Project Title:

Sentence to Sentence semantic similarity

Dataset:

Quora question pairs with similar questions marked

Project Idea:

Quora has millions of users that post questions, and lots of these questions are similar. Identifying these duplicated questions can provide a better experience for active seekers and writers. Our project idea is to use the NLP technology we learned in class to identify semantic similarities for these question pairs.

The approach you will use:

Data clean up: stemming, lemmatization, contractions etc

Feature extraction: Word Embedding vi word2vec or Glove

Deep learning model: Implement the neural network model that learns the similarity.

Model Optimization: Using several optimization approaches to improve model performance.

Evaluation: Evaluate our model based on metrics (Intrinsic/Extrinsic).

Results visualization: Visualization of training results.

Software you will use:

Python, PyTorch

References:

Dataset: https://www.kaggle.com/competitions/quora-question-pairs/overview

Teammates:

Ruihao Wei, Jiahui Cao

Timeline:

- Data preparation, May 14, Week 6
- Word Embedding, May 21, Week 7

- Implement neural network, May 28, Week 8
- Evaluate and Optimization, Jun 4, Week 9
- Presentation, June 8, Week 10
- Final Report, June 15, Week 11