Natural Language Processing
Assignment 2
Sentiment Analysis
Due Thursday 30th of April 2020 before 8:00pm

Guidelines:

1- This assignment should be done individually. Collaboration is encouraged, however your submission must include a statement describing the contributions of each collaborator (if any).
2- Assignment must be submitted on the eLearning gateway.
3- No hard copies are accepted.
4- Assignment should include your full name and student ID.
5- Cutoff date for this homework is Thursday the 30th of April 2020 (before 8:00pm).
6- Any late assignment will not be accepted.
7- Academic Fraud: Cases of plagiarism will be deal with according to university regulations. A tool will be used to detect similarities between submissions. The tool will be run on all submissions, across all sections. Submissions that are flagged by the system will receive a mark of zero.
8- Warning: It's your responsibility to ensure that the submission is indeed received by the eLearning gateway. If the submission is not there by the deadline, it will obviously not be marked.
9- **Python** and the **NLTK** will be used in the implementation of this assignment.

The goal for this homework is to perform **Sentiment Analysis** by classifying entire movie reviews as being either positive or negative.

Train a **Naïve Bayes** classifier on the Internet Movie Database review (**imdb1)** data set provided to you with Laplace (add-1) smoothing. Your classifier will use words as features (unigram). Use log probability scores.

Task 1: Your first task is to use Multinomial Naïve Bayes classifier to train and test using 10 fold cross-validation mechanism. Evaluate your classifier by reporting the confusion matrix and average accuracy.

Task2: Remove stop words from your train and test sets. You have been provided with file (english.stop) that includes the stop-words you need to remove. Evaluate your model again with the stop words removed. Does this approach affect average accuracy?

Task3: Repeat task one reporting the results of using Boolean Naive Bayes classifier. This means removing duplicate words in each document (review) before training. Does this approach affect average accuracy?

Task4: Repeat task 1 using a bigram and trigrams features. Report results and discuss.

Submit on the eLearning a zip file of your assignment. It should include the source code and a separate file (word document) that includes a printout of the results from your four tasks, along with a short discussion on the results.