BI CAPSTONE PROJECT PROPOSAL
CSC591 – ALGORITHMS FOR DATA GUIDED BUSINESS INTELLIGENCE

Team Members:
Rahul Gutal : rmgutal
Kushal Nawalakha : kbnawala
Animesh Sinsinwal : assinsin

QUORA QUESTION PAIRS: IDENTIFY QUESTION PAIRS THAT HAVE THE SAME INTENT

1. **Learning Objectives**
   a. Main Theme
   We have touched a challenging problem in natural language processing and machine learning – Mitigating the inefficiencies of having duplicate questions pages at scale, we need an automated way of detecting if pairs of questions text actually correspond to semantically equivalent queries.

   What is the objective behind our topic?
   An important product principle for Quora is that there should be a single question page for each logically distinct question. With duplicate pair of questions, Quora provides two different threads for solutions. So, this approach should help Quora team to merge these two threads in future.

   b. Covered Topics
   Word2vec
   Word2vec is a two-layer neural net that processes text. Input given is a text corpus and result is a feature vector for words in the corpus. Mostly feature vector is in numerical form that deep nets can understand.

   Glove
   We are using pre-trained Glove model which comes with spacy, which was trained on Wikipedia which was stronger in terms of word semantics.

   Siamese Neural Network
   Siamese neural networks is a class of neural network architectures that contain two or more identical subnetworks. Siamese Neural Networks are popular for tasks that involve finding similarity or a relationship between two comparable things.

   c. Research Papers
   We went through some research papers to get the gist of the techniques we are using. Thus, made
   1. https://papers.nips.cc/paper/769-signature-verification-using-a-siamese-time-delay-neural-network.pdf
   2. https://arxiv.org/pdf/1512.05193v2.pdf
   3. https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf

4. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

d. Quantitative Understanding – Proc and Cons of performance metrics
Pros of using Word2vec:
- Given enough data, usage and contexts, Word2vec can make highly accurate guesses about a word's meaning based on past appearances.
- The usefulness of Word2vec is to group the vectors of similar words together in vectorspace, i.e. it detects similarities mathematically.
- Word2vec creates vectors that are distributed numerical representations of word features, features such as the context of individual words. It does so without human intervention.

Pros of using Siamese Architecture:
- Sharing weights across subnetworks will have less parameters to train, which in turn means less data required and less tendency to overfit.
- Each subnetwork essentially produces a representation of its input. By this, we will have representation vectors with the same semantics, making them easier to compare.

2. BI Use Case

Today, internet has answers for billions of questions. As an end user and the platform hosting the questions like Quora, Stackoverflow, etc. have always faced the issue of multiple duplicates of logical questions.
Through this project, we want to address this issue and provide an efficient machine learning technique to detect this duplicity.

3. Data Sets

Dataset taken from – http://qim.ec.quoracdn.net/quora...
- Dataset consists of over 400,000 lines of potential question duplicate pairs.
- Each line contains IDs for each question in the pair, the full text for each question, and a binary value that indicates whether the line truly contains a duplicate pair.

Constraints and Keynotes:
- Original sampling method by Quora returned an imbalanced dataset with many more true examples of duplicate pairs than non-duplicates. Therefore, they added negative examples. So, the whole dataset is not truly semantically equivalent.
- The ground-truth labels contain some amount of noise: they are not guaranteed to be perfect.

4. Demo

Demo will consist of real-time prediction of question duplicity. You can provide pair of questions in an input file and we will generate corresponding output file stating whether questions are duplicates or not.

5. Methods/Algorithms and packages

Proposed Algorithm:

- We will be using either Word2vec or GLOVE to create feature vectors for the questions to be compared.
- We will be dividing the feature vectors of question-pairs into training data and testing data.
- We will train Siamese NN model and will predict the outcome for test data.
- At last we will compute the accuracy.

We are using the below libraries in our python code:
numpy, keras, sklearn, tf-idf, tqdm, genism, pandas, spacy

6. Contribution

Rahul Gutal
Working on Word2Vec to generate feature vectors of questions. Feature vectors are enhanced by using TF-IDF scores.

Kushal Nawalakha
Working on GLOVE to generate feature vectors of questions.

Animesh Sinsinwal
Working on Siamese neural networks which will take feature vectors of question as input and decide whether the questions are similar.