Fall 2018 ITCS 4111/511
Introduction to Natural Language Processing
Project Proposal Template

1. Introduction

After working for eight hours, commuting for another hour, and spending another four for daily commitments and responsibilities, the last thing a busy person wants to do is browse through half-a-dozen websites to get a well-rounded perspective of the news. Our proposed program seeks to meet user needs by being a one-stop solution to users for a quick round-up of the day's events.

By collecting articles from a variety of news sources (to account for bias and promote higher relevance) and providing a quick summary of the articles, our program will be an effective alternative for users to catch up on news, improving user's reading experience. The following sections of the proposal seek to provide more details on our proposed solution to solve this problem.

2. Related Work

As a team, we have done extensive research on automatic text summarization to better strategize how we are going to tackle this problem.

During the preliminary stages of our research, we used NLP Progress website as a starting point to look into current state-of-the-art approaches, which led us to some of the following papers and ideas.

In "A survey of text summarization techniques", the authors outlined numerous approaches for identifying important content for automatic text summarization. The authors Nenkova and McKeown particularly suggest frequency-based model, Bayesian model, and sentence clustering for topic representation. As this document mainly provided an overview of techniques we could potentially use for own solution rather than providing the specific details, we used this paper as a guide to direct us in researching for specific details in implementation and effectiveness of the techniques.

In "Single Document Automatic Text Summarization Using Term Frequency-Inverse Document Frequency (TF_IDF)", the authors presented more details on the TF-IDF approach including the two types of summarization (extraction and abstraction) and specific formulas to calculate various evaluation metrics like precision, recall, and f-measure. While extraction-based summarization seeks to extract words, phrases, and sentences directly from the corpus, abstractive summary seeks to generate sentences from a semantic representation of the text. In the paper, the authors go on to describe their own implementation and effectiveness of extraction based summarization using TF-IDF technique.

3. Project Topic and Proposed Solution

Since, text summarization is highly subjective and we are new to the field, as a team we decided to explore just one technique to get a better feel. We are planning to implement our program using the frequently used technique: term frequency-inverse document frequency (TF-IDF) approach. The Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic which reflects how important a word is in a document in the corpus (Salton et al., 1988). This method is often used as a weighting factor in information retrieval and text mining.

As part of data preprocessing, a basic linguistic analysis would be carried out which comprises of tokenization, sentence segmentation, part-of-speech tagging, stop word identification and elimination. This would be followed by feature extraction, which ranks each sentence in a text document based on its importance between a value of zero and one. Third, sentence selection and assembly are when the sentences are stored in descending order of the rank, and the highest rank is considered as the summary. Lastly, summary is generated by choosing the sentences and ordering them according to their position in the original document.

As we are planning to collect data from various news sources and there is high chance for articles/subjects to overlap among these sources, we are also planning to eliminate redundant summaries before displaying them to user. This similarity analysis among summaries will be done using cosine similarity measure. Details regarding program evaluation follows in the next section.

As our program parses news articles from various sources, we seek to gather data from well-recognized and credible sources like CNN, Fox News, and MSNBC. In addition to articles from these channels, we aim to find a dataset with summaries to evaluate our TF-IDF model.

Evaluation:

Given such subjectivity in text summarization, it is critical to devise a solid system to evaluate our model and decipher its usefulness when it comes to text summarization part of evaluation, we choose to use online text summarizer tool provided by www.tools4noobs.com as the ground truth measure for us to compare our model's performance. This is one of the top recommended online free tools for summarizing (MakeTechEasier). Like our application, it also has the option for users to choose the summary length.

According to Nedunchelian et al. (2011), the evaluation process of text summarization is performed by using three parameters from a typical confusion matrix: precision, recall, and f-measure. Following are the formulas and how they will translate to our application:

Precision (P) is the fraction of retrieved documents that are relevant.
Precision = P(relevant|retrieved) = #(relevant items retrieved) / #(retrieved items)

Recall (R) is the fraction of relevant documents that are retrieved.
Recall = P(retrieved|relevant) = #(relevant items retrieved) / #(relevant items)

The F-measure conveys the balance between the precision and the recall.
F-measure = 2*((precision*recall)/(precision+recall))

Final result:

As of now, we envision our system to display article titles followed by a small summary (five to ten sentences) for each of the articles. Each of these summaries will also be followed by a few links pointing to the original article in the source website(s). User may be able to choose the number of sentences they would like to see for their news digest from a drop-down menu. In addition, users would also be able to copy/paste an article in free text area for summarization.

4. Project Timeline (fill in this table, no additional text required)

	Date	Milestones	Status
1	Sept 6	Project Proposal Submission	Done
2	Sept 18	Complete data preprocessing	Done
3	Sept 28	Implement TF-IDF measure	Done
4	Sept 30	Project at least 50% done, project progress report due	Done
5	Oct 12	Refine the statistical approach to improve accuracy	Not yet begun
6	Oct 26	Implement cosine similarity measure to eliminate redundant summaries in system	Not yet begun
7	Nov 14	Complete evaluation and testing, consider implementing another model to compare/contrast the models	Not yet begun
8	Nov 25	Project Complete, final report due, draft and prepare for final presentation	Not yet begun

5. Team Roles and Contributions (fill in the table, no additional text required)

Milestones Accomplished

	Team Member Name	Task Completed
1	Shashikant Jaiswal	Gathered data (input articles), research, TF-IDF implementation, interim report and presentation
2	Tejaswini Naredla	Research related work, gathered two input data articles
3	Anusha Balaji	Data preprocessing, application design (tools to be used), interim report and presentation

Future Milestones

Sr. No.	Team Member Name	Responsible For
1	Shashikant Jaiswal	Web application back-end (if we're using a database), develop the front-end for text area for user given articles, final report/presentation
2	Tejaswini Naredla	Research, implementing statistical approach and TF-IDF, perform model evaluation and testing (against Tools4Noobs summaries), final report/presentation
3	Anusha Balaji	Cosine similarity measure to eliminate redundant articles/summaries, web app front-end to display articles and their summaries, gather Tools4Noobs output summaries for testing, final report/presentation

6. **Changes/Updates**

   In addition to our original proposal, we have come up with couple of new features to improve user experience. With our new approach, users should be able to choose number of sentences they would like to see in their summaries. Apart from viewing our top picks of news articles, users shall also have an option to copy/paste additional articles they would like a summary of. This way our application is extensible to new content.

7. **References (provide any background references, mandatory section)**

1. NLP Progress website https://nlpprogress.com/summarization.html
2. A survey of text summarization techniques https://link.springer.com/chapter/10.1007/978-1-4614-3223-4_3
3. COMPENDIUM: A text summarization system for generating abstracts of research papers https://www.sciencedirect.com/science/article/pii/S0169023X13000815
4. Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF) http://journal.binus.ac.id/index.php/comtech/article/view/3746
5. Top 4 online text summarizing tools https://www.maketecheasier.com/5-useful-tools-to-summarize-articles-online/