Assignment 9 - Text Toolkit and Document Analysis

Due Date: Thursday, December 5th, 9am

Assignment Overview

In this assignment, you will:

1. Learn how to use a toolkit for text analysis
2. Learn about stemming and using stop words
3. Create an index for documents in a collection
4. Process a search query and return relevant documents
5. Reflect on your experiences

Submission Requirements

You will submit a file called **YourNameA9report.pdf**, which will include

- Short answers to each of the reflection prompts below.
- The chart that you will create in Part 4.
- Your vector plot of your search query from Reflection 4.

Instructions

Peter Holme's word stemmer is a web-based text analysis tool which walks you through the processes of stemming and analyzing documents. These are the steps one would go through while creating an index of a document which can be used for IR.

Part 0

1. Go to http://holme.se/stem/
2. Open A9_text.txt This file contains text you know well, the first three paragraphs of TDO.
3. Read the text, copy it and paste it into the Text field in the word stemmer.

Part 1

- Click the “Don’t use any stopwords” checkbox in **Extra super fun stuff**.
- Increase the number of words for word count to 100.
- Leave all other settings at their default.
- Click Send.
- Look at the results in Step 1 on the right.

 Reflection 1 (1 point)

- Why did some words disappear?

- Why do we want to remove these terms/words and what are some problems that may arise when we do so?

Part 2.1

- Change the minimum number of characters in a word to 1.
- Run the analysis again. Look at the results in Step 1, and note what got added back and what is still missing. Now, look at Step 2 and check out how words were stemmed.

 Reflection 2.1 (1 point)
Pick *three stemmed words* from the list, and for each, answer the following:
- Why was that stem chosen?
- What problems might arise with that stem?

Part 2.2

- Next, look at the Word Count list in Step 4: Notice the number of words and the frequency.
- Are there any words that surprise you in this list?

 Reflection 2.2 (1 point)
- How much about the article can be inferred from the words on this list?
- Looking at the word frequency, where would you draw the line to eliminate stop words?

Part 3

- Uncheck the *don't use any stopwords* checkbox.
- Take the top 10 words from your list that you think should be used as stopwords. Enter them into the stopwords box and then run the analysis.
- Copy the resulting Word Count list (Step 4) into a table or spreadsheet.
- Next, delete the list you just added to the stopwords and run the analysis again. (When the stopwords box is empty, it will run using a default list of stopwords, which you can see in Step 3).
- Copy these new Word Count results (Step 4) into your table and compare them with your previous results to see how the list and frequencies changed.

 Reflection 3 (1 point)
- Why do you think the system chose some of the stopwords? Are there any words that surprise you?
- Why are the stopwords stemmed?

Part 4

- Run A9_text2.txt, A9_text3.txt, and A9_text4.txt through the system and copy each resulting word count list into separate parts of a spreadsheet. Now you have an index for each document in your collection (note that in a real system you would want the full index - for simplicity's sake, we've cut it off at 100 terms).
- Now you get to play the part of a search engine. You will now manually run the query "Organizing System" on your collection of three documents.
- For each document, find the term frequency (tf) for each term in the query. Record each calculation in a chart like the one below.
- Calculate the idf for each term. List it in your chart and show your work in Reflection 4.
- Calculate the tf-idf for each document. Again, show your work in Reflection 4.

Reflection 4 (6 points)

- As indicated before, show us your work in calculating tf, idf, and tf-idf. (2 points)
- In a simple graph, draw the vector plot for your search query (you can do this by hand or in software). (2 points)
- Which document appears to best fit your query? Why? (1 points)
- What does tf-idf allow you to deduce that using term frequency alone does not? (1 point)

Term	**tf** for doc 1	**tf** for doc 2	**tf** for doc 3	**idf**	**tf-idf** for doc 1	**tf-idf** for doc 2	**tf-idf** for doc 3
organizing							
system							

Extra Credit (1 point):

- Calculate the cosine similarity between your top two documents and the query. Show your work. Which is now the best fit for your query? Why?