



GRAPH THEORETIC TEXT ANALYSIS

Under the Guidance of
Prof. Susanta Chakraborty

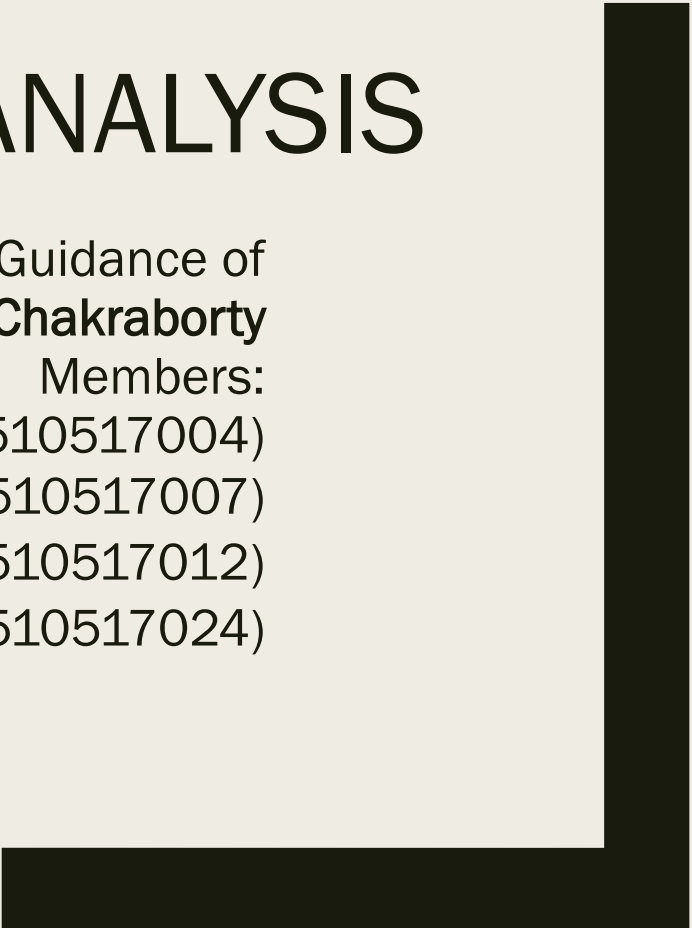
Members:

Arijit Saha (510517004)

Rohit Kumar (510517007)

Akash Singh (510517012)

Subhajyoti Saha (510517024)



Outline

- Introduction
- Motivation
- Objective
- Preliminaries
- Our Approach
- Example
- Application
- Conclusions
- Future Scope

INTRODUCTION



INTRODUCTION

- In recent years, the web has become an essential resource for obtaining information associated with any topic or domain.
- The amount of text produced by interactions on social media, blogs, URLs, etc., has made essential to use advanced techniques to be able to understand and obtain valuable patterns from these large volumes of data, which is mainly text.

