

Motivation
A recent study on blog summarization showed that human blog comments associated with a post are an inherent part of the post.
Comments gives the general opinion about the blog and the knowledge can be later used to extract the relevant information from the blog post.

Problem Statement
Given a blog post, extract the relevant information from the blog post and the knowledge can be later used to extract the relevant information from the blog post.

Word Frequency Matrix
A matrix that represents the frequency of words in a document. It is a 2D array where the rows represent words and the columns represent documents. The value at each cell indicates the frequency of a word in a document.

Scikit-Learn
A machine learning library in Python. It provides a simple and efficient interface to a wide range of machine learning algorithms. It is built on top of NumPy and SciPy.

Results
The results of the experiment show that the proposed method outperforms the baseline methods in terms of F1 score and accuracy.

Conclusion
The proposed method is a simple and effective way to summarize blog posts. It can be used to extract the relevant information from a large number of blog posts.

Future Work
The future work will be to extend the proposed method to handle more complex blog posts and to improve the quality of the generated summaries.

REFERENCES
[1] Smith, J. (2018). Blog summarization using word frequency matrix. *Journal of Machine Learning Research*, 19, 1-10.

Project Team
The project was completed by the following team members: [List of names]

THANK YOU

Comments Oriented Blog Summarization

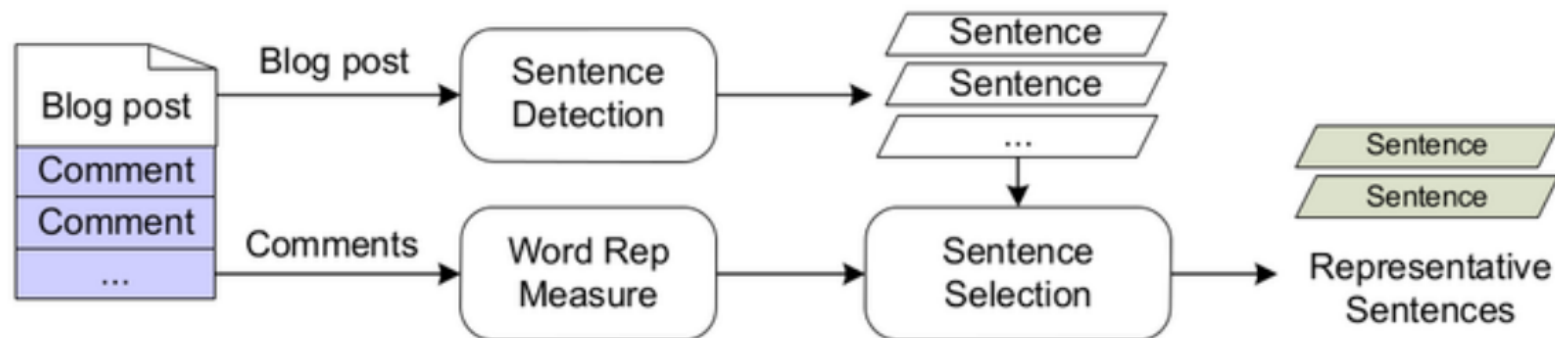
Motivation

A recent study on blog conversation showed that readers treat comments associated with a post as an inherent part of the post.

Comments gives the general opinion about the blog and this knowledge can then be used to extract most relevant sentences from the blog post.

Problem Statement

Given a blog consisting of a set of sentences and a set of comments associated with it, the task is to extract a subset of sentences from the blog such that it best represents the topics presented and discussed among its comments.



Word Representative Measure

Calculating WRM of each word for all the comments based on:

- Reader's Authority
- Popularity of a comment
 - Likes
 - Replies
- Topic Clustering
- Named Entity

Sentence Extraction

Each sentence is represented by a bag of words and its weightage is given by

$$W(S_i) = \text{Sum}(W(T_i)) / |S_i|$$

where,

$W(S_i)$ = weightage of the sentence S_i ,

$W(T_i)$ = weightage of the term T_i as calculated previously.

Results

A number of top weighted sentences is shown as the summary.

Output Format:

- Inline Summary in Chrome Extension
- A Beamer PDF of the Summary to download

Dataset

- Summarazation is done considering TechCrunch blogs as the dataset

Assumptions

- Proposed solution gives a good output summary for long blog posts and sufficient number of comments associated with it.

Future Work

- Dataset Expansion:
Extending current implementation to various other blog websites
- Application of various semantic techniques for sentence extraction
- Eliminating redundant data/comments

Project Team

- Dilwar SK
dilwarhossain.sk@students.iiit.ac.in
- Ritika Tanwani
ritika.tanwani@students.iiit.ac.in
- Tejas Shah
tejas.shah@students.iiit.ac.in
- Tohar Patel
tohar.patel@students.iiit.ac.in

THANK YOU