Team Members : Lakshmi Vandana Nunna , Yaswanth Bandaru , Tanuja Polineni

GitHub Link: https://github.com/LakshmiVandanaNunna/ICE-1\_WordFrequencyDistribution/wiki

1. Why we use stopwords? Why stopwords are not necessary for NLP frequency distribution.

Stopwords is a library that comes with “ntlk.corpus” package. This helps in providing a list of words that are trivial ie; do not reflect the information/crux of the content but they just help in framing sentences to align with the grammatical structure. With stopwords.words() function, we could configure a specific language. Once we pass the name of the language as an argument to this function, it would return a list of stop words for that specific language. By using the output of this function, we could filter the word list that we have derived from the target content.

The reason we try to find the frequency distribution of words in any given text, is to understand the overall idea of the content. If we do not eliminate stopwords from this list, they would always come up as the high frequent words since they are the connecting words that frame a sentence in any given language. So, by not eliminating stopwords, we will never be able to estimate the words that reflect the information of the content. Once we purge the stopwords and visualize the frequency distribution of the rest of the words, we would be able to identify the words that occur more frequently in the text. This helps us to evaluate and estimate the general context and idea of the content.

1. Based on high frequency words what information you can extract from the graph?

After eliminating stop words/special characters/digits, the frequency distribution would have a list of words that reflect the information/idea of the content. We could perform multiple techniques to identify the final list of desired words. For example, consider a specific number of top frequent words or choose a specific frequency threshold above which we would consider the words etc. In the use case that was presented, we were asked to pick a web page about SpaceX. So, ideally the highest frequent word should be ‘SpaceX’. But, if any other word has similar or same frequency then, the web page might be about both ‘SpaceX’ and the other word. By deriving ratios of frequencies, we could also estimate the percentage of text that aligns towards each word/idea and identify the central topic.

On another note, we could also identify the words that are less frequent to understand what the content does not cover, or is not focused on. This helps derive many evaluations. For instance, the intent of the author and speculate why he/she is subtle about specific information.

If a specific word dominates all the other words by a very big margin of frequency, then, we could consider a higher probability for the content to be focused on that specific information- related to that word.

I could only answer this question based on my understanding, but there could be many more complex derivations that could be done based on the frequency distribution.

1. Can you provide different visualization for frequency distribution? If yes, please perform. If no, why?

Yes, we could perform multiple other visualizations for the frequency distribution. In the code, I choose to implement a box plot of all the top 10 frequent words. As opposed to a connected line, this plot would allocate confined box spaces for each word where the width of the box is constant and the height of the box represents the frequency of the word.

Chart, bar chart

Description automatically generated

To provide another example, we could also perform wordcloud representation, which would display a poster with a cloud of textual words in different font/color, so that the size of the font is proportional to the frequency of the word. The downside of this plot is that the user will not be able to identify the exact frequency count of each word except for to visually compare it against other words to estimate the proportions