Duplicate Questions P21

1 Project idea

Every online Q&A website faces the issue of duplicate questions. In our project we take this NLP problem, evaluate and apply techniques to classify whether question pairs are duplicates or not.

Our work will be divided into following modules

- Data processing: Tokenization, Stemming and Lemmatization.
- Evaluation and selection of word representation techniques between word2vec or GloVe
- Evaluation of classical ML techniques such as Random Forests and SVM
- Evaluation of deep learning techniques for text classification such as RNN

2 Data set

Question pairs data set provided by Quora will be used for the project. (1)

3 Software and tools

Following list of software and relevant tools will be leveraged for this project:

- Python libraries : NLTK, Sklearn, keras, matplotlib, Numpy, Pandas, TensorFlow
- Word2Vec, GloVe

4 Referenced papers

1. GloVe: Global Vectors for Word Representation (2)
2. From Word Embeddings To Document Distances (3)
3. Detecting Semantically Equivalent Questions in Online User Forums (4)

5 Team and work division

1. Darpan Dodiya (dpdodiya) - Pre-processing and exploratory analysis of data
2. Mohit Vinchoo (mvincho) - Evaluation and implementation of deep learning models
3. Shantanu Sharma (ssharm34) - Evaluation and implementation of Word2Vec and GloVe
4. Shrijeet Joshi (sjoshi22) - Evaluation classifier techniques such as Random Forest and SVM.

6 Midterm milestone

We aim to accomplish following by midterm milestone:

- We should be able to finalize word representation and chalk out differences between them.
- Baseline model should be created for comparison with other models.
- We should be done with at least 3 more models and justify which one works better and why.

References

[1] https://www.kaggle.com/c/quora-question-pairs/data

[2] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.
https://nlp.stanford.edu/pubs/glove.pdf

[3] Matt J. Kusner , Yu Sun , Nicholas I. Kolkin , Kilian Q. Weinberger, From word embeddings to document distances, Proceedings of the 32nd International Conference on International Conference on Machine Learning, July 06-11, 2015, Lille, France
http://proceedings.mlr.press/v37/kusnerb15.pdf

[4] Bogdanova, Dasha et al. “Detecting Semantically Equivalent Questions in Online User Forums.” CoNLL (2015).
https://aclweb.org/anthology/K15-1013