# *Probability Based Theme Detection*

## *Using Markov Chain*

***Abstract—In this paper we propose a model to extract themes from large corpus of text data. Suggested technique use Markov Chain and rudimentary probability based analysis. Markov processes are widely used to generate sequences that imitate a given style. The results obtained have been discussed at the end of the paper. Most probable phrases were generated with the help of Markov Chain algorithm. 1-Gram and 2-Gram models were used. Some interesting 3 word phrases were obtained which reflected the theme the large text data.***

***Keywords—Markov Chain; Bayes Theorem; Theme Detection; Stochastic Approach;1-gram and 2-gram model,***

## I. INTRODUCTION

Probability based theme detection could be used on documents to determine the most probable phrases. These phrases reflect the theme of the document. These phrases generated can be used to generate more documents with similar themes. This technique could be used to assign topic to the documents which would help in further clustering of text data. Instead of reading the whole document one could get a brief idea by reading these phrases. This technique could be used to determine most important sentences of a document. Documents with similar themes could then be classified into different clusters. In the field of medicine this technique could be used to get automated suggestions on medicines based on medicines prescribed by other doctors in a given period of time and for a given region. We could automatically determine trending news topics by looking at current news. We worked on IMDB movie review dataset to determine the most probable phrases. Markov chains are a powerful, widely-used technique to analyze and generate sequences that imitate a given style with applications to many areas of automatic content generation such as music, text, line drawing and more generally any kind of sequential data. A typical use of such models is to generate novel sequences that "look" like or "sound" like the original. In practice, higher order models offer a better compromise between expressivity and representation cost. Indeed, increasing the Markov order produces sequences that replicate larger chunks of the original corpus, thereby improving the impression of style imitation.

## II. LITERATURE REVIEW

In (**I**), the author has proposed a data-driven method for concept-to-text generation, the task of automatically producing textual output from non-linguistic input. He defined a probabilistic context-free grammar that described the structure of the input (a corpus of database records and text describing some of them) and represented it compactly as a weighted hyper-graph. We have produced three word phrases from text data which depicts the theme of the entire text data.

In (**II**), the author has presented an approach to generate summaries, a hidden Markov model that judges the likelihood that each sentence should be contained in the summary. He then compared the results of this method with summaries generated by humans, showing that he obtained significantly higher agreement than the earlier methods. We have tried to form three word phrases based on the probability of occurrence of the third word along with the previous two words. We achieved this using Baye's theorem along with the concept of Markov Chain rule.

In (**III**), the author has introduced an innovative approach for analyzing textual content that conveys multiple themes. It focused on efficiently segregating the context switches in text, and then accurately mining the different opinions present. He utilized three categories of features namely positional, lexical-semantic and sentiment polarity for theme co-referent text segmentation within a document. His experimental results demonstrated the proficiency of the proposed scheme to segregate textual content by themes, identify meaningful topics inherent in the themes, compute the polarity, and also suggest the applicability of the method to query-based retrieval systems. We have used the concept of Text Mining in order to retrieve high quality information from the text.

In **(IV),** the author has discussed the main facets and directions in designing error detection and repairing techniques which is required for cleaning the text. He proposed a taxonomy of current anomaly detection techniques, including error types, the automation of the detection process, and error propagation. He also proposed a taxonomy of current data repairing techniques, including the repair target, the automation of the repair process, and the update model. He concluded his research by highlighting current trends in "big data" cleaning. We applied similar process in order to clean the given text, like, removing special characters and digits, converting all characters to lower case and then tokenizing them.

## III. Working Methodology

**Markov chain** is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules. The defining characteristic of a Markov chain is that no matter *how* the process arrived at its present state, the possible future states are fixed. The **state space**, or set of all possible states, can be anything: letters, words, numbers, weather conditions, baseball scores, or stock performances. We have used two models, 1-gram and 2-gram. In 1-gram model we start with a given word and select the most probable word which is likely to follow the previous given word. In 2-gram model we chose the next word based upon the previous two words. Idea is very similar to Bayes Theorem of Probability.

