

News Article Summarization

Shashwat Sanghavi
sanghavi.s@husky.neu.edu

Shraddha Phadnis
phadnis.s@husky.neu.edu

CS6120 Natural Language Processing, Spring 2018

College of Computer & Information Science
Northeastern University
Boston, MA

April 6, 2018

1 Introduction

- What is text summarization?
- Extractive summarization for news articles

2 Related Work

- Luhn's Approach
- TextRank

3 Proposed Methods

- Cleaning and Preprocessing
- Improvements on Luhn's method
- Improvements in TextRank

4 Experiments and Results

- Dataset
- Evaluation Method
- Results

What is text summarization?

- **wiki:** Text summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document
- Lossy compression of the text

Types of Text Summarization

- 1 Extractive summarization: Extract important sentences as it is from the original text
- 2 Abstractive summarization: Produces the summary in more human-like manner by paraphrasing the original text

The Daman and Diu administration on Wednesday withdrew a circular that asked women staff to tie rakhis on male colleagues after the order triggered a backlash from employees and was ripped apart on social media. The union territory's administration was forced to retreat within 24 hours of issuing the circular that made it compulsory for its staff to celebrate Rakshabandhan at workplace. It has been decided to celebrate the festival of Rakshabandhan on August 7. In this connection, all offices/ departments shall remain open and celebrate the festival collectively at a suitable time wherein all the lady staff shall tie rakhis to their colleagues, the order, issued on August 1 by Gurpreet Singh, deputy secretary (personnel), had said. To ensure that no one skipped office, an attendance report was to be sent to the government the next evening. The circular was withdrawn through a one-line order issued late in the evening by the UTs department of personnel and administrative reforms. The circular is ridiculous. There are sensitivities involved. How can the government dictate who I should tie rakhi to? We should maintain the professionalism of a workplace an official told Hindustan Times earlier in the day. She refused to be identified. **The notice was issued on Daman and Diu administrator and former Gujarat home minister Pratul Kodabhai Patels direction, sources said.** Rakshabandhan, a celebration of the bond between brothers and sisters, is one of several Hindu festivities and rituals that are no longer confined of private, family affairs but have become tools to push political ideologies. In 2014, the year BJP stormed to power at the Centre, Rashtriya Swayamsevak Sangh (RSS) chief Mohan Bhagwat said the festival had national significance and should be celebrated widely to protect Hindu culture and live by the values enshrined in it. The RSS is the ideological parent of the ruling BJP. Last year, women ministers in the Modi government went to the border areas to celebrate the festival with soldiers. A year before, all cabinet ministers were asked to go to their constituencies for the festival.

Abstractive Summary

The Administration of Union Territory Daman and Diu has revoked its order that made it compulsory for women to tie rakhis to their male colleagues on the occasion of Rakshabandhan on August 7. The administration was forced to withdraw the decision within 24 hours of issuing the circular after it received flak from employees and was slammed on social media.

Extractive summarization for news articles



Background

- Extractive summarization can be performed using **supervised** as well as **unsupervised approach**.
- **Supervised Approach:**
 - Requires dataset which has input text and important sentences in that text
 - If such data is not available, one needs to select and align sentences manually
 - Train a binary classifier (Include a sentence in summary or not?)
- **Unsupervised Approach:**
 - Performance is better than the above
 - Unsupervised approach is more common
 - Example: Luhn's Approach, TextRank

- Prof. H P Luhn gave an algorithm based on important words in the text to obtain the summary of the text
- **Overview:**
 - Remove stopwords from the text
 - Find word frequency, sort them and get top N words as the most important words
 - for each sentence, find sentence score by counting the number of important words in the sentences divided by the sentence length.
 - Top N1 sentence with the highest sentence score are the summary
- This approach works for short and simple text, but terrible for long and jargon heavy text
- Requires additional assumptions depending on the nature of the text

TextRank

- Graph-based ranking algorithm like PageRank can be extended to rank sentences in the text
- In PageRank, rank of the page is the probability of opening the page from other pages
- TextRank is derived from PageRank, where each sentence is equivalent to a page
 - Build a graph with each sentence as a node
 - Create links between all the sentences with weights as the sentence similarity scores
 - Run PageRank on this weighted graph to obtain sentence ranking
 - N top ranked sentences will be included in the summary

Cleaning and Preprocessing

- Obtain HTML file from url
- Parse the file and separate out each tag
- Find appropriate tag which contains article text, extract text from that tag
- Tokenize the text
 - Available methods can tokenize sentences separated by a punctuation (?!.) followed by a white space
 - Text contains <sentence> . <sentence> which can not be tokenized by any available tokenization tool
 - Instead, extract each sentence separately from HTML which lies in its own <p> tag.

