Capstone Project Two

Summarize text information with news articles

Proposal

by Yiding Weng
April 07, 2019

Problem

With the rapid growth of internet and online information services, printing medias, more and more information is available and accessible. Explosion of information has caused a well recognized information overload problem. There is no time to read everything and yet we have to make critical decisions based on whatever information is available. Without absorbing the text information line by line, how to know what is the significant writing worth paying attention? What is the central theme?

Therefore, we need a tool to build an automatic text summarization and extraction system. I will use news articles as the source of information, which means news will be summarized in a short paragraph to provide distilled information.

Data

Proposed Data Collection:

All the news: 143,000 articles from 15 American publications

Data Location: https://www.kaggle.com/snapcrack/all-the-news

Approach

highlights important key words in the text

The approach to the problem will be considered to have two parts:
1. Summarize the new text.
2. Evaluate the methods for summarization .

For the first part, I will first experiment extractive summarization, which is extracting the important key words and phrases in the text as the representation of the whole text body, without adding new vocabulary. Then I will dive into abstractive summarization, which means the summary is generated by rephrasing with novel sentences and using the new words, instead of simply extracting the important sentences. The abstractive summarization techniques I will explore include sequence to sequence, textRank, deep-reinforced model, etc.

For the second part I will evaluate the result of different abstractive summarization techniques I mentioned above with Rouge and other evaluation systems. And conclude the best scenarios for applying each of the summarization method.