# Homework 2

Please email your report as a *pdf file* as a **separate attachment** and a tarball of your code to nlp.qc.cuny@gmail.com. The email should be sent before 5:00 pm. If you send multiple emails, **only your first (earliest) submission will be graded**.

- Feel free to talk to other members of the class in doing the homework. You should, however, write down your solution yourself. Please try to keep the solution brief and clear.

- **The programming assignment is to be done in Python. Only standard Python libraries are to be used for this homework**. For the programming assignment, in addition to the results (see below), you need to turn in a *short* report describing what you did, what were the difficulties, and what were your conclusions.

1. [Movie review classification using Naïve Bayes - 10 points]

   Assume that you have trained a Naïve Bayes classifier for the task of sentiment classification (please refer to Chapter 4 in the J&M book). The classifier uses only bag-of-word features. Assume the following parameters for each word being part of a positive or negative movie review, and the prior probabilities are 0.4 for the positive class and 0.6 for the negative class.

   |         | **pos** | **neg** |
   |---------|---------|---------|
   | I       | 0.09    | 0.16    |
   | always  | 0.07    | 0.06    |
   | like    | 0.29    | 0.06    |
   | foreign | 0.04    | 0.15    |
   | films   | 0.08    | 0.11    |

   Question: What class will Naïve Bayes assign to the sentence "I always like foreign films"? **Show your work.**

2. [Implementing the Naïve Bayes classifier for movie review classification – 90 points]
In this assignment, you will write 2 scripts: NB.py and pre-process.py. NB.py
should take the following parameters: the training file, the test file, the file where
the parameters of the resulting model will be saved, and the *output* file where
you will write predictions made by the classifier on the test data (one example
per line). The last line in the output file should list the overall accuracy of the
classifier on the test data. The training and the test files should have the following
format: each line corresponds to an example; first column is the label, and the
other columns are feature values.

pre-process.py should take the training (or test) directory containing movie re-
views, should perform pre-processing[1] on each file and output the files in the
vector format to be used by NB.py.

a) Implement in Python a Naïve Bayes classifier with bag-of-word (BOW) fea-
tures and Add-one smoothing. Note: Do not use smoothing for the prior
parameters. You should implement the algorithm from scratch and should
not use off-the-shelf software. [35 points]

b) Use the following small corpus of movie reviews to train your classifier. Save
the parameters of your model in a file called movie-review-small.NB (you can
manually convert this small corpus into the vector format, so that you can
run NB.py on it). [10 points]

    i. fun, couple, love, love **pos**

    ii. fast, furious, shoot **neg**

    iii. couple, fly, fast, fun, fun **pos**

    iv. furious, shoot, shoot, fun **neg**

    v. fly, fast, shoot, love **pos**

c) Test you classifier on the new document below: {*fast, couple, shoot, fly*}.
Compute the most likely class. Report the probabilities for each class. [5
points]

d) Now use the movie review dataset provided with this homework to train a
Naive Bayes classifier for the real task. You will train your classifier on the
training data and will test it on the test data. The dataset contains movie
reviews; each review is saved as a separate file in the folder "neg" or "pos"
(which are located in "train" and "test" folders, respectively). You should
use these raw files and represent each review bag-of-word features, where each
feature corresponds to a word from the vocabulary file (also provided), and
the value of the feature is the count of that word in the review file (please
consult the lecture on Naive Bayes for how to represent each movie review
using BOW features).

---

[1]Please read below for how to do the pre-processing.

*Pre-processing*: prior to building feature vectors, you should separate punctuation from words and lowercase the words in the reviews. You will train NB classifier on the training partition using the BOW features (use add-one smoothing, as we did in class). You will evaluate your classifier on the test partition. In addition to BOW features, you should experiment with additional features. In that case, please provide a description of the features in your report. Save the parameters of your BOW model in a file called movie-review-BOW.NB. Report the accuracy of your program on the test data with BOW features.

Investigate your results. For the reviews for which your program made incorrect predictions, were there any trends that you observed? That is, can you explain why these incorrect predictions were made? [40 points]