The Bayes theorem describes the probability of an event based on the prior knowledge of the conditions that might be related to the event. If we know the conditional probability $P\left(\frac{A}{B}\right)$, we can use the bayes rule to find out the reverse probabilities $P\left(\frac{B}{A}\right)$ .

$$P\left(\frac{A}{B}\right) = \frac{P\ (A \cap B)}{P\ (B)}$$

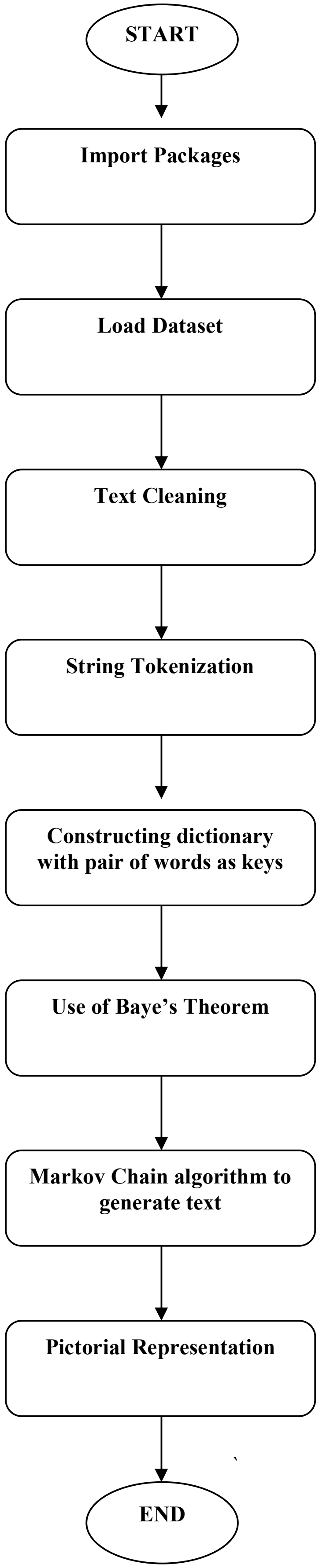
$$P\left(\frac{B}{A}\right) = \frac{P\ (A \cap B)}{P\ (A)}$$

$$P\ (A \cap B) = P\left(\frac{A}{B}\right) * P\ (B) = P\left(\frac{B}{A}\right) * P\ (A)$$

$$P\left(\frac{B}{A}\right) = P\left(\frac{A}{B}\right) * \frac{P\ (B)}{P\ (A)}$$

The above statement is the general representation of the Bayes rule. We can generalize the formula further.If multiple events $A_i$ form an exhaustive set with another event B.We can write the equation as

$$P\ (A_i/B) = \frac{P(B|Ai) * P(Ai)}{\sum(i=1\ to\ n)\ P(B|Ai) * P(Ai)}$$



The entire process is divided into various steps. Each step would be described separately for understanding of reader.

First step consists of importing python libraries such as numpy, pandas, matplotlib, re and wordcloud. Next we import the IMDB review dataset. We then use pandas to store the reviews in dataframes.

Next step consists of text formatting and cleaning the reviews. We will take help of Natural Language Toolkit Library ( nltk ) of python. We first removed special characters and digits. We then converted the text into lower case. We then split the text into lists of different words.

We now construct a dictionary with words occurring in texts as keys. The words which follow a given key are added to the list of that particular key. This will help us in choosing the most probable word following a word when it already occurs in text. This is 1-gram model.

In 2-gram model keys used are pairs of two words. These two words occur simultaneously in text. Now we add all the words which occur only when these two words were already present in the text. This follows Bayes Theorem of Probability.

In next step we generate text based on the above principles. We then considered keys one by one and then chose the most probable word following the present word in 1-gram model. In case of 2-gram model we chose the most probable word when previous two words are already present in the text. Now we have generated phrases.

