Yonathan Mekonnen
Nathan West
Derrick Adjei

Project 4 Report

2.1

Open feature_vector.csv file to view the feature vector

Open tdm.csv to open the Term Document Matrix

2.2

A: Tokenize Sentences

It is used because before you can began processing text you need to split it into units(words). While applying this text-mining step every sentence in the document is split up into sub-lists inside of a list. In this step we are ridding of the white space separating two differing sentences.

B: Remove Punctuation and Special Characters

It is used to reduce dimensionality and make our dataset more manageable. It is also used before you can began processing text because we are focusing on the words in each sentence, so punctuation and special characters is pointless to analyze in this scenario. Any punctuation or special characters in each sentence is removed during this text-mining step.

C: Remove Numbers

It is used before you can began processing text because we are focusing on the words in each sentence, so numbers are pointless to analyze in this scenario. They do not add any context to deciphering a word. Any numbers in each sentence is removed during this text-mining step.

D: Convert Upper-Case to Lower-Case

When looking at a word if the same word is both uppercase and lowercase it will be analyzed as two separate words. This step is important to rid of excessive words that can be counted as the same. In this step we are ridding of all upper-case's in a word and potential duplicate words that may be processed as different words but be the exact same.

E: Stop Word Removal

It is used to remove words that are frequent in any given language but do not convey any discriminatory information and are equally distributed across the document. It rids of the stop words given to us in the txt file and reduces the number of words in our document that are not as needed making our analysis simpler.

Yonathan Mekonnen
Nathan West
Derrick Adjei

F: Preforming Stemming

It is used to remove irrelevant information and order linguistic forms by dimension reduction. During our processing we want to group words that are only differing with each other based off its suffix as the root of that word. This further reduces excessive words.

G: Combine Stemmed Words

It is used to count multiple occurrences of a word from the whole txt document which will rid of duplicate words when processing the data. While applying this text mining step we are ridding of all duplicates of the stemmed words in each sentence.

3

In order to obtain different results you can change the amount of clusters that the sentences could be grouped into. You could also affect the results by defining a minimum number of occurrences to keep track, so the Term Document Matrix can shrink and affect the sentences that get clustered