Assessment 8 : CCCS660 Computational Intelligence

Quentin Lao — Matricule : 261233739

HANDS-ON OPTION

I implemented my own Python code.
The code is provided at https://github.com/soniquentin/CCCS660_McGill_Assignment8.

The notebook is divided in 4 parts :
— Part 1 : Import of libraries and open the text file containing the speech
— Part 2 : Preprocessing (cleaning text by removing punctuation, tokenization, remove stop words and stemming)
— Part 3 : Building of a basic word cloud based on the tokens
— Part 4 : Modifying the word cloud with sentiment classification of the tokens

1   Text chosen

For the text, we chose the speech *The Gettysburg Address*[1] delivered by President Abraham Lincoln in 1863 during the dedication of the cemetery at the Gettysburg battlefield in Pennsylvania. He wanted to memorialize the Union dead and highlight the redemptive power of their sacrifice.

This text is interesting because it contains both positive words to inspires hope and negative words because of the somber realities of war.

2   Preprocessing

In the preprocessing stage (part 2), we used the famous library `NLTK` to transform the text into a well-formed list of tokens. This process is summarized in the provided diagram in figure 1.
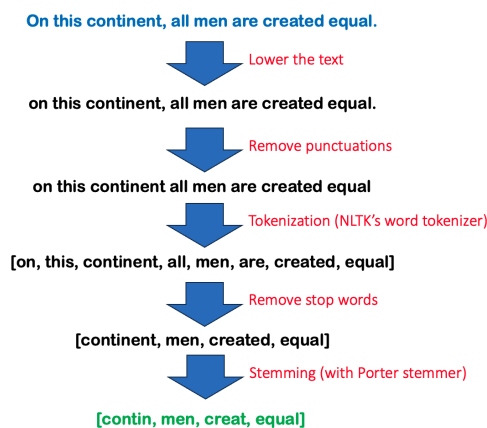

FIGURE 1 – Summary of the different steps of preprocess

---

1. https://www.abrahamlincolnonline.org/lincoln/speeches/gettysburg.htm

3 Basic word cloud

We could make a word cloud based on the stemmed words obtained during preprocessing to find out the most prominent words from "The Gettysburg Address." We used the library `wordcloud` and get the following word cloud shown in figure 2 :

FIGURE 2 – Basic word cloud made from the speech *The Gettysburg Address*

4 Sentiment classification of words (in-depth stage for the hands-on option)

To classify the words in the speech, we use a publicly available sentiment lexicon from Kaggle [2]. We first stemmed the vocabulary words as done in the preprocessing stage.

To classify a stemmed word of our speech, we simply compated it with the stemmed words in our vocabulary. If the word falls under the positive category, it's marked as positive ; similarly, if it's in the negative category, it's marked as negative. Words not found in either category are classified as neutral. This straightforward approach is quite effective. The figure 3 shows the resulting word cloud which uses green to denote positive words and red for negative words.
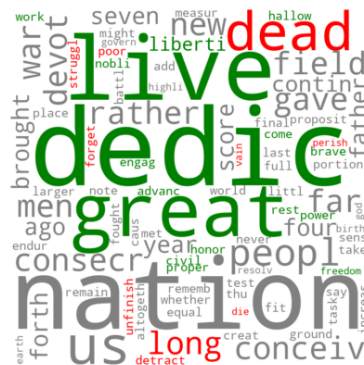

FIGURE 3 – Word cloud made from the speech *The Gettysburg Address* with word classification

Coloring the color in a word cloud according to the word sentiment is very effective to visualize instantly the emotional tone of the speech. This representation suggests that despite the grim context of war, Lincoln's speech predominantly conveys hope : "*live*" is bigger that "*dead*"