复旦大学大数据学院
School of Data Science, Fudan University

Assignment 4
Word2vec & Sentiment Analysis

Description

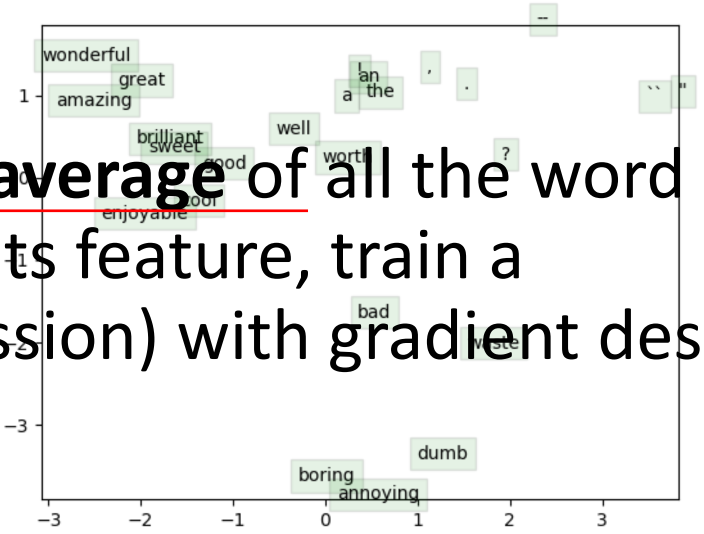
- In this assignment, you will need to use word2vec models for sentiment analysis.

- 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

- This task is separated as two subtasks:
- Word2vec: 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 method.(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.



Formula

- word prediction formula：

$$\hat{\boldsymbol{y}}_o = p(\boldsymbol{o} \mid \boldsymbol{c}) = \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w=1}^{W} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

- Softmax-CE loss function:

$$J_{softmax-CE}(\boldsymbol{o}, \boldsymbol{v}_c, \boldsymbol{U}) = CE(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

- negative sampling loss function：

$$J_{neg-sample}(\boldsymbol{o}, \boldsymbol{v}_c, \boldsymbol{U}) = -\log(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \sum_{k=1}^{K} \log(\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c))$$

- Skip gram cost：

$$J_{\text{skip-gram}}(\text{word}_{c-m\ldots c+m}) = \sum_{-m \le j \le m, j \ne 0} F(\boldsymbol{w}_{c+j}, \boldsymbol{v}_c)$$

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:
- [0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8], (0.8, 1.0]
- for very negative, negative, neutral, positive, very positive, respectively.
- 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(Given)

- 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

Provided Files(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.py
  - This file is used to complete the sentiment analysis mission. Your work is to find the best hyper parameter and regularization parameter.

Submission

- Generate a zip file and name it as “sid_homework-4.zip”.

- It should include all python files mentioned above, a figure of the visualization of your word vectors named “word_vectors.png”, a figure of the visualization of your sentiment analysis named “reg_acc.png”,and a written report named “word2vec and sentiment analysis.pdf”.

- Program: codes should be written in python.

- Report: the report needs to be written in English with no more than 4 pages.

Evaluation

- We will mark your homework based on the criteria mentioned on the “assignment4.pdf” :
    - Gradient Calculating(30%)
    - Program (40%)
    - Report (30%)

Due

- Submit your homework via E-learning system.
- Deadline: Mid-night at **December 26th 2017**

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

- TA in Charge
  - 顾云帆(aleck16@163.com)