Text Summarization - An Extractive Method

Dinesh Kumar Dhamotharan, Harsh Rajesh Darji, Namitha Chandrashekaraiah Reddy, Sakthi Priya Rajendran and Yesaswi Avula

Syracuse University

Problem Statement

- With big amount of data circulating in the digital space and very less time, there is a great need to reduce much of this text data to shorter, focused summaries that capture the salient details, both so we can navigate it more effectively as well as check whether the larger documents contain the information that we are looking for
- This summarization will not be manually possible for this data volume and hence needs automatic methods / algorithms
- The algorithms that could be developed for this purpose would reduce time and speeds up the process of researching for information, and increases the amount of information that can fit in an area

Objectives

- To get familiarized with different text summarization methods
- To apply extractive text summarization on a large text to identify the most important information from the given text and present it

Approaches

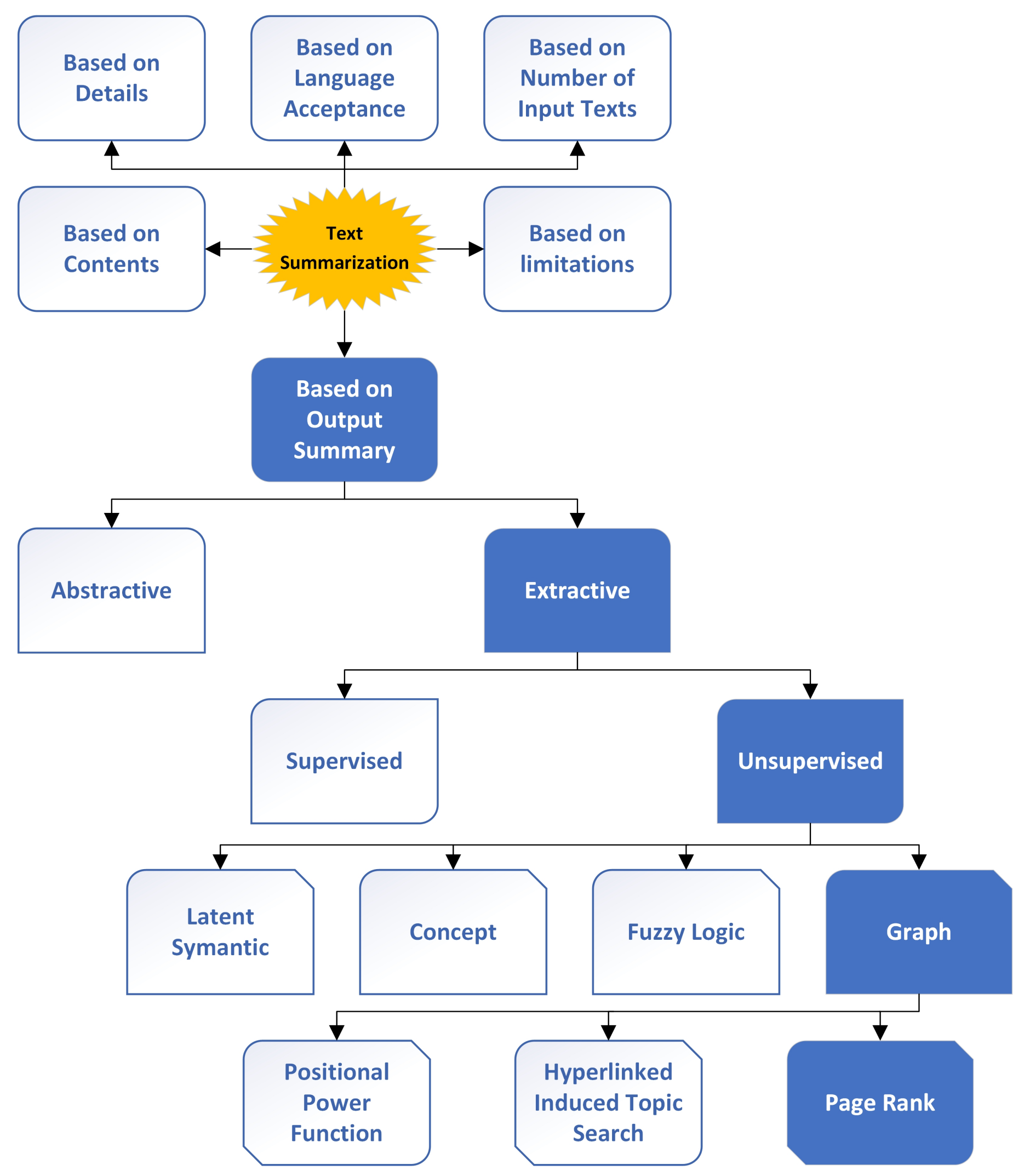

Figure 1: Text Summarization Approaches

Algorithm Flow

1. Split the text into individual sentences
2. Find vector representation (word embeddings) for each and every sentence
3. Calculate and store similarities between sentence vectors in a matrix
4. Convert similarity matrix into a graph, with sentences as vertices and similarity scores as edges, for sentence rank calculation
5. Finally, a certain number of top-ranked sentences form the final summary

Algorithm Flow

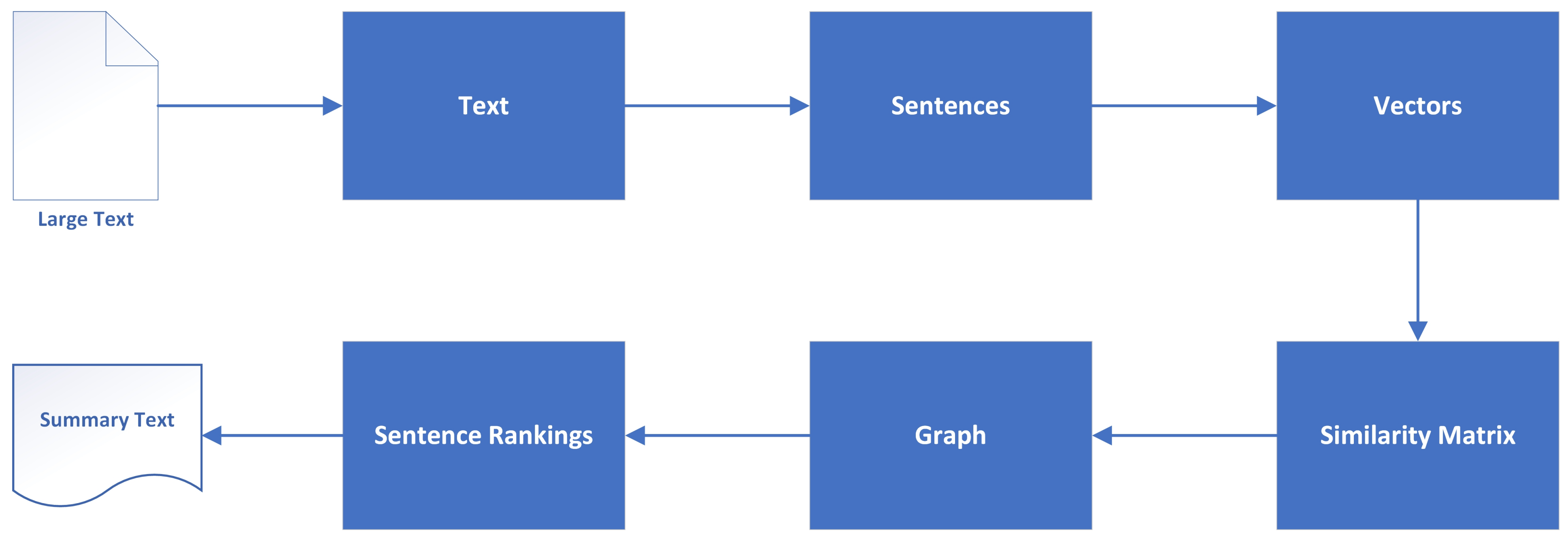

Figure 2: Algorithm

Input

The Brown Corpus - Chapter 2 text has been given as input for summarization:

Figure 3: Input - Brown Corpus

Output

Following is the output summary represented as a word cloud

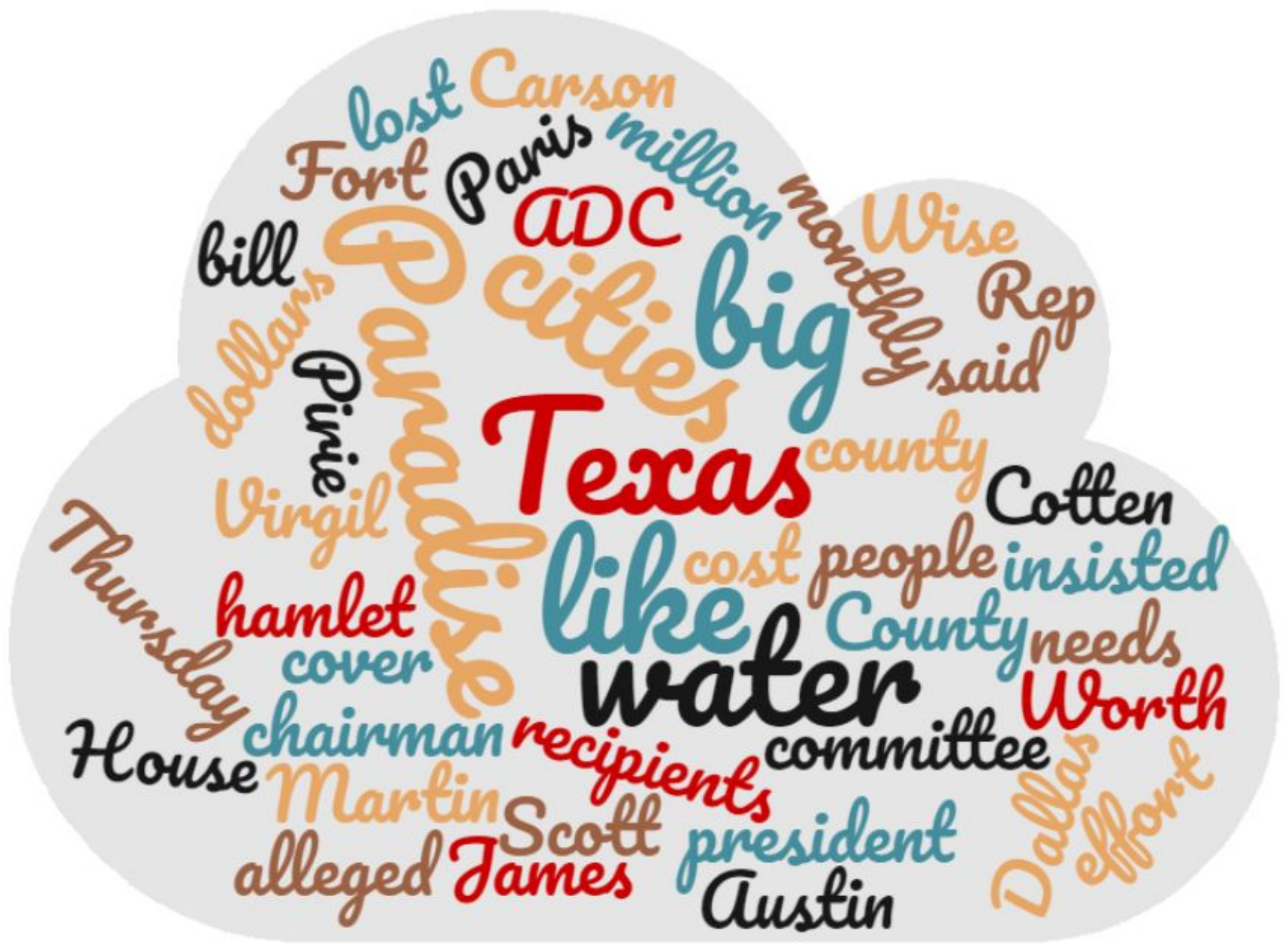

Figure 4: Output

Processing

Following is the graphical representation of the Sentence numbers vs their importance

