# NLP Project Report

Varun V 01FB15ECS341
Vinay G B 01FB15ECS350

## Extractive Summarization

Extractive summarization techniques produce summaries by choosing a subset of the sentences in the original text. These summaries contain the most important sentences of the input. Input can be a single document or multiple documents. The process can be divided into the following steps.

1) Score the sentences in the article.

We experimented with two word scoring schemes.

- tf-idf
- Bayes' theorem

2) Select a summary comprising of a number of sentences having highest scores.

## Dataset

We experimented with the following datasets

- Inshorts https://www.kaggle.com/sunnysai12345/news-summary/data
- CNN https://cs.nyu.edu/~kcho/DMQA/

The inshorts dataset has around 4000 articles and their summaries in around 60 words. The CNN dataset has around 92000 stories and highlights which we are treating as the summary. We experimented with only a subset of this dataset (10000 samples).

## Preprocessing

1. The articles and summaries were segmented into sentences.
2. The sentences were tokenized into sequence of words.
3. Stop words were removed.

## Word Scores

### tf-idf

Term frequency is the number of times a particular term occurs in a document. It is defined as,

$$tf(t,\ d)\ =\ count(t,\ d)$$

Where $t$ is a term, $d$ is a document.

Document frequency is the number of documents in which a term occurs in a corpus. It is defined as,

$$df(t,\ D)\ =\ count(d\ \varepsilon\ D\ and\ t\ \varepsilon\ d)$$

Where $D$ is the corpus.

Inverse document frequency measures how common or rare a particular term is. It is defined as,

$$idf(t,\ D) = log_{10}\frac{N}{1+\ df(t,\ D)}$$

Where $N$ is the number of documents in the corpus.

The denominator is adjusted with +1 to prevent terms that are not in the corpus, from giving a division-by-zero error.

The score of a word is the product of its term frequency and inverse document frequency.

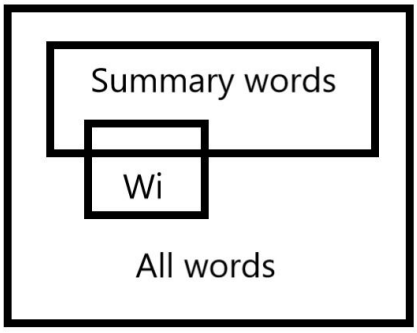
$$score(t,\ d,\ D) = tf(t,\ d)\ .\ idf(t,\ D)$$

tf-idf scoring scheme is unsupervised in the sense that it does not require summaries of articles to compute word scores.

**Bayes' theorem**

Bayes' theorem helps us calculate probabilities of events based on prior knowledge of conditions that might be related to the event.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$




V = Vocabulary
Sw = Summary word
Wi = Occurence of a particular word

Each word in V is given a score as follows.

$$score(Wi) = probability\ that\ particular\ word\ is\ a\ summary\ word$$

$$score(Wi) = P(Sw \mid Wi)$$

$$P(Sw \mid Wi) = \frac{P(Wi \mid Sw)\ P(Sw)}{P(Wi)}$$

$$P(Wi) = \frac{Number\ of\ occurrences\ of\ the\ word}{Total\ number\ words\ in\ the\ dataset}$$

$$P(Sw) = Probability\ a\ random\ word\ appears\ in\ the\ summary$$

$$P(Sw) = \frac{Number\ of\ words\ in\ the\ summaries}{Total\ number\ of\ words\ in\ the\ dataset}$$

$$P(Wi \mid Sw) = \frac{Number\ of\ occurrences\ of\ the\ word\ in\ summaries}{Total\ number\ of\ words\ in\ summaries}$$

Bayes' theorem uses the article summary pairs in the dataset to compute the word scores.

## Generating the summary

Now we can compute the score of a given word. We can define sentence score to be

$$sentence\ score = \frac{\Sigma\ score(word)}{Number\ of\ words\ in\ the\ sentence}$$

We pick the sentences from article based on the calculated sentence score to appear in the summary.

## Experimental Setup

We performed all our experiments with python3. We also used nltk for sentence segmentation, word tokenization and stopword removal. The pickle library was used to store the preprocessed dataset, to avoid the preprocessing again and again.

## Conclusion

We tried generating summaries of random news articles for different combination of scoring schemes and corpora. As there is no exact way of evaluating an automatically generated summary, we cannot claim anything about the summaries generated. However most of the times the important sentences were included in the summary. Below is a sample.

Original article

President Trump lashed out Tuesday at the publication of questions that special counsel Robert S. Mueller III was said to be interested in asking him as part of the Russia probe and possible attempts to obstruct the inquiry.
In a morning tweet, Trump said it was “disgraceful” that the 49 questions were provided to the New York Times, which published them Monday night.
“So disgraceful that the questions concerning the Russian Witch Hunt were ‘leaked’ to the media,” he wrote on Twitter.
It appears that the leak did not come from Mueller’s office. The Times reported that the questions were relayed to Trump’s attorneys as part of negotiations over the terms of a potential interview with the president. The list was then provided to the Times by a person outside Trump’s legal team, the paper said.

In his tweet, Trump also falsely asserts that there are no questions about “Collusion.” Among those is a query about Trump’s knowledge of any outreach by his former campaign chairman Paul Manafort to Russia “about potential assistance to the campaign.” A court filing this month revealed that Mueller had sought authorization to expand his probe into allegations that Manafort “committed a crime or crimes by colluding with Russian government officials.”
Another question asks about Trump’s knowledge of a June 2016 meeting in Trump Tower between his aides and a Russian lawyer who offered politically damaging information on Trump’s Democratic opponent, Hillary Clinton.
And another asks what Trump knew about “Russian hacking, use of social media or other acts aimed at the campaign?”
In his tweet, Trump calls collusion “a phony crime” and repeats his claim that none existed. The president also derides Mueller’s investigation as having “begun with illegally leaked classified information,” adding: “Nice!”

Summaries generated using sentence count of 3.

SUMMARY USING CNN CORPUS AND TF-IDF SCORING

So disgraceful that the questions concerning the Russian Witch Hunt were ‘leaked’ to the media, he wrote on Twitter. The Times reported that the questions were relayed to Trump’s attorneys as part of negotiations over the terms of a potential interview with the president. The list was then provided to the Times by a person outside Trump’s legal team, the paper said.

SUMMARY USING CNN CORPUS AND BAYES SCORING

Another question asks about Trump’s knowledge of a June 2016 meeting in Trump Tower between his aides and a Russian lawyer who offered politically damaging information on Trump’s Democratic opponent, Hillary Clinton. In his tweet, Trump calls collusion a phony crime and repeats his claim that none existed. The president also derides Mueller’s investigation as having begun with illegally leaked classified information, adding: Nice!

SUMMARY USING INSHORTS CORPUS AND TF-IDF SCORING

So disgraceful that the questions concerning the Russian Witch Hunt were ‘leaked’ to the media, he wrote on Twitter. The Times reported that the questions were relayed to Trump’s attorneys as part of negotiations over the terms of a potential interview with the president. The list was then provided to the Times by a person outside Trump’s legal team, the paper said.

SUMMARY USING INSHORTS CORPUS AND BAYES SCORING

A court filing this month revealed that Mueller had sought authorization to expand his probe into allegations that Manafort committed a crime or crimes by colluding with Russian government officials. Another question asks about Trump’s knowledge of a June 2016 meeting in Trump Tower between his aides and a Russian lawyer who offered politically damaging information on Trump’s Democratic opponent, Hillary Clinton. The president also derides Mueller’s investigation as having begun with illegally leaked classified information, adding: Nice!

## References


Tf-idf
https://en.wikipedia.org/wiki/Tf-idf

Bayes Theorem
https://en.wikipedia.org/wiki/Bayes_theorem

Text Summarization Techniques: A Brief Survey
https://arxiv.org/pdf/1707.02268.pdf

How to Prepare News Articles for Text Summarization
https://machinelearningmastery.com/prepare-news-articles-text-summarization/