CSCI381/CSCI780 Natural Language Processing

Homework 2

Due 05/04/2021 at 3:10pm

Due Wednesday, 05/04/2022 3:10pm in class.
Please submit your **report** and a printout of your code. If you are unable to attend class, please slide a copy of the assignment under the door of my office (SB A332) by the due date and time.
Please also email your report as a separate attachment and a tarball of your code to nlp.qc.cuny@gmail.com. The email should be sent before 3:10 pm.
If you do not submit a paper copy of the report and code, the assignment will get a grade of 0. If you do not submit an electronic version, the assignment will get a grade of 0. If you do not submit code both electronically and as a hard copy, the assignment will get a grade of 0. Please see further submission instructions at the end of the email.

- 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: one example per line; 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 reviews, 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) features 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 **comedy**

      ii. fast, furious, shoot **action**

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

      iv. furious, shoot, shoot, fun **action**

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

   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]

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

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 using a vector of 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.

   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]

Submission
Please include all the required code files in a tarball and email the tarball and the report to `nlp.qc.cuny@gmail.com` using subject line CSCI381/CSCI780 Homework 2:

- The tarball should include the Python code along with a README file that has instructions on how to run it in order to obtain the answers to questions in Part 2.

- The report that should be attached separately to the same email should include the answers to the questions in Part 1 and Part 2.

Your grade will be based on the correctness of your answers, the clarity and completeness of your responses, and the quality of the code that you submitted. Please refer to the course webpage on late submission policy.