Duplicate Questions Detection P21

Darpan Dodiya (dpdodiya)
Mohit Vinchoo (mvincho)
Shantanu Sharma (ssharm34)
Shrijeet Joshi (sjoshi22)

1 Outline

Why, this problem is important to solve

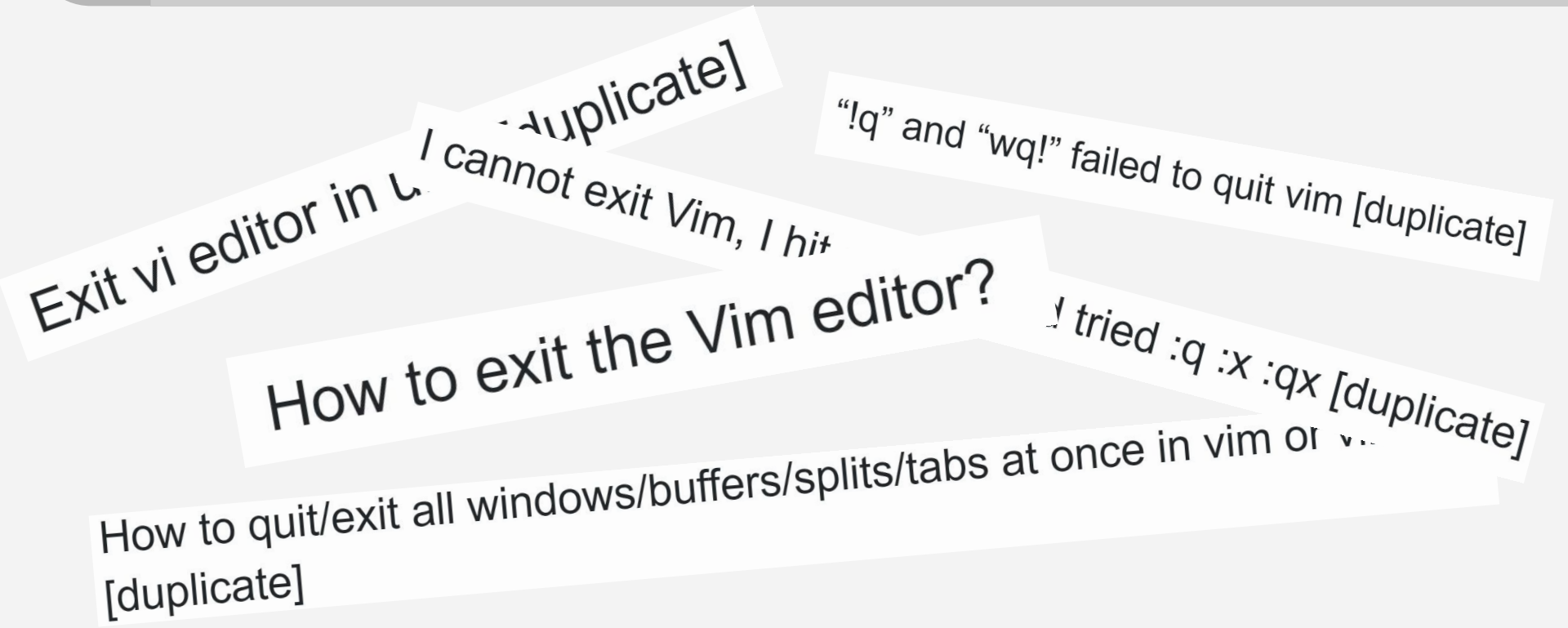
What, experiments we performed

How, details of the implementation

Given questions (Q1 | Q2)

- How to exit the Vim editor?
- Why can’t I exit Vim, I hit escape and tried :q :x :qx?

Are these questions *Duplicates* ?


Exit vi editor in
Duplicate]
I cannot exit Vim, I hit
"!q" and "wq!" failed to quit vim [duplicate]
How to exit the Vim editor?
I tried :q :x :qx [duplicate]
How to quit/exit all windows/buffers/splits/tabs at once in vim or
[duplicate]

- Information duplication
- Bad user experience for both question seekers and writers
- Moderators are required to scrutinize posted questions

