# CMSC 409: Artificial Intelligence

Fall 2021, Instructor: Dr. Milos Manic, http://www.people.vcu.edu/~mmanic

## *Project 4*

***CMSC 409: Artificial Intelligence***
***Project No. 4***
**Due: Nov. 11, 2021, noon**

*Student certification:*
*Team member 1:*
*Print Name:* _____ *Date:* _____
*I have contributed by doing the following:* _____
*Signed:* _____ *(you can sign/scan or use e-signature)*

*Team member 2:*
*Print Name:* _____ *Date:* _____
*I have contributed by doing the following:* _____
*Signed:* _____ *(you can sign/scan or use e-signature)*

*Team member 3:*
*Print Name:* _____ *Date:* _____
*I have contributed by doing the following:* _____
*Signed:* _____ *(you can sign/scan or use e-signature)*

**Pr.4.**

1. Create the feature vector by writing a program that applies the following text mining techniques to this set of paragraphs. (4 pts)

Download and unzip "Project4_code.zip" files. A set of paragraphs is given in the file "Project4_paragraphs.*txt*". Proceed with the following steps.

**A.** Tokenize paragraphs
**B.** Remove punctuation and special characters (including html tags)
**C.** Remove numbers
**D.** Convert upper-case to lower-case
**E.** Remove stop words. A set of stop words is provided in the file "Project4_*stop_words.txt*"
**F.** Perform stemming. Use the Porter stemming code provided in the file "*Porter_Stemmer_X.txt*"
**G.** Combine stemmed words (in case Porter stemming did not catch some words and you would need to stem them by yourself).
**H.** Extract the most frequent words (i.e. words for the feature vector, or most characteristic, distinct words).

**I. Provide the feature vector in your report.**

# CMSC 409: Artificial Intelligence

Fall 2021, Instructor: Dr. Milos Manic, http://www.people.vcu.edu/~mmanic

## *Project 4*

**Note**:
The feature vector contains unique sets of words that appear in the set of paragraphs 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 the first task, write a program that generates the Term Document Matrix (TDM) for ALL of the paragraphs in "Project4_paragraphs.*txt*", similar to TDM below. (5 pts)

**Example TDM**

| **Keyword set** | review | watch | scene | **…** |
|---|---|---|---|---|
| **Paragraph 1** | 1 | 4 | | … |
| **Paragraph 2** | 2 | 0 | 1 | … |
| ….. | … | … | … | … |
| **Paragraph 20** | 2 | 0 | 0 | … |

**a) Provide the TDM in your report.** (3 pts)

**b)** For each of the text mining steps (A to H), explain the purpose of each step and what sort of information is lost while applying each of those text-mining steps. (2 pts)

3. Write a program implementing the clustering algorithm of your choice (Kohonen WTA or FCAN). Apply that algorithm to TDM to group similar paragraphs together. (6 pts)

**a)** How many clusters/topics have you identified? (2 pts)

**b)** What drives the dimensionality of TDM? What can you do to reduce that dimensionality? Does the order of data being fed to the algorithm matter? (2 pts)

**c)** Show and comment on the results. (2 pts)

---

**Note:**

1. Your code 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:
    i. Written report with answers to the questions above in pdf format.
    ii. The source code.
3. Submit your zip file to Canvas. Please name the zip file as GroupName_Project4.zip.