CS 410 Project Progress Report

Christopher Dimitri Sastropranoto

Project description

Create a google chrome extension that can extract the text from a CNN news article and give a text summary about it.

Progress

My main goal for the early part of the project was to gain some experience using python's NLTK library to process and summarize a given piece of text. I've spent most of the time going over online tutorials and documentation about NLTK. In addition to this, I've also spent time researching the different types of ways that text is usually summarized. These methods include Abstractive and Extractive methods. Abstractive summarization is done by analyzing the text to determine it's meaning, which is then used to generate a summary. Extractive methods rely on scoring the words and sentences in a passage to determine important parts. These scoring functions are similar to those presented in CS 410. For the purposes of this project I will be using the latter method.

After going through the documentation, I went ahead and created a Jupyter Notebook to start getting some hands on experience using NLTK. I've done some basic preprocessing steps to tokenize and build a document-term matrix out of a sample article I pulled from CNN.

Challenges

The biggest challenge so far has been familiarizing myself with the NLTL library. One thing that was hard was to setup the library in the beginning as the installation steps were not as simple as just using pip.

Remaining Work

The last remaining step to completing the Jupyter notebook is to implement the scoring function and actual text summarization steps. Fortunately NLTK has some built in libraries that can help with this meaning that it should not take up too much time. Once this step has been completed, the following things need to be done.

1. Package Jupyter Notebook code as an API.
2. Build google chrome extension. The extension will consist of a simple UI with a button allowing the user to summarize the article. All it does is just scrape the text from the screen, pass it into the API from step 1 and finally outputting the result.