HW 1

Knowledge Discovery in Social and Information Networks - 2018
Summer 2018
Total points: 100
Issued: 06/06/2018 Due: 06/09/2018

The CiteSeer UMD collection is a standard text document collection, consisting of abstracts of research articles from Computer Science, which are sampled from the CiteSeer digital library. The dataset is available for download from sicuaplus.

Tasks:

1. Write a program that preprocesses the collection. In doing so, tokenize on whitespace and remove punctuation.

2. Determine the frequency of occurrence for all the words in the collection. Answer the following questions:

   a. What is the total number of words in the collection?

   b. What is the vocabulary size? (i.e., number of unique terms).

   c. What are the top 20 words in the ranking? (i.e., the words with the highest frequencies).

   d. From these top 20 words, which ones are stop-words?

   e. What is the minimum number of unique words accounting for 15% of the total number of words in the collection?
   Example: if the total number of words in the collection is 100, and we have the following word-frequency pairs:

word	*tf*
the	20
of	10
a	10
data	8
mining	7
...	...

   the answer to this question will be (1 word accounts for 15% of the total 100 words).

3. Integrate the Porter stemmer and a stopword eliminator into your code. Answer again questions a.-e. from the previous point. (See below a link to a Java Porter stemmer implementation and to a stopwords list).

   https://www.dropbox.com/s/rexuzz3j56vi4bt/Porter.java
   https://www.dropbox.com/s/5789sj8v07j2id0/stopwords.txt

4. Encode each document using the sparse TF-IDF representation.

Note: It is highly recommended that your code is as modularized as possible.

Submission instructions:

1. write a README file including:
   - a detailed note about the functionality of each of the above programs,
   - complete instructions on how to run them
   - answers to the questions above
2. make sure you include your name in each program and in the README file.
3. make sure all your programs run correctly on the virtual machines.
4. submit your assignment through sicuaplus.