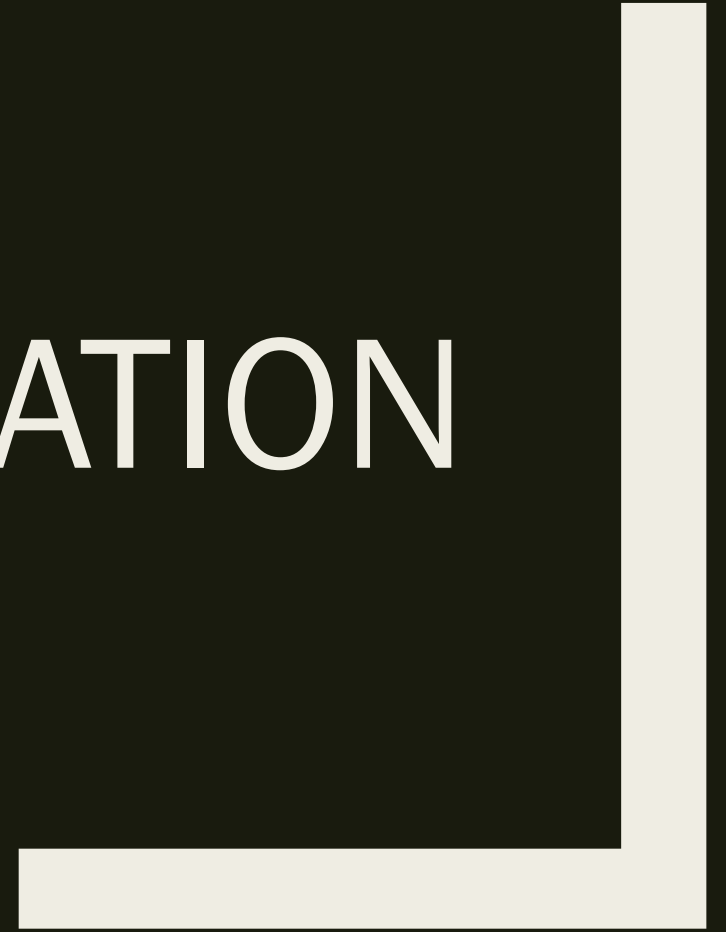
INTRODUCTION

- It is estimated that around 80% of all information is unstructured, with text being one of the most common types of unstructured data.
- Because of the messy nature of text, analyzing, understanding, organizing, and sorting through text data is hard and time-consuming so most companies fail to extract value from that.

INTRODUCTION

- By using graph theoretic text analysis, we can structure business information such as email, legal documents, web pages, chat conversations, and social media messages in a fast and cost-effective way.
- This allows us to save time when analyzing text data, help inform business decisions, and automate business processes.

MOTIVATION



MOTIVATION

- The motivation of this analysis is to show how co-occurrence graphs can be used to represent text documents in a practical manner independently of the text analysis task.
- To show this kind of representation can be a valuable asset to extract features/patterns, due to its structural simplicity.

APPLICATIONS OF TEXT ANALYSIS

- Finding representative keywords in text.
- Text Summarisation
- Finding keywords for tagging News Articles.
- DNA Analysis
- Text Classification

OBJECTIVE



OBJECTIVE

- Determination of Important nodes of a text co-occurrence graph generated using n-gram model.
- Applying algorithm to literary text to detect words of significance to check credibility.
- Applying algorithm to news articles to find keywords for tagging and sorting to check credibility.

