Homework 2: Sentiment Analysis (Feature engineering based && Word2Vec based)

Description

- In this homework, you will need to use feature engineering and word2vec based models for sentiment analysis.

- Each sentence in our data has a sentiment label to represent its sentiment level.

- The sentiment level of the sentences are defined as five classes:
  - “very negative”, ”negative”, ”neutral”, ”positive”, ”very positive” which are represented by 0 to 4 in our task

Description

Finish this task with two methods:

(1) Feature engineering based sentiment analysis:

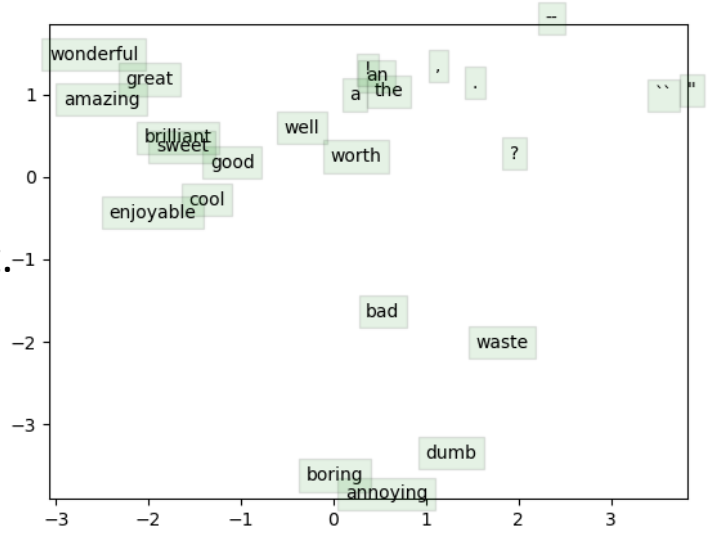
- Feature extraction: use the bag of words features

- Sentiment analysis: use the naïve bayes classifier

Description

(2) **Word2vec based sentiment analysis:**

- Word2vec training: use word2vec model (Skip-gram in this task) to **train** your own word vectors, and **visualize** your word vectors.
  - The framework of word2vec model:
    - **Calculate** the loss function and gradients
    - **Train** your word vectors with gradient descent. (SGD and BGD are also recommended)
    - **Visualize** your word vectors



- Sentiment analysis: use the **average** of all the word vectors in each sentence as its feature, train a **classifier** (e.g. softmax regression) with gradient descent method.

Provided Files (Dataset)

Dataset: Stanford Sentiment Treebank (SST) dataset

1. **original_rt_snippets.txt** contains 10,605 processed snippets from the original pool of Rotten Tomatoes HTML files. Please note that some snippet may contain multiple sentences.

2. **dictionary.txt** contains all phrases and their IDs, separated by a vertical line |

3. **sentiment_labels.txt** contains all phrase ids and the corresponding sentiment labels, separated by a vertical line.

- Note that you can recover the 5 classes by mapping the positivity probability using the following cut-offs for very negative, negative, neutral, positive, very positive respectively:

  [0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8], (0.8, 1.0]

- Please note that phrase ids and sentence ids are not the same.

Provided Files (Dataset)

- 4. **datasetSentences.txt** contains the sentence index, followed by the sentence string separated by a tab. These are the sentences of the train/dev/test sets.

- 5. **datasetSplit.txt** contains the sentence index (corresponding to the index in datasetSentences.txt file) followed by the set label separated by a comma:

  1 = train
  2 = test
  3 = dev

8,544 , 2,210 and 1,101 instances for training , development and testing respectively.

- Please note that the datasetSentences.txt file has more sentences/lines than the original_rt_snippet.txt.

Provided Files (for word2vec based method)

- **data_utils.py**
  - This file is used to read data from our dataset.
- **gradcheck.py**
  - This file is used to check whether your grad is right or not.
- **sgd.py**
  - This file is used to run stochastic gradient descent.
- **run.py**
  - Train your own word vectors and visualize it.
  - This file can be edited if you want to change the hyperparameter for better performance

Files to be updated (for word2vec based method): your work

- **word2vec.py**
  - This file is used to build your word2vec model , including calculation of your cost and gradient.
- **softmaxreg.py**
  - This file is used to train a softmax regression model, and the softmax regression part is given. Your work is to implement the feature extraction part.
- **sentiment_word2vec.py**
  - This file is used to complete the sentiment analysis mission. Your work is to find the best hyper parameter and regularization parameter. (This file can run without any implement)

Files to be added (for feature engineering based method): your work

- **Sentiment_bagofwords.py**
  - This file is used to complete the sentiment analysis with feature engineering based method. You can use the naïve bayes classifier with bag of words features.

Submission

- Generate a zip file and name it as "sid_homework-2.zip".

- It should include all python files mentioned above as well as the following files:

✓ a figure of the visualization of your word vectors named "word_vectors.png"

✓ a figure of the accuracy of your word2vec based sentiment analysis named "word2vec_acc.png" on the train and dev set

✓ a written report named "sentiment analysis based on feature engineering and word2vec.pdf" which describes your two methods, the results and analyses.

- Program: codes should be written in python.

- Report: in English with no more than 3 pages.

Evaluation

- We will mark your homework based on the criterias:
  - Accuracy (10%)
  - Program (60%): Feature engineering based sentiment analysis: 20%; word2vec based sentiment analysis: 40%
  - Report (30%)

Due

- Submit your homework via E-learning system.
- Deadline: Mid-night at **November 20th 2019**

- If you have any questions about this homework, send email to TA or me.

TA:
陈　伟：　18110980003@fudan.edu.cn
王红瑞：　18210180087@fudan.edu.cn