NLU Assignment 1

Due Date: 11 Mar, 2019

Weightage: 15%

In this assignment, you will implement and experiment with prediction-based models for learning word embeddings.

Datasets:
You'll be using the Reuters corpus for this. NLTK provides an interface to access the dataset. Use the train/test split as provided there.

You can split the train set further to create a validation set for model selection.

Task 1

Implement a word2vec skipgram model [1]. While you can use ML libraries of your choice, you should implement the core logic of word2vec yourself.

EDIT: To evaluate the models, you should use the $SimLex-999$ word similarity task [2].

Select models using performance on validation set and report values of the objective function and performance on $SimLex-999$ task on the different splits of the dataset.

Experiment with the hyper-parameters of your model such as batch size, number of negative samples, embedding size. Report your observations along with your explanations for the findings.

Task 2

Using the best model learnt in Task 1, verify the claims in the word2vec paper about capturing relationships between words through the analogical reasoning

task.

Report your findings quantitatively as well as qualitatively.

[For bonus points] Can you identify biases in the learnt word embeddings? For example, if you consider the embeddings of "man" and "woman" and explore the neighbourhood, what kind of biases do you observe?

– – – – –

Provide plots and figures to support claims.

Source code: Please submit your code on Github and share the link in the report. Include a README.

Report: Please use Latex to typeset your report in the ACL format ($http://www.acl2019.org/medias/340-acl2019-latex.zip$). Please try to keep the report brief and to the point.

Further details on the submission will be shared in a while.

– – – – –

References:

[1] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013. [2] Hill, Felix, Roi Reichart, and Anna Korhonen. "Simlex-999: Evaluating semantic models with (genuine) similarity estimation." Computational Linguistics 41.4 (2015): 665-695.