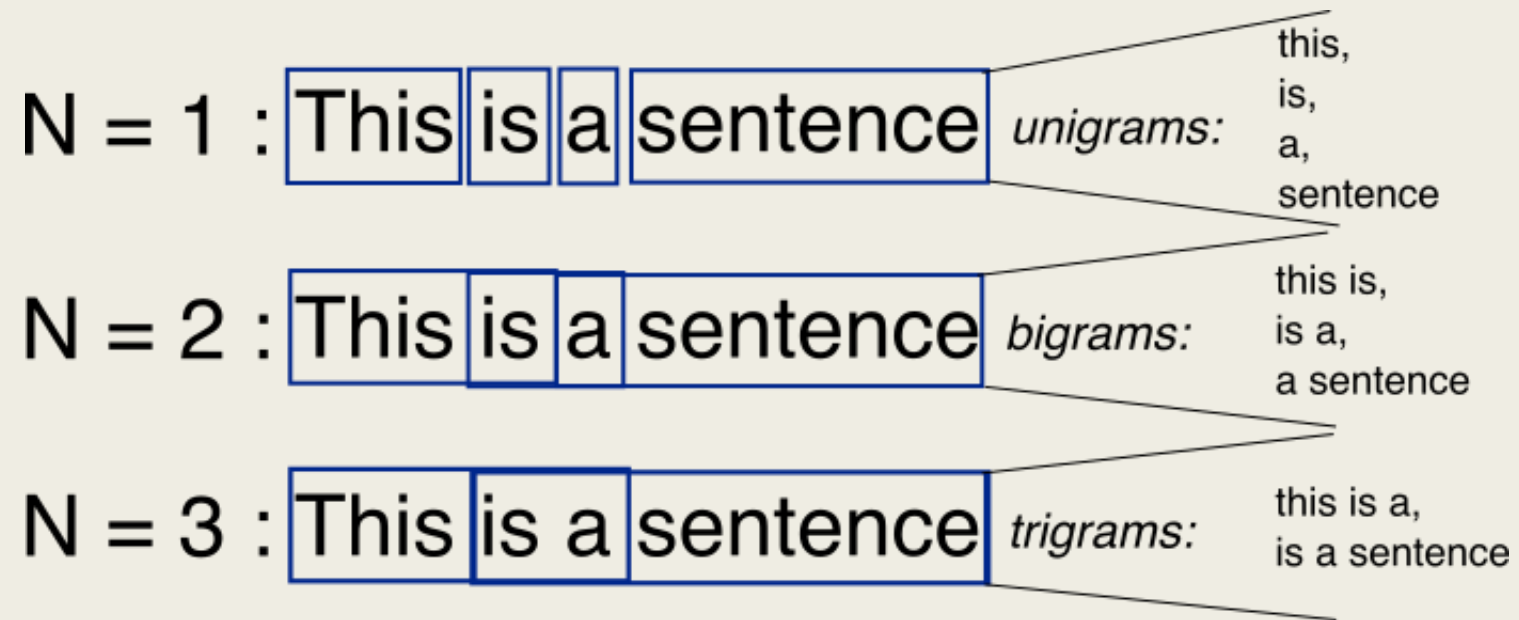
PRELIMINARIES



N-gram

- In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text or speech.
- The items can be syllables, letters, words or base pairs according to the application.
- The n-grams typically are collected from a text or speech corpus.
- An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram"; size 3 is a "trigram".

N-gram



NLTK Text Processing

- We have used the Natural Language Tool Kit (NLTK) Text Processing Library available in python for pre-processing the text beforehand.
- We have used the following features:
 - *Wordnet*
 - *Tokenization*
 - *Stemming*
 - *Lemmatization*

WORDNET

- It is a Lexical Database for the English language, which was created by Princeton, and is part of the NLTK corpus.
- It is used alongside the NLTK module to find the meanings of words, synonyms, antonyms, and more.

TOKENIZATION

- Tokenization is the process by which big quantity of text is divided into smaller parts called tokens.
- These tokens are very useful for finding patterns in the text so that text analysis can be performed on them.
- Tokenization is also considered as is considered as a base step for stemming and lemmatization.

STEMMING

- Stemming is the process of reducing inflected words to their word stem, base or root form.
- The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.
- We have used the Porter Stemmer Algorithm.
- *For Example, the words argue, argued, argues, arguing, and argus are reduced to the stem argu.*

LEMMATIZATION

- Lemmatization is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.
- Unlike stemming, lemmatisation attempts to select the correct lemma depending on the context.
- The word "*better*" has "*good*" as its lemma. This link is missed by stemming.
- We have primarily used this for Synonym Replacement to compress the vocabulary of the text.

OUR APPROACH



OUR APPROACH

- First, the text was tokenised in forms of sentences which in turn were tokenised further into words.
- Words were converted into lowercase letters.
- The word tokens were lemmatised, according to the Parts of Speech Tag.
- Stop-words (like *the*, *of*, *and*, *or*, etc) and punctuations were removed.

