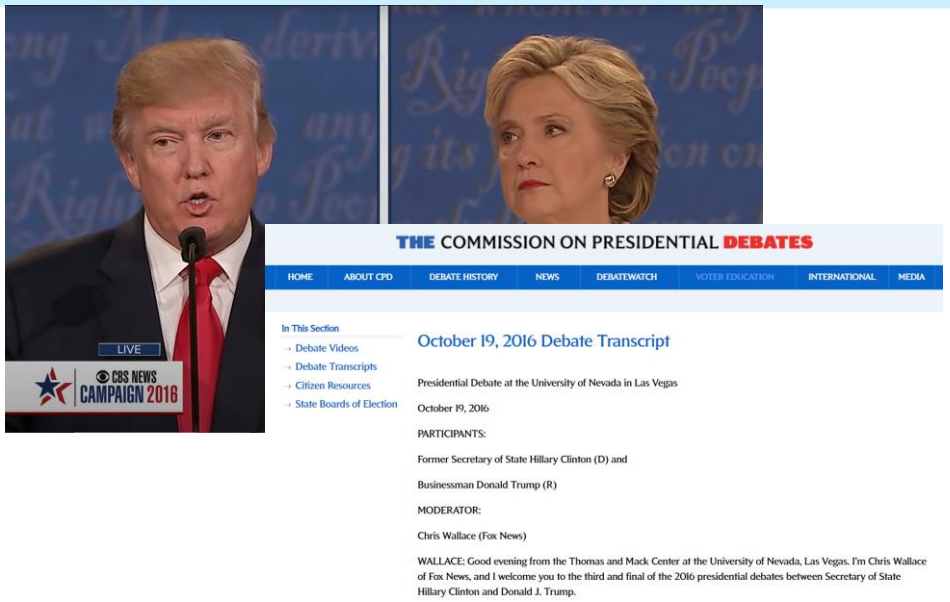
23/24 WS Introduction to Programming: Final Project

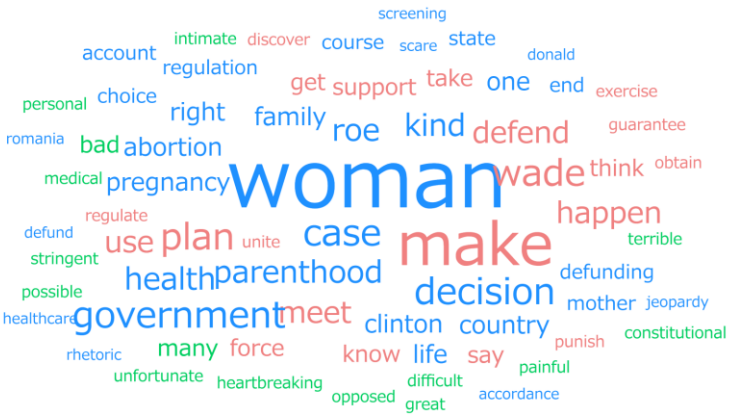
Making Word Clouds of the US Presidential Debates

Koji Okumura

WHAT will I do?

I am planning to make word clouds of the US presidential debates. A word cloud like the figure below illustrates what words appear and how often the words are used in the text. The transcripts of the US presidential debates from 1976 to 2020 are archived at https://www.debates.org/voter-education/debate-transcripts/. In my plan, if you write in the URL where the script is archived, you can see word clouds of two candidates. The lists of frequent words with part-of-speech information are also provided for linguistic analysis.





WHY did I make this plan?

My major is linguistics, and my specific interest is discourse analysis of politicians. Since the US presidents have been very influential in the world, their discourse has been analyzed by many researchers. A word cloud is an effective way to visualize the text characteristics. That is why I will try to make word clouds of US presidential candidates.

HOW will I reach the goal?

Scraping

First, the data should be collected. I get the transcript data from the website by scraping. I will use Requests and Beautiful Soup modules for scraping. The two modules are commonly used for the aim.

Text Cleansing

Since the transcript includes all the remarks of two candidates and moderators, I extract only one candidate's remarks. In the transcripts, if the speaker changes, a description like "TRUMP: " appears.

Analysis

To make word clouds, I will use WordCloud module. Furthermore, I use Nltk, NumPy and Matplotlib modules to analyze the text, just as I did in the recent homework. I also use Pandas library to treat tables skillfully.