Introduction | Data Description

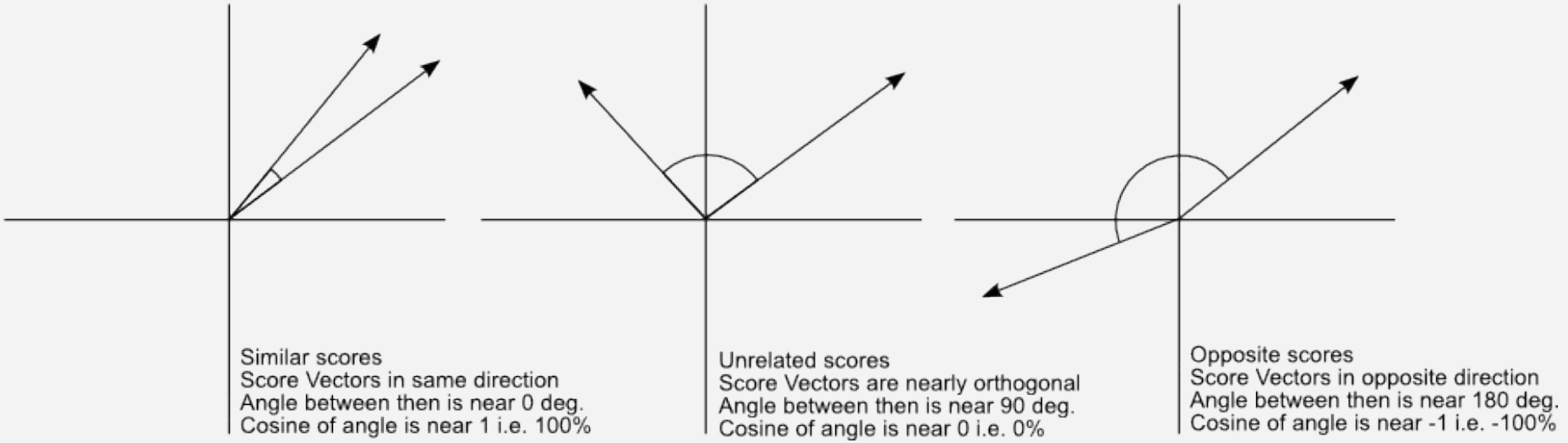
We have used dataset provided by *Quora on Kaggle platform*

Data Description	**Comments**
Question Distribution	404,290 (Duplicates 149,263, Non Duplicates 255,027)
Features	*Tuple* (qid1, qid2, question 1 text, question 2 text) *Class:* Label Binary (1/0)
Mean question length	Question 1 = 59, Question 2 = 60
Median question length	Question 1 = 52, Question 2 = 51
# rows with missing data	3
Duplicate rows	0

6 Related Work

- Bogdanova, Dasha et al. “**Detecting Semantically Equivalent Questions in Online User Forums**.” CoNLL (2015)
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning **GloVe: Global Vectors for Word Representation**. (2014)
- Matt J. Kusner , Yu Sun , Nicholas I. Kolkin , Kilian Q. Weinberger, **From word embeddings to document distances** (2015)
- Jonas Mueller. Aditya Thyagarajan, **From Siamese Recurrent Architectures for Learning Sentence Similarity**, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16).

Duplicate questions have more ***cosine similarity*** than non-duplicate questions



Duplicate questions have more ***cosine similarity*** than non-duplicate questions: 63% accuracy

Question Type	Median Cosine Similarity
Duplicate	0.69
Non-duplicates	0.48

9 Methods | Cleaning data

Missing data, Duplicates, Non-alpha	Tokenize, Stop words	Stemming, Lemmatization
• Remove data with missing values • Remove duplicate rows • Remove non-alphanumeric characters	• Tokenize words • Remove stop words	• Stemming of words • Lemmatization of words • Vectorizing words

Methods | Improving baseline

K-NN Classifier

Apply KNN classifier

3-NN has the most favorable results

Feature Engineering

Created feature such as noun count similarity

This simple feature had improvements

N-grams

Use n-grams while vectorizing questions

WordNet

Find semantic similarity of two questions using WordNet

Use Word2Vec for vectorization

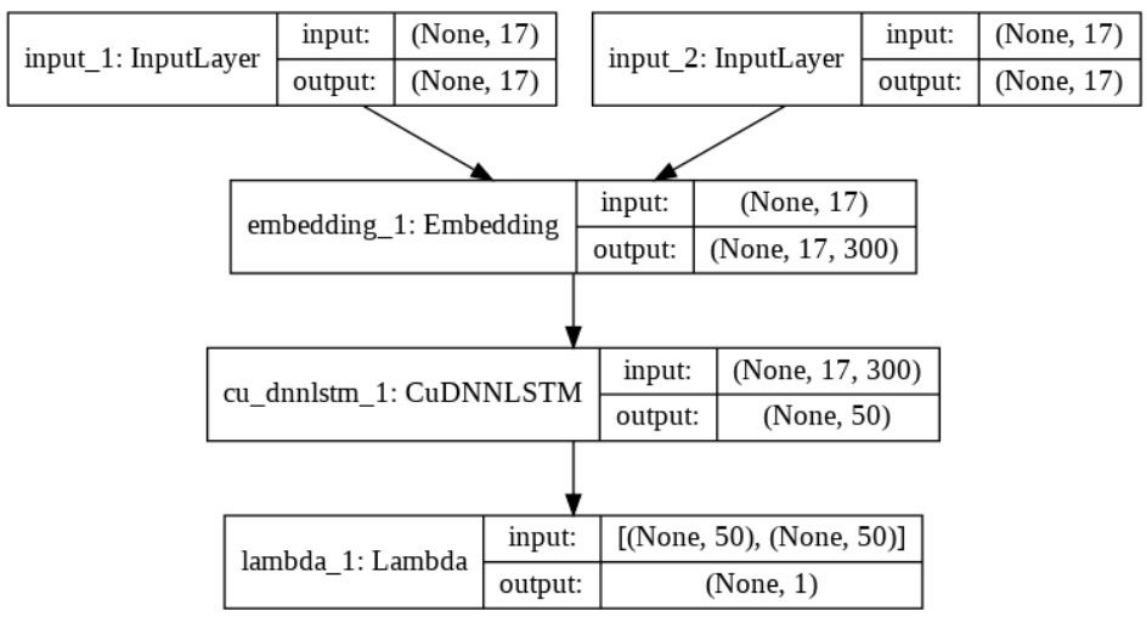
Methods | LSTM Neural Network

- What are LSTM Networks
- Motivation behind using LSTM
- LSTM implementation approach in current problem space

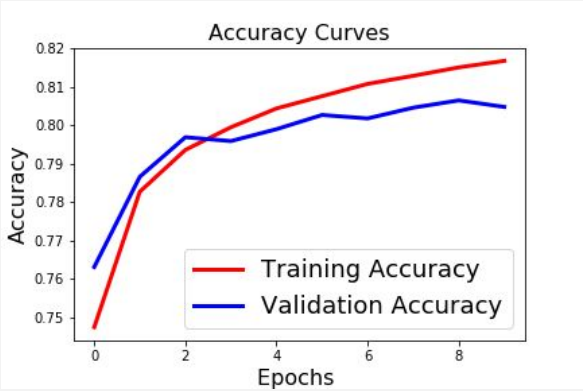
Hyperparameters | LSTM Neural Network

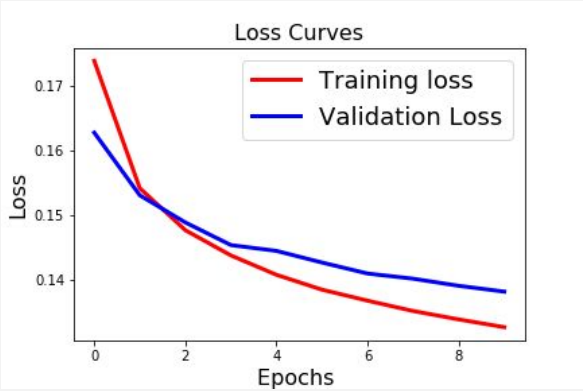
HYPERPARAMETER	VALUE
Number of Hidden Layers	50
Number of epochs	10
Optimizer	Adadelta
loss	mean_squared_error
Gradient clipping norm	1.25

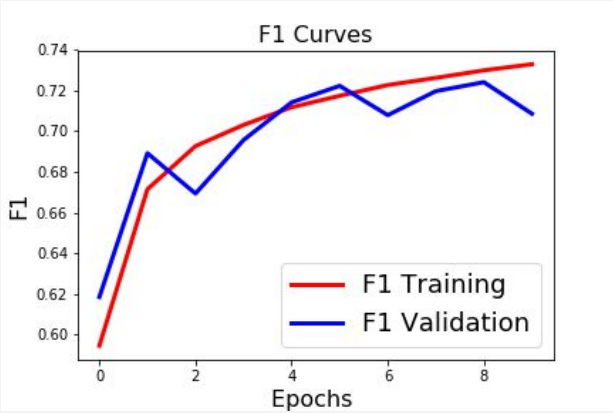
LSTM Network Graph



Performance vs Epochs | LSTM Neural Network







Results

Method	Accuracy	F1 Score
Cosine Similarity	63%	0.59
3NN + Feature Engineering	67%	0.46
SVM	65%	0.50
Semantic Similarity Wordnet	72%	0.51
Word2Vec + WMD	67%	0.61
MaLSTM Neural Network	**82%**	**0.72**

16 Future Scope

- Neural Network have used pre-trained word vectors
- It can be fine tuned further as follows
  - With more data
  - Train word embeddings
  - Explore other Gradient Descent Optimizers

17 Discussions / Learning

- Explored and gained insight into many classification algorithms
- Even simple feature engineering can improve accuracy
- How to implement Neural Networks
- How to tune and select hyper parameters
- Usage of Google Colab, TPU and CUDA for computational intensive tasks

Thank You