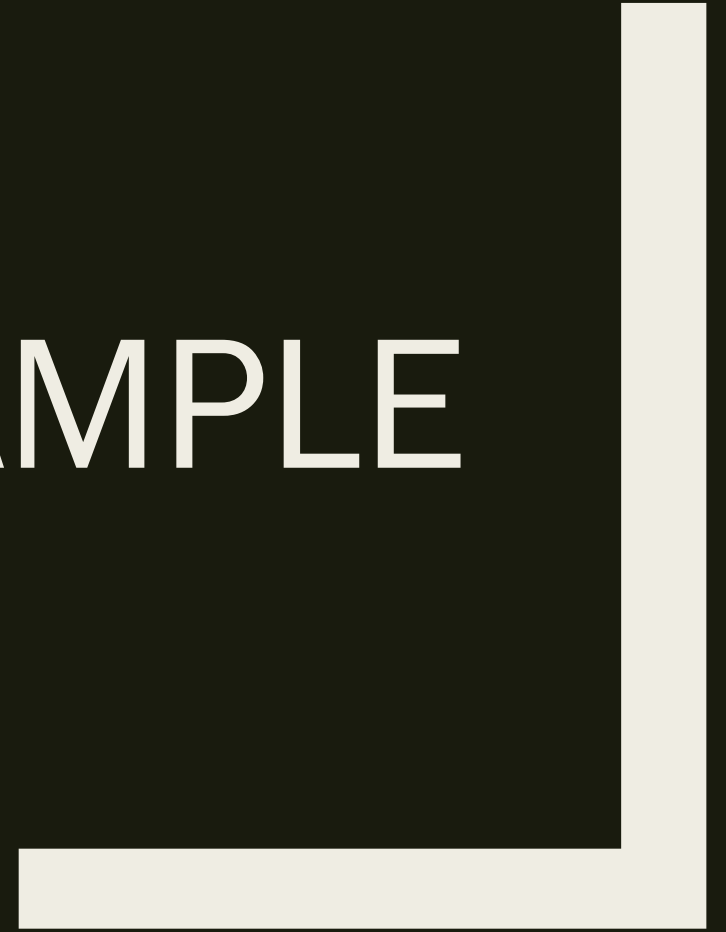
OUR APPROACH

- Words were replaced by their synonyms according to word frequency in that particular text by employing the synsets from NLTK Wordnet Library.
- Tri-gram was generated from the text.
- An undirected graph was created on the basis of the Tri-gram where each Tri-gram was made into a fully connected component.
- Any edge added twice was merged into original edge with increased weight.

OUR APPROACH

- The centrality factors – Betweenness Centrality, clustering Coefficient, Degree Centrality were calculated in the resultant graph.
- The most influential (top 10%) word-nodes were taken, on a weighted voting system of the above factors.

