CSCE470 Written Assignment 1 (5 Points!)
Due by 11:59 pm on Sep. 7th

1. Short answers (2 Points! 1/3 Point per answer):

a) What are the main conclusions and implications of zipf's law when used to analyze a text corpus and the relation between term frequencies and their ranks?

b) Name one positive impact of eliminating stopwords from the index of a search engine.

c) Name one negative impact of eliminating stopwords from the index of a search engine.

d) How does the text processing task "stemming" work?

e) Please explain why's inverted index called "INVERTED"?

f) A positional index is helpful for what type of queries?

2. Problem solving: Positional indexes (3 Points! 1 Point for solving each sub-problem)

Consider the following documents:
Doc 1: today you are you. that is truer than true.
Doc 2: you have brains in your head. you have feet in your shoes. you can steer yourself any direction you choose. you are on your own. and you know what you know. and you are the one who'll decide where to go.

1). Positional indexes are very useful to search against documents. Let's build positional indexes based on these documents using the format DocID: <pos1,pos2,pos3,…>; …, for example, the positional indexes for the words "**are**" and "**those**" are as follows.

are: 1: <3>; 2: <22, 34>
today: 1: <1>

No need to consider any type of normalization of tokens, in addition, the punctuation marks should be stripped off from words and ignored when you count tokens.

Now please show the positional indexes for the words "**you**", "**head**" and "**feet**".

you:

head:

feet:

2). A phrase query "word1 word2" retrieves documents where word1 is immediately followed by word2, A /k query "word1 /k word2" (k is a positive integer) retrieves documents where word1 occurs within k words of word2 on either side. For example, k=1 demands that word1 be adjacent to word2, but word1 may come either before or after word2.

For the following queries, return all the docs and corresponding positions (phrase starting positions) for which the query conditions are met. If no document meets the criteria, return none.

"you are you"

"head feet"

"you /2 you"

3). Let's say we want to find documents in which "you /2 you", and the two words "you" and "you" are in the same sentence. This condition only applies to document 1.

How would you modify the positional index to support queries that demand the terms to be in the same sentence? You can assume that the parsing step is able to identify the sentences in a document. Please write down an example of the modified postings list for the words "**you**".

you: