Hong Kong Baptist University
Department of Computer Science
COMP4075 Social Computing and Web Intelligence /
COMP7630 Web Intelligence and Its Applications
Semester 2, 2018-19

Assignment 1 (Due March 13, 2019 23:59)

Web Data Acquisition, Pre-processing and Analysis using Scikit-Learn

Submit via the link provided at the course web site.

* Note: This is a programming assignment. In case you find that there is something not yet clearly specified, make reasonable assumptions and move on with your development accordingly.

Part A:

1) Develop a (python) program that can
    a. Extract the "movie names", and "storylines" of at least 100 movies from https://www.imdb.com/
    b. Perform all the preprocessing steps you think appropriate
    c. Compute the tf-idf features to represent each storyline
    d. Perform k-means clustering on the movie storylines
    e. Compute the silhouette coefficient (that shows the clustering quality)
    f. Compute the wordcloud for each cluster.
2) Use the program to perform clustering given three different number of clusters, compute their silhouette coefficients, pick the clustering result with the highest value, and compute the wordcloud for each cluster.

Part B:

1) Develop a (python) program that can
    a. Read in the dataset fetch_20newsgroups (use ALL the categories).
    b. Perform classification using Logistic Regression (find it out from sklearn).
    c. Output the classification results in terms of accuracy, precision, recall, F1-score, and confusion matrix.
    d. Perform document classification using Support Vector Machine (find it out from sklearn) and output the results.
2) Use the program to perform the classification and compare the classification results obtained by the two different classification methods.

What to Submit:

Submit the following in one zipped file named as [YourStudentID_ass1].zip

a. Source code of Part A1 and all the results of Part A2 in a MSWord file
b. Source code of Part B1 and all the results of Part B2 in a MSWord file

Bonus:

Extract also "genres" from https://www.imdb.com/ for each movie, use them as the class category labels, and repeat Part B. Submit that as [YourStudentID_ass1_bonus].zip

-- The End --