Business Requirements Document

Zachary Fitzgerald – 100769868

AIDI-2005 Capstone II

Marcos Bittencourt

January 24, 2020

**Summary statement**

Students and employees spend hundreds, if not thousands, of hours researching and summarizing articles each year. I want to create a system that will eliminate the need for humans to spend time summarizing text for their understanding. This will improve productivity and allow individuals to spend more time on tasks that AI systems cannot yet solve. To do this I will create a system that uses natural language processing techniques to take a text document and return a clear and concise summary of what was included in that document. If met with success, the summary will be on par, if not better, than what a human could produce from the same body of text. Often times it is challenging to understand a new topic one is studying because of the sheer breadth of the information within a document. If an individual imports the document into this system, then it will produce a concise, but intelligent, summary of the information within the text. This can improve the ability for someone to learn new topics without being scared away by immense amounts of information.

This program will also aid the business world by allowing a user to input multiple text documents and have the system return the top-5 most used words and the overall subject of the file. With this data, the system will compare across all the documents entered and return what is common between them all. Also, the program will tell the user whether the documents sentiment is overall positive or negative. This will be very beneficial to those in the business world as it will streamline the document intake process and save time for all parties involved.

This system is specifically created to aid both those in the academic world, as well as the corporate world. Whether it be for summarizing a research paper, or trying to better understand a new topic, this system has something that could aid any user. The end goal of the project is to create a system that is indistinguishable from a human in terms of its ability to summarize intelligent text documents. If met with success, this system has the ability to be implemented in any home or business and will help improve productivity and increase the access to knowledge of its users.

**Project objectives**

There is a total of six (6) objectives for this project, they are as follows.

1. To aid students in their studies. The system will accomplish this by allowing students to input or import a text document and having the algorithm produce a condensed high-quality summary of it. This will save students valuable time studying and allow them to focus on other aspects of their studies.
2. To be used in the corporate world. I.e.) Law Firms. The system can also be used in the corporate world in order to summarize documents. In professions that deal with thousands of documents that need to be read, the system can be very beneficial. This will save businesses potentially hundreds of hours that can be better used to tasks that AI cannot solve as of yet.
3. To be able to intake multiple documents and return the top-5 most used words and the overall subject of the document. Also, using this data, the program will be able to compare across multiple documents to see what is common among them all.
4. To return the sentiment of the document, whether it is positive or negative.
5. To be able to take difficult topics and summarize them in an easy to understand way. By summarizing text, the system can take dense articles and present a concise summary. One of the biggest barriers to knowledge is the way it is presented to the reader. If there is a concise summary, then the user could potentially be able to understand it clearer. If the system allows knowledge to be understood easier then it will be deemed a success.
6. To be as accurate as a human in summarizing documents. This is the overall goal of the project. If the system can summarize an article with the same, or better, clarity as a human, then it can be seen as a success. In order to do this, the system will need to summarize documents that other humans have already summarized, then we can compare the results.

**Needs statement**

This project is needed in the academic world as well as the business world for reasons which were presented in the project objectives. This system can help individuals save time on their studies by summarizing documents for them. For example, if a student had a professor’s lecture notes, they could put it into the system and then it will return a condensed summary of it. The program will also be able to take multiple documents and provide the top-5 words that are used in each one and the overall topic of the document. Also, the program will be able to take the top words in each document and compare them to the other ones inputted. This will allow businesses to be able to streamline the document reading process. The project will also save the students time and energy that can be better used for tasks that an AI system cannot yet complete.

Furthermore, the system can be used in the business world along the same lines as it is used for the academic world. The system is very useful for companies that have to go through countless text documents each day. We want to eliminate the need for people to sit down and read for hours on end to summarize papers. Workers time can be better used on tasks such as understanding the material presented rather than reading information that might not be useful. It is our understanding that the system will increase productivity in both the business and academic world.

**Project scope**

The scope of the project depends on the success of the original model. If the model that is created can accurately produce a summary that is indistinguishable from a human’s work, then the scope of the project will be large.

I want to have a system that is useful to both business users and regular students. This is why I would like to have two separate functions of the project. One that returns the top words and is able to be used as a comparative system. And another function that is able to be used as a text summarizer.

The goal of the project is to create a system that is indistinguishable from a human in its ability to summarize articles of text clearly and intelligently. The system will make use of both abstractive and extractive text summary to create an output that most closely mimics that of a human.

The system should be able to return the top words used in any document that is fed into the program. It will also return the sentiment, and overall subject of the file. Using this data, the system will be able to compare across all files recently inputted by the user. Thus, the program will be very beneficial to users in the business world as well.

The easiest way to think about extractive versus abstractive text summary is to think of extractive as a highlighter and abstractive as a pen. Extractive simply highlights the important words of phrases, whereas abstractive will take these important sections and write its own summary based off of them. In order to do this natural language processing tools will be used as well as TensorFlow and Keras to create neural networks.

**Dataset**

I will be using the WikiHow-Dataset. This is a large-scale dataset that is taken directly from wikihow.com. In total there are over 1.5 million entries within the dataset. The website is one of the world’s largest collections of ‘how-to’ manuals. I believe that this dataset will provide a good starting point for text summarization and also word counting.

I chose to use the Wikihow dataset because it is written by ordinary people. Other options for text summarization would be the CNN/Daily Mail dataset but these are written by journalists. I believe it is important to use articles written by ordinary people because they will all be written in different styles and this is far more representable of the end goal that the project hopes to reach.

The dataset that I will be using is split into four different parts. These are: Title, Overview, Headline, and Text. The title is simply the title of the article as it appeared on the Wikihow website. The overview is the introduction section of the WikiHow articles represented before the paragraphs corresponding to procedures. This section is not as important for my project as I am far more focused on the following two sections. The headline is a summary sentence of the paragraph which will be the reference point for any summary my model comes up with. Lastly, the text is the body of the paragraph (minus the headline) that the summary, sentiment, and word count will be based off of.

In order to use this dataset, I am going to have to complete numerous data cleaning processes. I will need to use the ‘re’ package on python to substitute letters and punctuation, make all character’s lowercase, and separate paragraphs from one another.

**Limitations:**

* My knowledge of NLP will need to improve in order to ensure that the performance of this model is the best that it can be.
* There are 1.5 million entries within the dataset. This will require a lot of computing power unless I want to use less of the dataset.
  + If I shorten the dataset, then the model might not perform as well.
* I will need to improve the NN that I made to predict sentiment because it was not trained on data like this.
* The data within the dataset is focuses on ‘HowTo’ articles. This could limit the generalizability of the results.

**Assumptions:**

* I assume that the text found within the dataset is complete and has actual sentences.
  + It is hard to manually check every single entry for completeness and coherency because there are over 1.5 million entries.

**Constraints**

* Time: in order to train a model successfully, I will need plenty of time since it will potentially take days of training to complete.

**Notes:**

I will be making use of a Neural Network which I created recently that was made to predict sentiment of text. I believe that this will be beneficial to include for this final project. A link to the Jupyter notebook file will be included on my GitHub.

Please see the linked Jupyter Notebook file on GitHub that will showcase what is included within this file.

**Link to the Jupyter Notebook file exploring the dataset on GitHub:**

<https://github.com/zaachfitz/Capstone/blob/master/Explore_dataset.ipynb>

**This is a link to the Neural Network which I created last semester that is for sentiment analysis:**

<https://github.com/zaachfitz/Capstone/blob/master/Module%202.ipynb>