Cairo University

Faculty of Engineering

Department of Computer Engineering

Big Data Analytics

Project Proposal
Team 7

Mohamed Shawky Zaky
SEC:2, BN:15

Remonda Talaat Eskarous
SEC:1, BN:19

Mohamed Ahmed Mohamed Ahmed
SEC:2, BN:10

Ahmed Mohamed Zakaria
SEC:1, BN:3

1 Idea

Quora is an online platform to connect with people, ask questions and get answers. With over 100 million people visit **Quora** every month, there must be a huge amount of duplicate questions that have the same words or carry the same meaning. This can be very confusing for the answer seekers and results in a longer search time. The huge amount of questions on **Quora** urges the usage of `Big Data Analytics` techniques. Also, the *inference* pipeline of such problem can be fit into a `MapReduce` workflow.

Our system offers a way to *cluster* and *classify* questions based on their content, in order to relate duplicate questions with each others for easier and quicker search.

2 Dataset

We are going to use the *dataset* from **Quora Question Pairs** `Kaggle`'s competition. The *dataset* consists of *questions pairs* with *labels* for whether they are related or not.

Dataset information :

- **Link :** https://www.kaggle.com/c/quora-question-pairs
- **Size :** $404,289$ `question pairs`.
- **Features :** `to be extracted from given questions text`.

3 Proposed Solution

In order to tackle the described problem, we need the following :

- *Find* the closest match(es) for a given question.
- *Predict* whether two questions are related or not.
- *Cluster* related questions together based on their content.

The solution is divided into **two** stages :

- Development stage *(non-distributed)* :
  - **Text processing and visualization :** *tokenization*, *stemming* and *statistics gathering*.
  - **Feature extraction :** *bag of words*, *n-grams* and *word embeddings*.
  - **Model training :** *KNN* `(for closest match finding)`, *XGBoost* `(for predicting relations)` and *K-Means* `(for clustering questions)`.
- Inference Stage *(pseudo-distributed)* :
  - The trained models are to be deployed in a **Hadoop** `MapReduce` environment for distributed batch processing and inference.