Text Summarization using Clustering and Submodular functions.

Introduction

Automatic Summarization of multiple documents into a single summary using various clustering algorithms and submodular functions.

We use a greedy approach to generate summary.

We calculate ROGUE score for each of the generated summaries by comparing them with human generated summaries.

Challenges

If we have n sentences then we have a set of 2^n possible candidate summaries.

We can’t follow this exponential approach and need a better approach.

For this we use a greedy approach which has a complexity of $O(n^2)$.

Adding a sentence may increase diversity of the summary but it may decrease the coverage and vice-versa.

Number of clusters are not defined for clustering algorithms.

Approach

We use a greedy approach and at each iteration we add a sentence to our candidate summary which has the best score when added to the candidate summary after previous iteration.

We use various clustering algorithms such as k-means, chinese whispers and hierarchical clustering.

At each iteration we use submodular functions and calculate score for coverage(L(S)) and diversity(R(S)) function for each sentence and choose the sentence with maximum score. We use this formula for scorning : $F(S)=L(S)+\lambda R(S)$

Approach

Basic NLP tasks such as tokenization, stemming, removal of stop words are applied for all the methods.

We calculate score for coverage and diversity functions for a candidate summary using clusters.

For similarity between sentences we use idf(inverse document frequency) of each token in sentences and term frequency of tokens in respective sentences.

Using these methods we generate our summary and calculate the ROGUE score.

Advantages of this approach

Decreases the complexity by a huge amount.

Very efficient and gives a pretty decent ROGUE score.

Helpful in other tasks such as information retrieval and web search.

Dataset and Results

Average ROUGE1 score:

K-Means: 0.3071840

Chinese Whisper: 0.306831

Hierarchical: 0.303747

References

http://melodi.ee.washington.edu/~bilmes/mypubs/lin2011-class-submod-sum.pdf

http://www.slideshare.net/kareemhashem/text-summarization

https://web.stanford.edu/class/cs345a/slides/12-clustering.pdf

Thanks!

Team-39

Aishwary Gupta (201302216)
B Prabhakar (201505618)
Sahil Swami (201302071)

Youtube