AN EXAMPLE



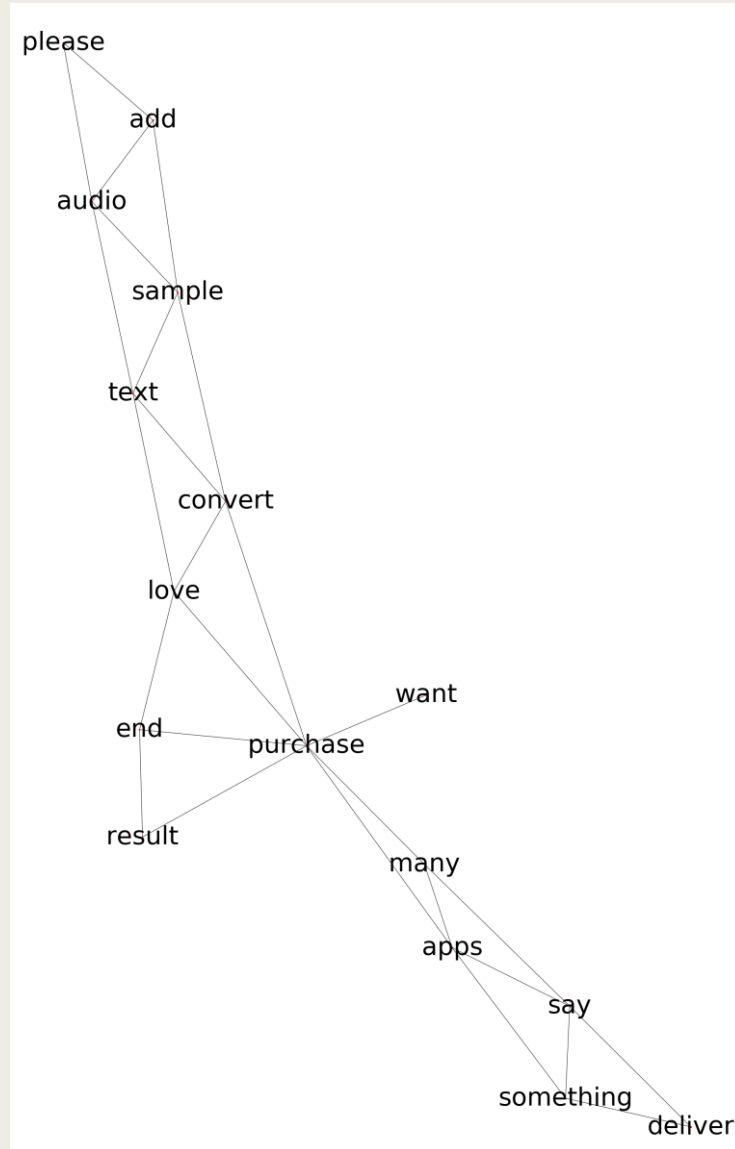
AN EXAMPLE

■ On Processing the text:

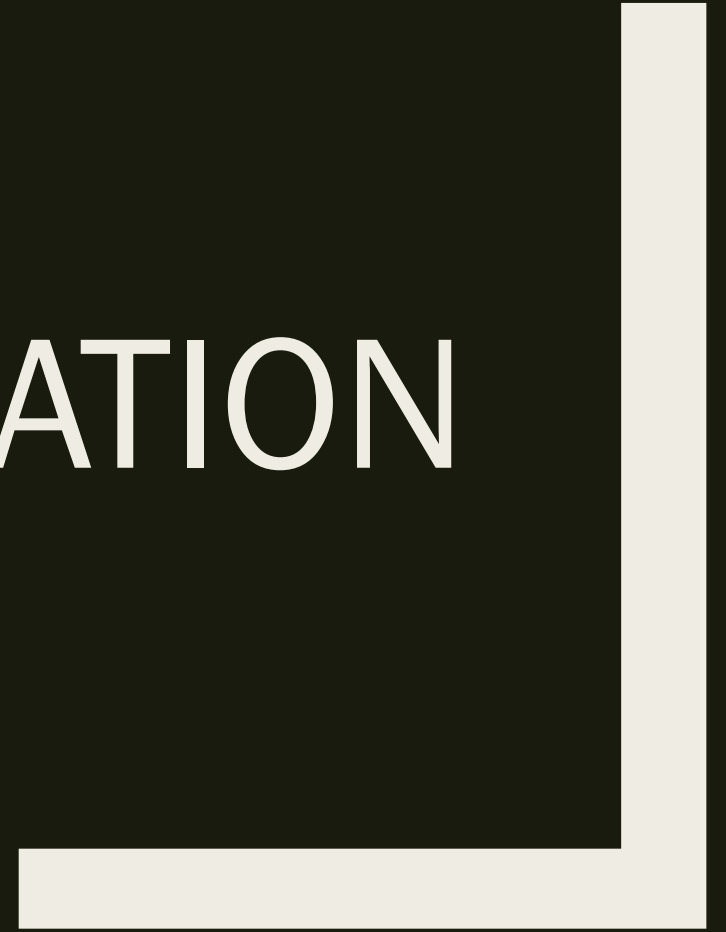
- *I would love to try or hear the sample audio your app can produce. I do not want to purchase, because I've purchased so many apps that say they do something and do not deliver. Can you please add audio samples with text you've converted? I'd love to purchase the end results.*

■ We get tokenised filtered text:

- *would love sample hear sample audio app produce*
- *want purchase purchase many apps say something deliver*
- *please add audio sample text convert*
- *love purchase end result*

