CMSC 409: Artificial Intelligence

Fall 2019, Instructor: Dr. Milos Manic, http://www.people.vcu.edu/~mmanic

Project 4

CMSC 409: Artificial Intelligence
Project No. 4
Due: Nov. 12, 2019, noon

Pr.4.

1. Download and unzip "Project4_sentences.zip" and "Project4_code.zip" files.
   A set of sentences is given in the file "*sentences.txt*". Each sentence is a line in the file. Create the feature vector by writing a program that applies the following text mining techniques to this set of sentences.
   A. Tokenize sentences
   B. Remove punctuation and special characters
   C. Remove numbers
   D. Convert upper-case to lower-case
   E. Remove stop words. A set of stop words is provided in the file "*stop_words.txt*"
   F. Perform stemming. Use the Porter stemming code provided in the file "*Porter_Stemmer_X.txt*"
   G. Combine stemmed words.
   H. Extract most frequent words.

Provide the feature vector in your report.

Note:
The feature vector contains unique sets of words that appear in the set of sentences provided.
The file "*Project4_code.zip*" contains implementations of the Porter Stemmer in several languages. You can use any version of the code provided (provided versions of the code are Java, Matlab, Python, and C). Make sure you rename your file accordingly. More source code for the Porter Stemmer can be found here: http://tartarus.org/martin/PorterStemmer/

2. Using the feature vector generated in first task, write a program that generates the Term Document Matrix (TDM) for ALL the sentences in "*sentences.txt*", similar to TDM below.

Example TDM

Keyword set	anonymous	identify	car	**…**
Sentence 1	1	4	3	…
Sentence 2	2	0	1	…
…..	…	…	…	…
Sentence 20	2	0	0	…

a) **Provide the TDM in your report.**
b) For each of the text mining steps (A to H), explain why they are used, and what sort of information is lost while applying each of the text-mining steps.

3. Write a program implementing the clustering algorithm of your choice (WTA or FCAN). Apply that algorithm to TDM to group similar sentences together.
    a. How many clusters/topics have you identified?
    b. What drives the dimensionality of TDM? What can you do to reduce that dimensionality? Does the order of data being fed to algorithm matter?
    c. Show and comment the results.

Note:

1. Your software must be user friendly. The TA must be able to test it simply by executing the code.
2. Project deliverable should be a zip file containing:
    a. Written report with answers to the questions above in pdf.
    b. The source code.
3. Submit your zip file to Blackboard. Please name the zip file as GroupName_Project4.zip.