Problem on Luhn's method

- Frequent words are not always important words
- **Eg.** *My name is John Doe. I am a singer. I can sing in English and Spanish. I like to play football as well. I can play football and sing at the same time.*
- Here, 'John Doe', 'English', 'Spanish', 'Football' appears only one time. 'I', 'play', 'can' etc are more frequent words
- In above, rather than 'I', 'Play' and 'Can', nouns should be more important
- TF-IDF can solve the above mentioned issue

TF-IDF & Stemming with Luhn's method

- $TF = \frac{\text{Word Count}}{|\text{Vocabulary}|}$
- $IDF = \log \frac{\text{number of Sentences}}{\text{number of sentences where the word occurs}}$
- $TF-IDF = TF \times IDF$
- Stemming is an approach to take care of words with similar roots
- **Eg.** The words 'stemming' and 'stemmed' are considered as different words.
- However, it is very rare that both of this word are referring to different meanings in the same context.
- Consider both of them as stem

Word2Vec

- Input: Vocabulary words
- Output: Probabilities of the words to be within $\pm N$ window distance from the input word
- Hidden Layer: Weight matrix contains the word vector

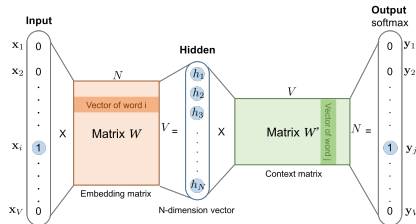


image from: <https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html>

Improvements on TextRank

- Sentence similarity is obtained by number of common N-grams in the sentences
- This neglects the context information for particular word
- This can be improved using Word2Vec followed by Sen2Vec
 - Word2Vec is calculated by considering words in $\pm N$ window
 - Takes context of the word in the consideration
- Sen2Vec can be calculated by adding word vectors(Word2Vec) of all the words in the sentence

Dataset

- The experiments are carried out with 4300 english news articles from 3 newspapers (Hindustan Times, India Today, The Guardian).
- The dataset includes 5 columns: Article No, Author, Headline, URL, Gold Summary

Evaluation Method

- Rouge-N Score is widely used method to analyze the quality of the summary

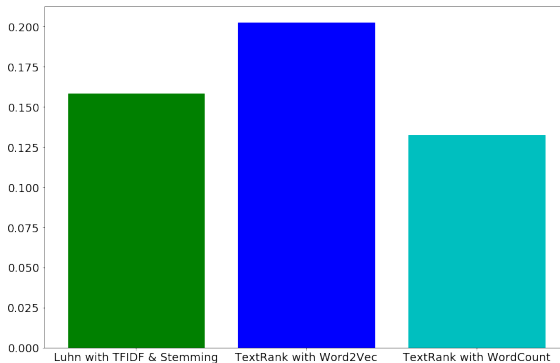
$$ROUGE - N = \frac{x}{y}$$

x = number of common N-grams in gold Summary and machine summary

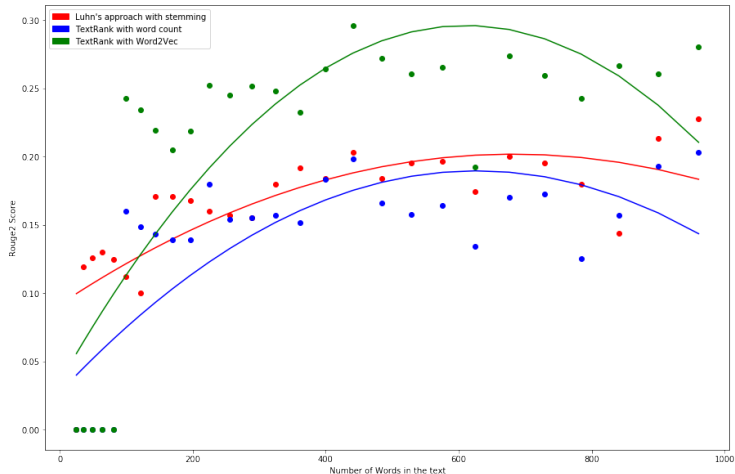
y = number of N-grams in gold summary

- Typically ROUGE-1 and ROUGE-2 are used for evaluating text summaries

Rouge2 score comparison for articles








Rouge2 score comparison for different lengths



Thank you.

References

-  Allahyari, Mehdi, et al. "Text summarization techniques: A brief survey." arXiv preprint arXiv:1707.02268 (2017).
-  Luhn, Hans Peter. "The automatic creation of literature abstracts." IBM Journal of research and development 2.2 (1958): 159-165.
-  Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.
-  Goldberg, Yoav, and Omer Levy. "word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method." arXiv preprint arXiv:1402.3722 (2014).
-  (2018) TextRank for Text Summarization - NLP-FOR-HACKERS. Retrieved April 06, 2018, from <http://nlpforhackers.io/textrank-text-summarization/>