MGT 8803: Understanding Markets with Data Science
Project Proposal

Project Title:

Predicting Question Pairs in Discussion Forums using Manhattan LSTM networks

Motivation:

This project is a sentence similarity prediction exercise. Sentence similarity is a common problem faced in the industry, where the goal is to identify similar pieces of text to avoid duplication of material on online forums such as Stack Overflow, Quora etc. The idea is to allow the user to be able to identify if a similar question already exists before they end up creating another duplicate entry, as this not only allows for keeping the content on the webpages clean and curated but also helps in reducing the system and servers loads by reducing the amount of data. This will also enable the user to get their question answered promptly.

For this project we intend to predict Quora Question Pairs i.e., predict which of the provided pairs of questions, from the online platform Quora, contain two questions with the same meaning.

Dataset:

For this problem, we have obtained a Kaggle dataset containing Quora question pairs as text, along with the label indicating whether the questions are similar or not.
Quora Question-Pairs Dataset

Methodology:

Since the problem involves capturing the inherent “meaning” of two given pieces of text, instead of directly comparing words or tokens and measuring their similarity, we aim to use a recurrent neural network framework in the form of Siamese-Manhattan LSTM (MaLSTM) networks. LSTM networks are typically useful for capturing the information (or meaning) and learning order dependence in sequence prediction problems (in this case text data). Siamese neural network is a network architecture used to tackle one-shot learning problems, where the goal is to identify whether two inputs are similar or not, as compared to traditional classification tasks where you assign a class to each input. Further, we shall use Manhattan distance in conjunction with the Siamese LSTM architecture in order measure similarity.

Group Members:

- Saurabh Aggarwal
- Siddharth Sen