**Identifying Duplicate Question Pairs in Quora.**

|  |  |
| --- | --- |
| Ankit Murarka | Balaji Radhakrishnan |
| [amurark@ncsu.edu](mailto:amurark@ncsu.edu) | bradhak@ncsu.edu |

1. **To examine and assess the business value of the Quora Dataset.**

The dataset includes the following versatile business values:

* Users seeking answers on the internet can be shown suggested questions asked by other users with the same intent.
* This can optimize search results, wherein partial questions written as the text phrase can be analyzed and matched with similar questions previously asked.
* Reducing redundancy by ensuring that each unique question exists only once will help build a high-quality knowledge base.
* Writers wouldn’t have to write the same answer to multiple versions of the same question hence reducing human labor.

1. **The list of business questions that could use data science techniques using Quora dataset are:**
   * Recommended similar questions: Recommend questions with similar meaning and intent which have been asked by other users when a user searches for a question. This can reduce the possibility of users asking similar questions.
   * Accumulate the answers to similar questions on the same page. This enhances the user experience of the users who are seeking answers to questions.
   * Avoiding redundant questions with similar meaning will ensure that writers wouldn’t have to write the same answer to multiple versions of the same questions.
2. **To find the survey papers and tutorials that summarize the most recent R&D advances in Sentiment Analytics and Recommendation Systems.**

We followed a few tutorials which are listed below but the main source of inspiration behind our approach and implementation of the Deep Learning Model was the tutorial by Abhishek Thakur. The link for the same is given below:

https://www.linkedin.com/pulse/duplicate-quora-question-abhishek-thakur

Some of the other tutorials we looked during the research phase include the following:

<http://www.mit.edu/~jonasm/info/MuellerThyagarajan_AAAI16.pdf>

<https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning>

https://www.kaggle.com/sudalairajkumar/quora-question-pairs/simple-exploration-notebook-quora-ques-pair

1. **To propose a target question your team will address using the Quora dataset with Sentiment Analysis.**

Quora is a place to gain and share knowledge. It’s a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and better understand the world.

Quora has over 100 million users. A lot of curious users leads to a lot of questions being asked each day. As of March, 2017, 13,304,529 questions have been asked on Quora. This has led to a humongous amount of data. With a corpus that large, there arises a possibility of duplication or redundancy. People coming from different backgrounds can ask a question with the same intent but can use completely different semantics and sentence structures to frame their questions. Multiple questions with the same intent can cause users to spend more time finding the best answer to a question. It even forces writers to answer multiple versions of the same question.

1. **To justify the possible business value behind answering such a question.**

The solution to this question will have quite an impact on the business value. This will make the lives of both sets of quora users better. The ‘answer seekers‘ and the ‘writers’. While the writers don’t have to answer multiple iterations of the same type of questions, the answer seeker can intuitively search through a collection of recommended similar questions which were already asked by other users. Hence preventing them from asking a redundant question. An intuitive way forward will ensure a greater User Base.

1. **To discuss how you may evaluate the quality of different answers: both from business and technology perspectives.**

We can evaluate the perspective of the proposed target question as follows:

* Business Perspective:
  + Users increasing: An intuitive and a user-friendly interface will only invite more intelectuals to the site hence increasing the user base.
  + Revenue: More users meaning more advertisements from sponsors and clients. This leads to increase in revenue.
* Techincal Perspective:
  + With an intuitive recommender system for similar questions and a best-in-industry duplicate-question-finding algorithm, this solution can go a long way in establishing Quoras hold in this kind of a market.

1. **The 2 papers that we read during the research phase were:**

Paper 1: A Decomposable Attention Model for Natural Language Inference

This paper proposes an approach that uses attention to decompose the problem into subproblems that can be solved separately, thus making it trivially parallelizable. This approach achieves impressive results even without considering the word order information. This paper creates embedding vectors for each of the two sentences of a fixed dimension d. The vanilla version which uses the sentences as is doesn’t rely on word order. There is a discussion about extending this further by including intra sentence attention, thereby utilizing a minimal amount of sequence information. This approach is split into three phases namely: attend, compare and approach. In this entire approach, the input representations are always single word embeddings. They utilized 300-dimensional GloVe embeddings to represent each word. One interesting aspect here was that out of vocabulary words were hashed to one of 100 random embeddings. All embeddings remained fixed during training. Each hyperparameter setting was run on a single machine with 10 asynchronous gradient-update threads. Dropout regularization (Srivastava et al., 2014) was used for all ReLU layers, but not for the final linear layer.

Paper 2: Siamese Recurrent Architectures for Learning Sentence Similarity

This model uses word embedding vectors supplemented with synonymic information to the LSTM’s, which use a fixed length vector to encode the underlying meaning expressed in a sentence. By restricting subsequent operations to rely on a simple Manhattan metric, they compel the sentence representations learned by the model to form a highly structured space whose geometry reflects complex semantic relationships. The core of this approach is the proposed Manhattan LSTM. Here each sentence of a given pair is fed to a LSTM. But the focus was on Siamese architectures which uses identical weights for each of the LSTM’s. This model predicts relatedness for a given pair of sentences. Training in this network happens via backpropagation through time under the mean squared error loss functions.

1. **To describe which features you have used from the `Quora` dataset**

We used Quora’s first public dataset to resolve the problem of identifying duplicate questions. There are over 400,000 lines of potential duplicate pairs. The dataset can be found using the following link:

https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

The dataset contains the following features:

* Id: The question pair ID.
* qid1: ID for question 1.
* qid2: ID for question 2.
* question1: The text for question 1.
* question2: The text for question 2.
* is\_duplicate: The output flag indicating if the questions have the same meaning and intent.