## IV. Result

Some meaningful phrases were obtained after the process. These phrases reflect the theme of document in a precise manner. By looking at these phrases one could determine the brief idea which the following texts hold. Even though the dataset we used did not contain large texts we were able to determine significant phrases of three words. Some phrases had frequency of more than 8. Highest frequency of phrase was 34.

We categorize the phrases based on their frequencies and finally plot them on a bar graph. We have also plotted a word cloud to give a better visual representation.

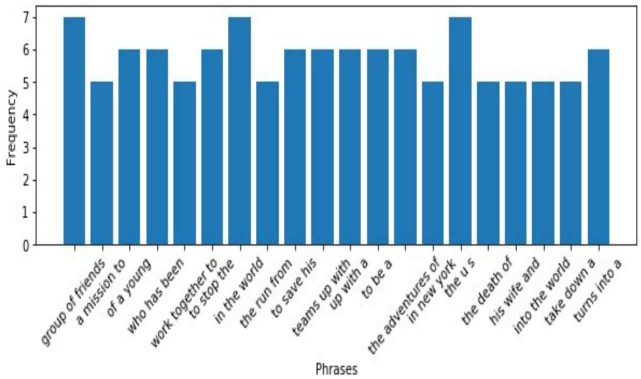


Fig 1: Bar graph representing some words with frequency between 5 and 7.

Some of the phrases with frequencies greater than 8 are:

| **Phrases** | **Frequencies** |
|---|---|
| A group of | 34 |
| In order to | 32 |
| The story of | 22 |
| A young women | 18 |
| On the run | 13 |
| A series of | 13 |
| With the help | 12 |
| The help of | 12 |
| To find the | 11 |
| The world of | 9 |
| A team of | 9 |
| New york City | 9 |

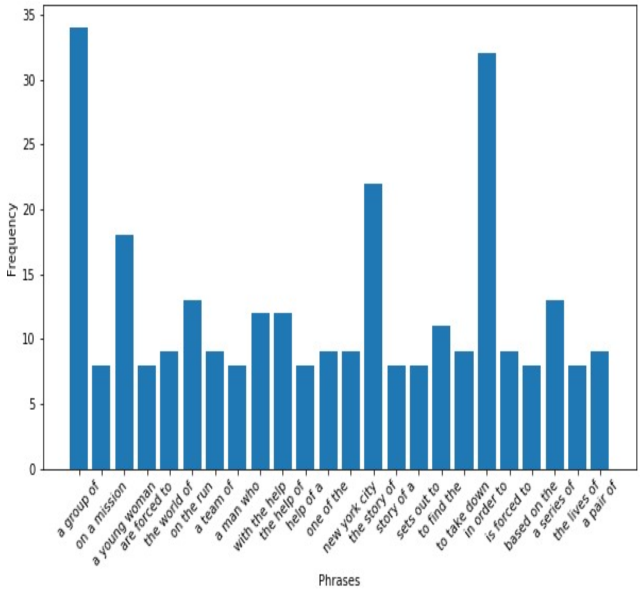


Fig 2: Bar graph representing words with frequencies greater than equal to 8.

