# CMSC 409: Artificial Intelligence

Fall 2015, Instructor: Dr. Milos Manic, http://www.people.vcu.edu/~mmanic

## *Project 4*

***CMSC 409: Artificial Intelligence***  
***Project No. 4***  
**Due: Nov. 30, 2015, noon**

**Pr.4.**

1. Download and unzip "Project4_sentences.zip" and "Project4_code.zip" files.
   A set of 5 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.

**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 1. write a program that generates the Term Document Matrix (TDM) for ALL the sentences in "*sentences_ tdm.txt*" AND "*sentences.txt*". **Provide the TDM in your report**.

Example TDM

| **Keyword set** | **Sentence 1** | **Sentence 2** | **…** | **Sentence 10** |
|---|---:|---:|---|---:|
| anonymous, anonymously | 2 | 0 | … | 1 |
| identify, identifies, identifying | 0 | 3 | | 1 |

3. For each of the text mining steps (A to G), explain why they are used, and what sort of information is lost while applying each of the text-mining steps.

----------------------------------------------------------------

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 word, pdf, ps, or txt format
   b. The data and separation lines in format as specified by Project1_data.zip
   c. The source code.
3. Submit your zip file to Instructor misko@vcu.edu and cc TA Daniel Mariño marinodl@vcu.edu. Use the subject line [CMSC 409, Family name, Project 4]