

Deep Learning for Natural Language Processing

-

Project

Alexis Conneau

February 2018

Abstract

In this assignment we will cover monolingual word and sentence embeddings, multilingual word embeddings and sentence classification with Bag-of-Vectors (BoV) and LSTMs. An ipython notebook for the full project is attached: **nlp_project.ipynb**. It contains instructions on what you must code. Please refer to the "Deliverable" part for a description of what we expect from you in terms of deliverable.

1 Monolingual embeddings (/6)

In **nlp_project.ipynb** you are asked to write functions for computing the nearest neighbors of any word, without using an external package. You will build two classes for word vectors and bag-of-words vectors, such that you get the desirable outputs (see code).

2 Multilingual word embeddings (/4)

The goal is to find a mapping W that will map a source word space (e.g French) to a target word space (e.g English), such that the mapped source words will be close to their translations in the target space. For this, we need a dictionary of "anchor points". Here, we will use the identical character strings in both languages. We can show that the solution of $\underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F$ has a closed form.

Question Using the orthogonality and the properties of the trace, prove that, for X and Y two matrices:

$$W^* = \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F = UV^T, \text{ with } U\Sigma V^T = \operatorname{SVD}(YX^T). \quad (1)$$

In `nlp_project.ipynb` you are asked to create X and Y using the identical character strings in each language, compute W and output target nearest neighbors of source words in the shared space.

3 Sentence classification with BoV (/4)

In this section and the following, we give you the train, dev and test sets of the Stanford Sentiment Treebank fine-grained sentiment analysis task. It consists of input sentences that you have to classify into 5 classes. For the test set, we only provide you with the input samples, not the ground-truth labels. You will have to produce your predictions using your best model, and send it to us. We will evaluate ourselves the quality of your predictions.

You are asked to use scikit-learn to learn a logistic regression on top of bag-of-words embeddings on the SST task.

Question : What is your training and dev errors using either the average of word vectors or the weighted-average?

4 Deep Learning models for classification (/6)

Question : Which loss did you use? Write the mathematical expression of the loss you used for the 5-class classification.

Question : Plot the evolution of train/dev results w.r.t the number of epochs.

Question : Be creative: use another encoder. What are your motivations for using this other model?

5 Deliverable

You should create a .zip file, named "P2_nlp_familyname_firstname.zip" which contains:

- answers.pdf (with your answers to the questions above)
- nlp_project.ipynb
- logreg_bov_y_test_sst.txt
- XXX_bov_y_test_sst.txt (bonus)
- logreg_lstm_y_test_sst.txt
- XXX_XXX_y_test_sst.txt

Please consider that having the same format for all the students save TAs a lot of time. We may consider penalties for submissions that do not follow these simple rules.. Thanks!