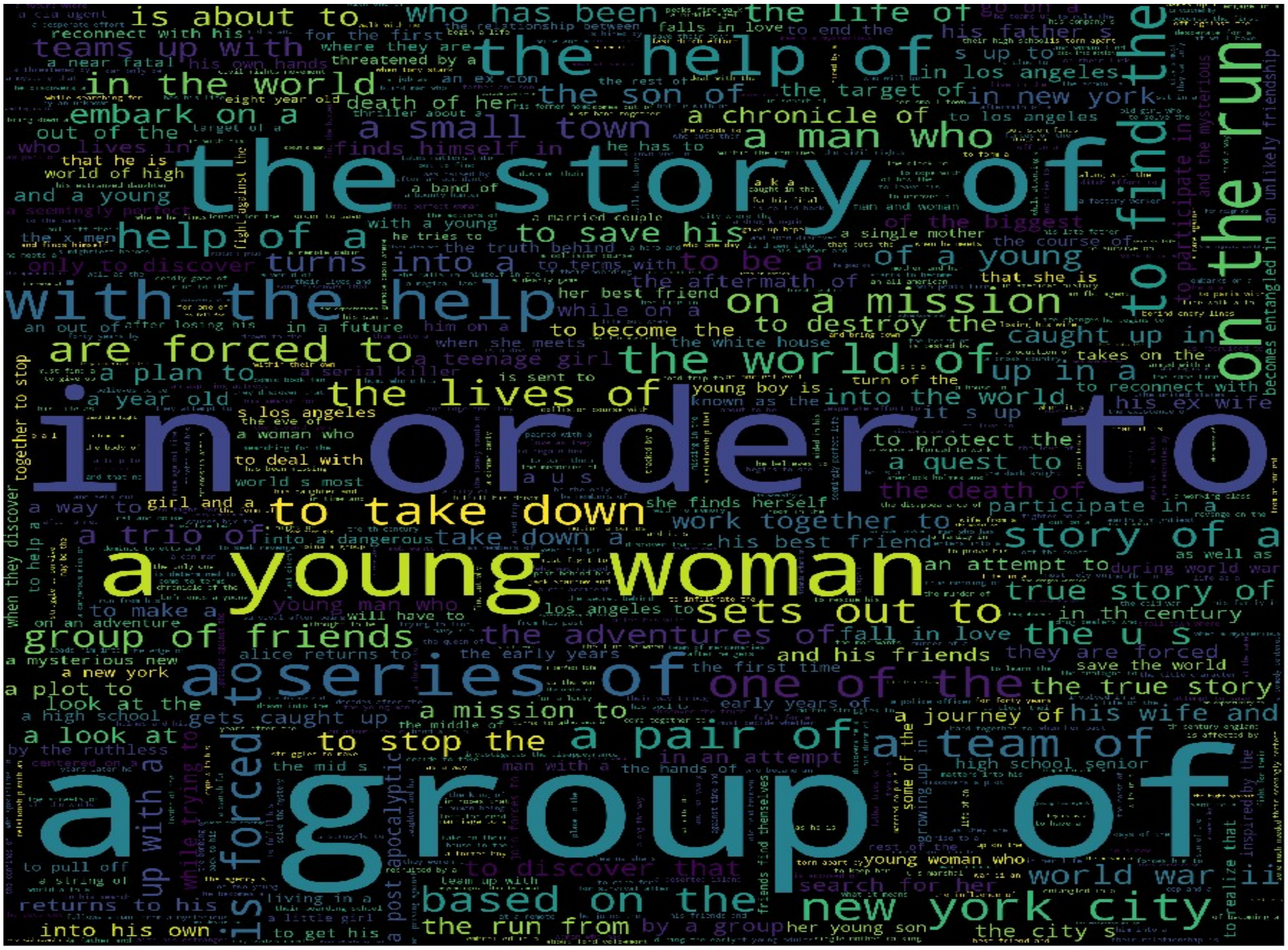


Fig 3: Word Cloud representing various phrases

## V. Conclusion

We could see that from text generation by Markov chain and probability, significant texts could be produced. It has wide applications in theme detection, determining the context of huge text documents, producing sample data sets with similar properties. Moreover it could be used in analysis of stream of documents to determine the trending topics. by doctors In the field of medicine this technique could be used to get automated suggestions based on medicines subscribed by other doctors. Our models could be improved by use of n-gram models where n is greater than 2. Sometimes the generated text show repetition of certain sequence of words. Not all starting keys can be used to produce the text. Many times meaningless texts are also produced. Increasing size of dataset could significantly improve the results. Results were quite significant but further research and improvement by changing of parameters could be done.

## VI. References

I. Concept-to-text Generation Via Discriminative Reranking-M Lapata, I Konstas.

II. Ext Summarization Via Hidden Markov Models-M Conroy, D. P. O'leary.

III. Tracking Context Switches in Text Documents and Its Application to Sentiment Analysis-S Sharma ,S chraverty.

IV. Trends In Cleaning Relational Data: Consistency And Deduplication- I F Ilyas