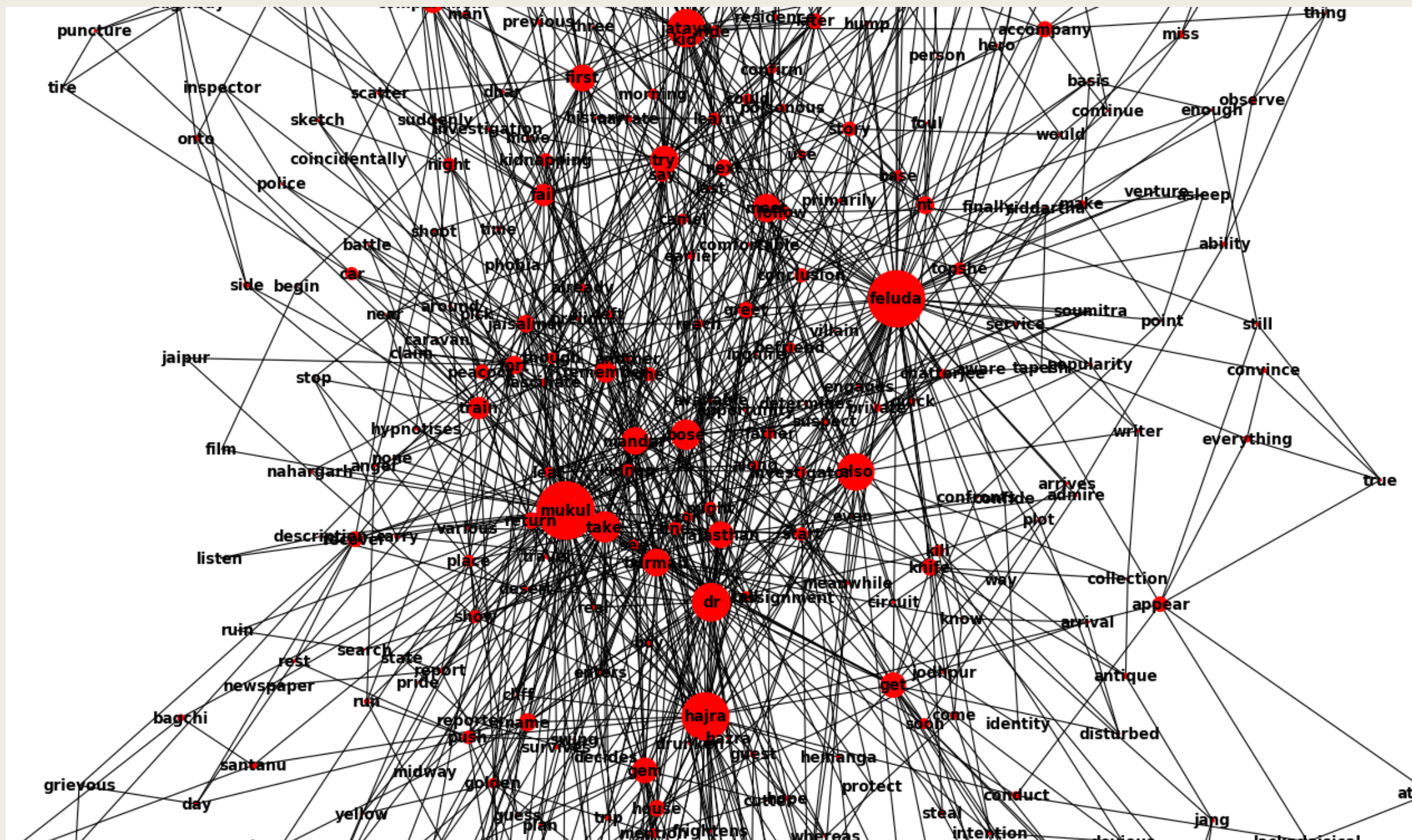


APPLICATION



APPLICATION IN LITERARY TEXTS

- We took the Story “*Sonar Kella*” by *Satyajit Ray* and applied our algorithm.
- Top 10 word Nodes after voting:
 - *mukul*
 - *hajra*
 - *dr*
 - *mandar*
 - *Also*
 - *feluda*
 - *jatayu*
 - *bose*
 - *rajasthan*
 - *fort*
- A part of the text graph generated is on the next slide. Here, more important nodes are represented using larger circles.

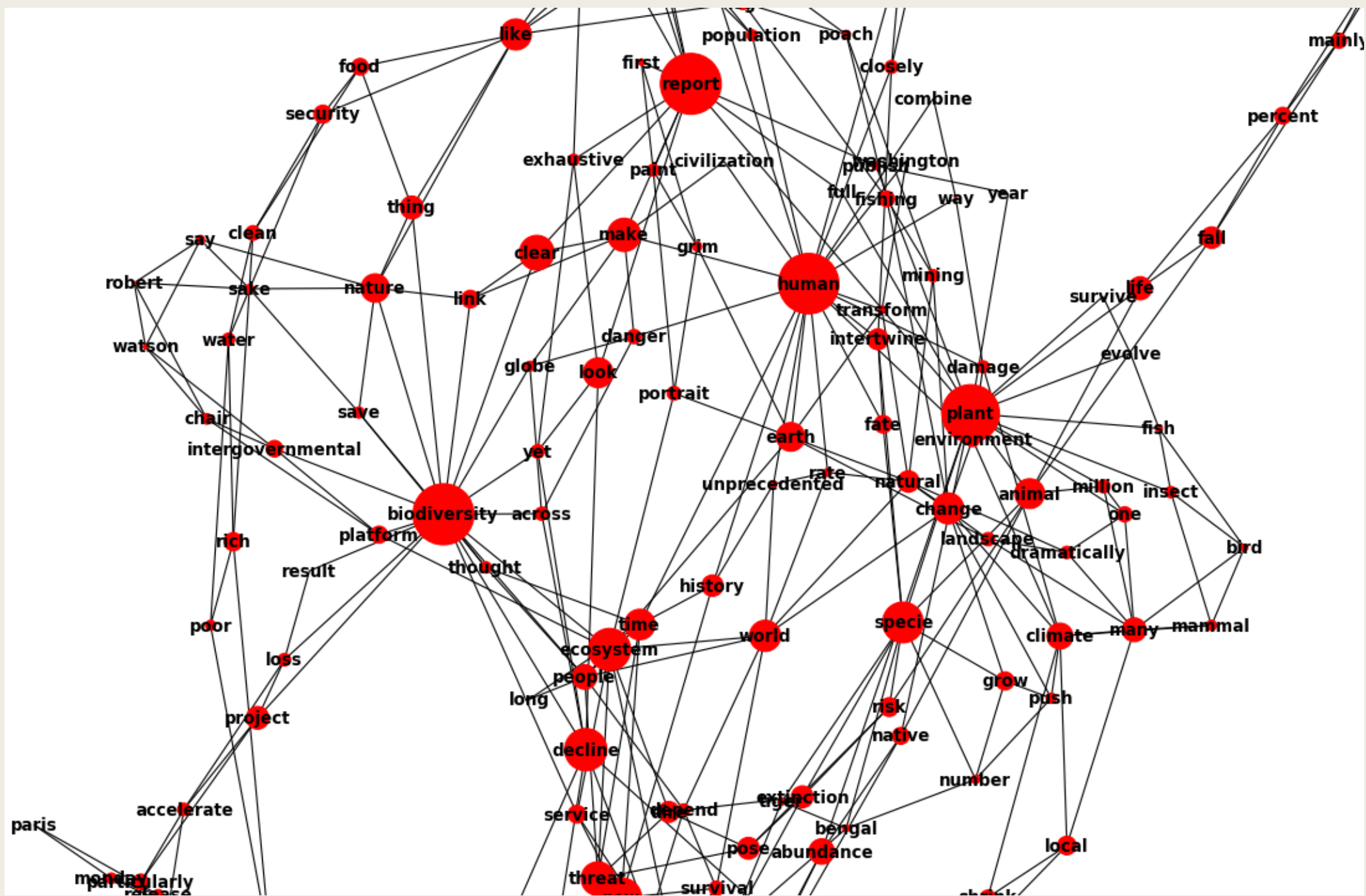


APPLICATION IN LITERARY TEXTS

- The words that are highlighted, are the keywords of the text. They are one of the following:
 - *Protagonist/Antagonist (eg. Mukul, Feluda, Hajra, etc.)*
 - *Geographical Location of Story Setting (eg. Rajasthan)*
 - *Other keywords (eg. Fort)*
- There are some discrepancies also as the words “*dr*” and “*also*” are also highlighted.