Some of the features of the dataset are:

* The original sampling method returned an imbalanced dataset with many more TRUE examples of duplicate pairs than non-duplicates. Therefore, we supplemented the dataset with negative examples. One source of negative examples were pairs of “related questions” which, although pertaining to similar topics, are not truly semantically equivalent.
* The ground-truth labels contain some amount of noise: they are not guaranteed to be perfect.

1. **To prototype a relatively simple, that solves this question from end-to-end perspective.**

We propose to pursue the challenge to tackle this NLP problem by applying a Deep Learning RNN Model to classify whether question pairs are duplicate or not. Using a Deep Neural Network comprising of two translation layers, one for each question. The different layers in the Model consists of Embedding using GloVe, an LSTM, a GRU and another LSTM in that order. This if followed by merging the layers for question 1 and question 2, adding a layer of batch normalization, a dense layer, a ReLU activation function, another dense layer to convert vectors to size 1 and a final activation(sigmoid) layer to determine the polarity of the output.

**GloVe**: Global Vectors for Word Representations. An unsupervised learning algorithm for obtaining vector representations for words.

**LSTM**: Long Short Term Memory networks, - usually just called “LSTMs” are a special kind of RNN capable of learning long-term dependencies. Like most RNNs, a LSTM network is [universal](https://en.wikipedia.org/wiki/Turing_completeness) in the sense that given enough network units it can compute anything a conventional computer can compute, provided it has the proper [weight](https://en.wikipedia.org/wiki/Weight) [matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)), which may be viewed as its program. Unlike traditional RNNs, an LSTM network is well-suited to learn from experience to [classify](https://en.wikipedia.org/wiki/Classification_in_machine_learning), [process](https://en.wikipedia.org/wiki/Computer_data_processing) and [predict](https://en.wikipedia.org/wiki/Predict) [time series](https://en.wikipedia.org/wiki/Time_series) when there are time lags of unknown size and bound between important events.

**GRU**: A gated recurrent unit. GRUs are related to LSTMs as both are utilizing different ways of gating information to prevent vanishing gradient problem. The GRU unit controls the flow of information like LSTM unit, but without having to use a memory unit. It just exposes the full hidden content without any control.

We trained our model for different number of epochs. A detailed analysis of the different epochs is listed below.

* We initially trained our model for a single epoch and got a ***training*** ***accuracy*** of ***0.7450*** and a ***loss*** of***0.5049.*** It took around 15-20 minutes to train the model for 1 epoch.

*loss: 0.5049 - acc: 0.7450 - val\_loss: 0.4778 - val\_acc: 0.7596*

* We next trained our model for 10 epochs and got a ***training accuracy*** of ***0.9548*** and a ***loss*** of ***0.1140***.

loss: 0.5032 - acc: 0.7464 - val\_loss: 0.4743 - val\_acc: 0.7611

loss: 0.4364 - acc: 0.7880 - val\_loss: 0.4269 - val\_acc: 0.7909

loss: 0.3925 - acc: 0.8138 - val\_loss: 0.4239 - val\_acc: 0.7965

loss: 0.3518 - acc: 0.8370 - val\_loss: 0.3896 - val\_acc: 0.8167

loss: 0.3106 - acc: 0.8600 - val\_loss: 0.3867 - val\_acc: 0.8178

loss: 0.2667 - acc: 0.8833 - val\_loss: 0.3980 - val\_acc: 0.8245

loss: 0.2218 - acc: 0.9055 - val\_loss: 0.4304 - val\_acc: 0.8207

loss: 0.1795 - acc: 0.9255 - val\_loss: 0.4725 - val\_acc: 0.8215

loss: 0.1424 - acc: 0.9423 - val\_loss: 0.5297 - val\_acc: 0.8275

loss: 0.1140 - acc: 0.9548 - val\_loss: 0.5877 - val\_acc: 0.8286

In this case, we see that the validation loss started increasing after the 7th epoch while the training accuracy didn’t change much indicating a clear case of overfitting. Hence we decided to run the algorithm for one final time for 5 epochs. WE got an ***accuracy*** of ***0.7453*** and an ***error*** of ***0.5032***.

**SAMPLE RESULTS:**

* How does the Surface Pro himself 4 compare with iPad Pro?,Why did Microsoft choose core m3 and not core i3 home Surface Pro 4?

Duplicate probability prediction: 0.0458052

* By scrapping the 500 and 1000 rupee notes, how is RBI planning to fight against issue black money?,How will the recent move to declare 500 and 1000 denomination lewin illegal will curb black money?

Duplicate probability prediction: 0.92498994

* What are the how best books of all time?,What are some of the military history books of all time?

Duplicate probability prediction: 0.87769759

* What if a cricket hits a batsman’s helmet and then goes to the boundary?,Should carbonated red balls and 8 yellow balls. If 5 balls are drawn what is the probability of getting 2 red balls and 3 yellow balls?

Duplicate probability prediction: 0.10641973

* Is it gouging and price fixing?,"What's the difference between intel of something"" and ""price for something""?

Duplicate probability prediction: 0.38428876

**TEAM MEMBER CONTRIBUTION:**

* During the initial research phase, both the team members, Ankit and Balaji read various research papers and tutorials to gather information about the project and about the prospective approach.
* After the initial research phase, while Balaji started working on deciding the actual model structure and defining the number of layers and components in the model.
* Ankit started implementing the Deep Neural Network model using python3.