Project Proposal

Problem Statement:

Given a document, extract a summary of the document. Summarization is an active area in Natural Language Processing where researchers have applied extractive as well abstractive techniques to generate the summary. I feel this is an exciting problem and can significantly reduce the amount of time people spend on reading documents.

Importance of the project:

When we are considering large scale text files like legal documents, summarizing them can be extremely useful to lawyers and paralegals as they wouldn't have to read them all to extract the important details. It's crucial to know that "importance" is extremely domain specific and something that is important is a particular domain might be irrelevant to another domain.

The techniques used for this project can be extended to other domains like video summarization. Example: In surveillance videos, one might want to extract the important events from the uneventful context.

Datasets:

I have identified a few datasets that I would like to use for summarization.

[1] https://archive.ics.uci.edu/ml/datasets/Legal+Case+Reports

This dataset contains Australian legal cases from the Federal Court of Australia (FCA) for the years 2006, 2007 and 2008. This is an open source datasets available on UCI repository and researchers have used the dataset for different NLP tasks.

[2] http://www-nlpir.nist.gov/related_projects/tipster_summac/cmp_lg.html

As part of the TIPSTER SUMMAC effort, a corpus of 183 documents has been made available as a general resource to the information retrieval, extraction, and summarization communities. The documents are scientific papers that appeared in Association for Computational Linguistics (ACL) sponsored conferences.

Techniques:

I am thinking of trying a few techniques to solve the problem:

[1] Unsupervised Algorithm: Here, we can calculate the similarity between sentences and identify key sentences that are highly relevant to the document topic and at the same time ensure that duplicate/similar statements are not chosen (this can be estimated using a sentence similarity metric – maybe cosine/tfidf)

[2] Supervised Algorithm: Here, we use the extractive summary of the document as labels for the machine learning problem. I think this is a more difficult approach than [1]

but can lead to better performance.

I am looking to do further literature review to learn more techniques and implement them on the dataset.

Potential Challenges:

I have taken a courses in NLP and ML and I think I can take a shot at solving this problem. However, I don’t have extensive research experience in NLP, I may encounter a few issues with model optimization.

References

[1] https://machinelearningmastery.com/datasets-natural-language-processing/

[2] https://en.wikipedia.org/wiki/Automatic_summarization

[3] Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering docu- ments and producing summaries. In *Proc. of SIGIR*.

[4] Mihalcea, Rada, and Paul Tarau. ”Textrank: Bringing order into text.” *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.

[5] Jagadeesh, J., Prasad Pingali, and Vasudeva Varma. ”Sentence extraction based single document summarization.” *International Institute of Information Technology, Hyderabad, India* 5 (2005).

[6] https://en.wikipedia.org/wiki/Automatic_summarization#TextRank_and_LexRank