APPLICATION IN NEWS ARTICLES

- We applied our algorithm to news articles and got the results as shown in the next slide.
- The example news article here is - *Humans Are Speeding Extinction and Altering the Natural World at an 'Unprecedented' Pace.*
- The keywords generated are:
 - *Human*
 - *Biodiversity*
 - *Plant*
 - *Report*
 - *Threat*
 - *Decline*



APPLICATION IN NEWS ARTICLES

- The keywords this produces can be used to summarise the news articles into small keywords so that the news articles may be tagged for sorting and management.
- The tags can be associated to news articles, so that searching softwares might use these tags to pull up news articles easily.
- Since, the news articles are smaller in size, the graphs tend to produce skewed results sometimes too.

RESULTS AND CONCLUSIONS



Conclusions

- The results obtained show the relevance of graphs compared with traditional text analysis approaches that use the same dataset.
- Co-occurrence graphs with a window of 3 words (Tri-Gram) works the best in comparison to lesser grams, which gives back very general keywords or higher grams, that tend to give more irrelevant keywords.
- Betweenness Centrality gives the best results as a node with high betweenness acts central to the flow of meaning in a text.

CONCLUSIONS

- Degree Centrality gives the next best result, as a node with high degree represents more connectivity with other words in the graph.
- Clustering Coefficient gives worst result as it is a measure of the surroundings rather than the node itself.

Limitations

- If the text size is very less, then it may result in skewed results.
- The Synonym replacement process is not perfect yet, and it produces undesired replacements in some cases.
- Very rich literature or archaic literature poses challenges as it becomes very tedious to handle words that are not present in the NLTK library.

FUTURE SCOPE



Future Scope

- Considering the growing amount of information in social media, the creation of proposed graphs will require tools capable of handling this kind of large datasets (Big Data analytics).
- Incorporation of Machine learning models to predict genre of text, and build more powerful classifiers of text.

References

1. **Esteban Castillo, Ofelia Cervantes, Darnes Vilarino (2017).** Computación y Sistemas, Vol. 21, No. 4, 2017, pp. 581–599 doi: 10.13053/CyS-21-4-2551
2. David A. Bader and Kamesh Madduri. Parallel algorithms for evaluating centrality indices in real-world networks. In ICPP, pages 539–550. IEEE Computer Society, 2006.
3. P. Hage and F. Harary. Eccentricity and centrality in networks. Social Networks, 17:57–63, 1995
4. Nieminen, Juhani. (1974). On centrality in a graph. Scandinavian journal of psychology. 15. 332-6. 10.1111/j.1467-9450.1974.tb00598.x.
5. L. C. Freeman. A set of measures of centrality based on betweenness. Sociometry, 40:35–41, 1977.

References

6. S. Murai, "Theoretically and Empirically High Quality Estimation of Closeness Centrality," 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, 2017, pp. 985-990.
doi: 10.1109/ICDM.2017.126
7. S. Muhuri, S. Chakraborty and S. K. Setua, "An Edge Contribution-Based Approach to Identify Influential Nodes from Online Social Networks," 2016 IEEE International Symposium on Nanoelectronic and Information Systems (iNIS), Gwalior, 2016, pp. 155-160.
8. M. Li and C. O'Riordan, "The effect of clustering coefficient and node degree on the robustness of cooperation," 2013 IEEE Congress on Evolutionary Computation, Cancun, 2013, pp. 2833-2839.
doi: 10.1109/CEC.2013.6557913
9. Ulrik Brandes, "A faster algorithm for betweenness centrality", Journal of Mathematical Sociology, 25(2):163–177, 2001.