CSE225 Data Structures, 2017(FALL)

PROJECT #1 (Deadline: 21.10.2017, 24:00)

Text similarity with common words and n-grams

One of the simple techniques for finding the similarity between two documents is based on Term-Frequency (TF) feature demonstration in real world cases. In this technique, simply the number of shared words are counted and a similarity ratio is generated based on this counter value. Another technique is called n-gram. A simple n-gram is a contiguous sequence of words. They generally give higher classification accuracy in comparison to TF since words can convey different semantics according to their context (i.e. polysemy, synonym).

This project is a programming assignment in C which aims to build an algorithm based on linked-lists that will compare two documents and find the similarity between them.

Your program needs to open and read two text files with the names 'file1.txt' and 'file2.txt' which are located in the same directory with the source code of the program. The number of terms in these documents will be arbitrary. In other words, the length of these files will be arbitrary.

Your program is expected to find the similarity between these two documents:

a) (35 points) based on TF. In other words, your program is expected to calculate the number of common terms these two documents have in common. For instance; the number of shared words between the first document and the second document is 18, then the output of the result will be:

   the number of common words: 18
   the common words are (in ascendingly sorted fashion)*:* .....

b) (65 points) based on the number of common 2-grams they have. For instance; the number of shared 2-grams between the first documents and the second document is 12, then the output of the result will be:

   the number of common 2-grams: 12
   the common 2-grams are(in ascendingly sorted fashion)*:* .....

A sample scenario:

Input1.txt

Term Frequency is a feature representation technique. Term frequency is often used in text mining field. Term Frequency measures how frequently a term occurs in a document.

Input2.txt

Term Frequency Inverse Document Frequency is another feature representation technique. Term Frequency Inverse Document Frequency is often used in text mining field. Term Frequency Inverse Document Frequency measures how important a term is.

a)

the number of common words: 16

the common words are:

a
document
feature
field
frequency
how
in
is
measures
mining
often
representation
technique
term
text
used

b)

the number of common 2-grams:12

the common 2-grams are:

a term

feature representation
frequency measures
is often
in text
measures how
mining field
often used
representation technique
term frequency
text mining
used in

Important Notes:

In your demo, we will run your program by **reading arbitrary-length input files**. **You need to store these files into two different linked-lists. Since the number of common terms and 2-grams will be arbitrary; we also want to see the results will be stored in another linked-list.**

In demo, we will use different input files and we will check if your program find the correct results or not. We will also check your data structure your design architecture.
Of course, other questions based on your implementation and coding structure will be asked you during your demo. These questions will be those kinds of questions which could be answered by only the students who really implement his/her project by himself/herself.

The main goal of this project is to be familiar with linked-list. So, if you use arrays instead of linked-lists then you will get zero, unfortunately.

You are responsible for demonstrating your program to your TA Berna Altınel on the scheduled day that will be announced later.

CODE SUBMISSION:

You should use the following email address in order to submit your code:

cse225.marmara.2017 at gmail dot com

Your any submission after the project submission due date, will not taken into consideration.

You are required to exhibit an ***individual effort*** on this project. Any potential violation of this rule will lead everyone involved to **failing from the course** and necessary disciplinary actions will be taken.

Good luck!!!