vectorization which returns a sparse matrix with the counting of the tokens appearances in each of the children pages, here, after applying a normalization method to rescale the results, we applied a PCA dimensionality reduction algorithm to get the 2 most important vector components out of the full vectorized list.

In the following figure, the top highest and lowest weighted features for each resulted PCA component are shown:

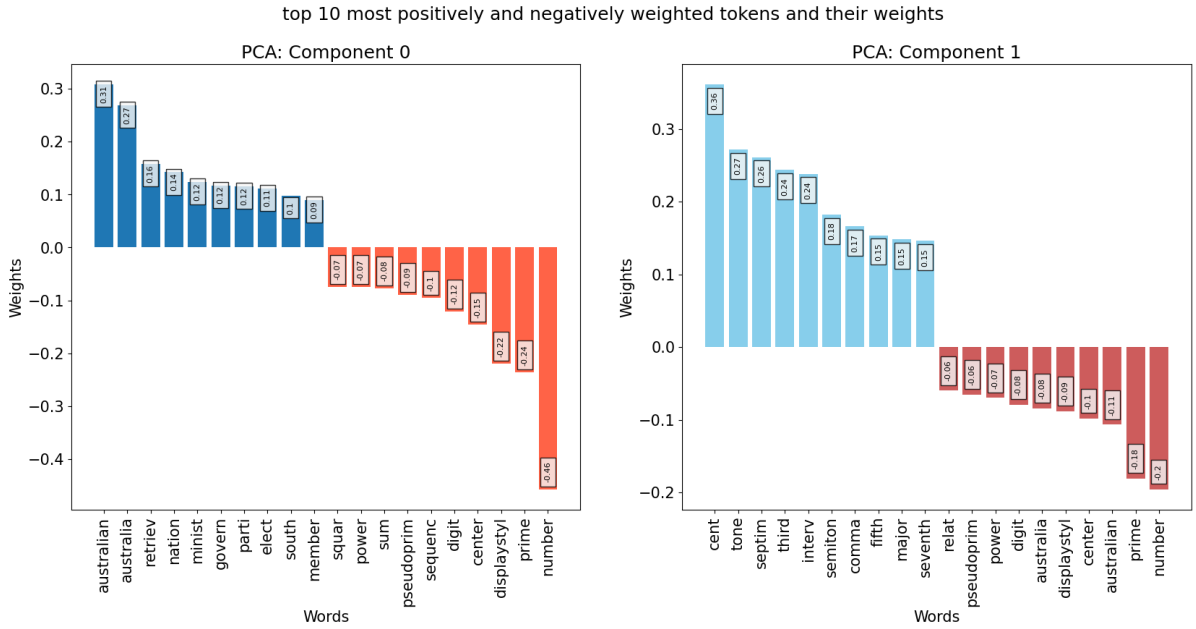

***Figure 2**: Top 10 most positively and negatively weighted tokens in each seed PCA component*

From the figure shown above, we can see that the First principal component, is weighting positively the words related to politics, and weighting negatively the words related to mathematics, where most of the words weighted in the respective category are directly related with the words classified by frequency in the Figure 1 graphic. In contrast, the Second principal component, is weighting positively some terms associated with the **A12_scale** that could be associated with musical scales and notations, while in the negative weight's component is mixing up terms from both Seed URL's.

Creating a graph of components weights in relation to the root URLs can give us more information about the interpretation of the PCA components.

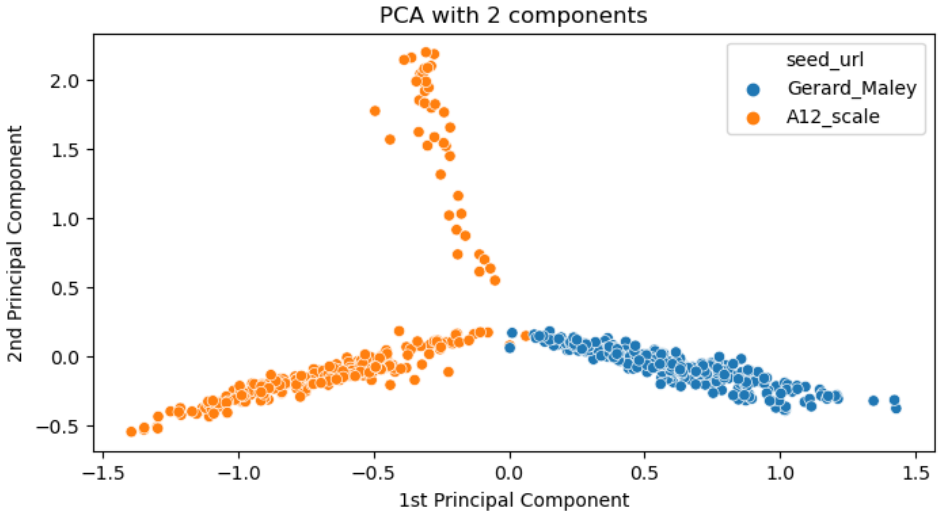

***Figure 3**: PCA components in the seed URLs*

In the figure above, and based on the insights from the figure 2, we can say that the first principal component of the PCA, seems to be great at dividing the topics among the seed URLs, conversely, the second component, does not differentiate the features in terms of theirs seed URL, however, it seems that this component is identifying another topic related to the **A12_scale** seed page, and as it was stated above, this topic is somehow related to musical scales and notations.

In conclusion, we can see that the **A12_scale** tokens are related to music and mathematical terms as we could expect, meanwhile, the **Gerard_Maley** tokens are more about politics, and state terms. It is also true, that the PCA algorithm 2D plotting in the Figure 3, can easily help us know to determine the association with the seed link for new unseen links.

Further Analysis

For further analysis, it might be positive to improve the dataset quality in terms of the text processing steps applied over it trying some different techniques than the applied so far, and applying some others that could improve interpretation. Some of the aspects that could provide further insights are:

- **Improve text extraction.** The HTML source of the web pages can change between one another, as it was the case in our solution, where some specific CSS selector texts were added to our solution tokens due to the exclusion rules did not filter them out. In relation to this, it is necessary to fully understand the text structure we are extracting from the HTML page in order to have a cleaner dataset.
- **Use different text processing approaches.** The lemmatization method applied over the dataset can change widely the results and interpretation of them, in this case the lemmatizing method can be better than stemming for word tokenization since it maps the words to its root forms rather than just chop words to produce root forms as the Stemmer does. Other useful hint could be to use ngrams to find hidden patterns within the data.
- **Use different vectorization techniques.** The use of the td.idf method, can also be useful to classify the specific topics that differentiate of the children pages of each seed among them.
- **Apply a classification algorithm.** As we found by applying the PCA algorithm, there are 3 visual groups in the figure 3, so apply a classification method can be helpful to identify the topics of each cluster.