Textual analysis code 2

filename = "TweeterData1.xlsx";

slCharacterEncoding('UTF-8')

T = readtable('TweeterData1.xlsx','Range','A1:F267','ReadVariableNames',true)

text = xlsread(filename,'Lapas2','F1:F267')

textData = T.text;

textData(1:5)

documents = preprocessText(textData);

documents(1:5)

bag = bagOfNgrams(documents)

figure

wordcloud(bag);

title("Text Data: Preprocessed Bigrams")

mdl = fitlda(bag,10,'Verbose',0);

figure

for i = 1:4

subplot(2,2,i)

wordcloud(mdl,i);

title("LDA Topic " + i)

end

cleanTextData = erasePunctuation(textData);

documents = tokenizedDocument(cleanTextData);

bag = bagOfNgrams(documents,'NGramLengths',3);

figure

wordcloud(bag);

title("Text Data: Trigrams")

tbl = topkngrams(bag,10)

function documents = preprocessText(textData)

% Convert the text data to lowercase.

cleanTextData = lower(textData);

% Tokenize the text.

documents = tokenizedDocument(cleanTextData);

% Erase punctuation.

documents = erasePunctuation(documents);

% Remove a list of stop words.

documents = removeStopWords(documents);

% Remove words with 2 or fewer characters, and words with 15 or greater

% characters.

documents = removeShortWords(documents,2);

documents = removeLongWords(documents,15);

% Lemmatize the words.

documents = addPartOfSpeechDetails(documents);

documents = normalizeWords(documents,'Style','lemma')

end