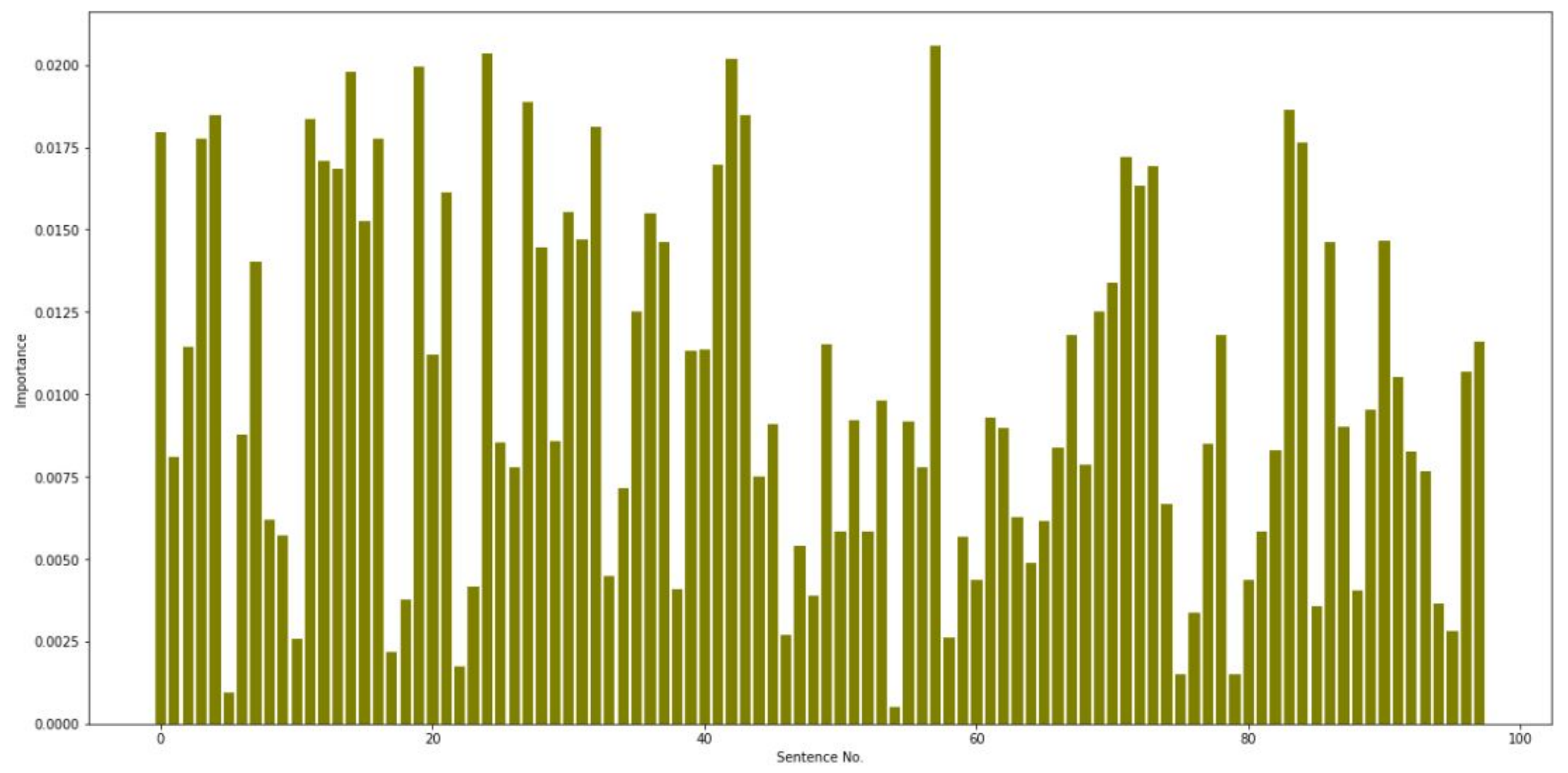

Figure 5: Sentences vs Importance

Conclusion

- The most important sentences give more information about the text as they have a higher score
- The algorithm is highly portable to other domains, genres or languages
- It checks the connections between the entities in a text and applies recommendation to them
- It is more simple and efficient than supervised learning techniques as the data do not need to be trained

Future Work

- Automatic text summarization is a crucial task in Natural Language Processing
- The task can further be done using the Fuzzy Logic approach and Latent Symantic approach
- Different ranking algorithms can be used for the selection of the important sentences and the accuracies of all the approaches can be compared

References

[1] Gyanit Singh Chandra Khatri and Nish Parikh. *Abstractive and Extractive Text Summarization using Document Context Vector and Recurrent Neural Networks.*