assignment1_22

May 4, 2022

Assignment 1 (NLU 22)

Sentiment analysis using logistic regression In this assignment you train a binary classifier for classifying movie reviews from IMDB database. The task is to classify them as positive or negative. The task has been divided into several subtasks.

1. You will need the following libraries. Do not import everything.
   - re
   - pandas
   - numpy
   - scipy
   - nltk
   - scikit-learn
2. The data consists of two directories–positive and negative reviews.

3. Each review is a document.
4. Process the texts so that you get rid of punctuation but keeping spaces. We have to be careful with stopwords. Completly removing them may lead to loss of crucial information. (How?)
5. You will have to map each document (email) to a vector.
6. You will need to use **tf/idf** weighting. You should create the tf-idf vectors from scratch. **Do not use library functions**.
7. Once you have the vectors for each document apply logistic regression to the training set to fix the weights. You *may* use **sklearn** logistic regression function from linear models.

8. Test your model with the test set and report accuracy, recall and precision.
9. Write a short report (about 250 words) on the model and how one may improve it.
10. You will be provided with a set of functions. Your task is to complete them.
11. Do not change anything in the structure of these functions. If you have print functions for testing comment them out.
12. Commenting your code is important. But not too much commenting.
13. The following are the basic steps:
    1. process the dataset.
    2. build a vocabulary
    3. convert documents to vectors by **tf/idf** weighting
    4. match the input/output vectors
    5. train the model (use logistic regression from sk-learn linear models)
    6. test the model and compute performance measures
    7. write short report, perhaps start with the report

[ ]: