Team Members:

1. Isaac Moris Habib

2. Beshoy Hani Badee

3. Mariv Ibrahim Abd El-Maseeh

4. Youssef Nagy Kamal

Sentence to Sentence semantic similarity

Team: 6

Project Title

Introduction:

Identifying similar questions which have the same meaning is a challenge nowadays. Everyday thousands of questions are being asked on different forums like Quora and Stackoverflow etc., lots of them are similar or asking for the same thing. Having a system that can identify similar questions means getting response faster than usual and sometimes in no time! Great, isn’t it?

Using NLP and Machine Learning techniques we will implement this system.

Methodology:

A set of steps were followed in order to achieve the best result. These steps are:

1. Studying and analyze the dataset.
2. Extracting information from the dataset.
3. Applying NLP approaches based on the extracted information, which includes:

* Replace all numbers with a unified number (number 1), since all numbers refer to the same intention.
* Reconstruct shortened expressions to their origin. For example: we’ve => we have.
* Apply lemmatization.
* Remove stop words.
* Remove punctuation.

1. Convert all documents into a matrix of token counts in order to be represented for ML model.
2. Split dataset into training and testing with 70% training and 30% testing.
3. Train the model with the training data.
4. Test the model with the testing data.

Data Set Summary:

1- The used dataset is Quora Question Pairs dataset.

2- Dataset description:

* The dataset consists of 6 columns as follow:

1. id: a unique id for each record in the dataset.
2. qid1: a unique id for each question in question1 column.
3. qid2: a unique id for each question in question2 column.
4. question1: first question in the pair questions.
5. question2: second question in the pair questions.
6. is\_duplicate: either the two questions have the same meaning (will be 1) or not (will be 0)

* Dataset consists of 404290 rows.
* There are 149263 records with question pairs have the same meaning and 255027 records haven’t.

3- Visualization of the dataset statistics

* Dataset Shape:

Graphical user interface, text, application

Description automatically generated

* Snapshot of the dataset Graphical user interface, application

  Description automatically generated
* Information of each column in the dataset Graphical user interface, text, application

  Description automatically generated
* is\_duplicate column plotting Graphical user interface, text, application

  Description automatically generatedChart, bar chart

  Description automatically generated

Results:

Several models with different parameters were experimented. The following table illustrates these experiments

|  |  |  |
| --- | --- | --- |
| Model | Parameters | Snapshot |
| XGB Classifier | Learning Rate: 0.1  Max Depth: 50  n estimators: 80 |  |
| XGB Classifier | Learning Rate: 0.2  Max Depth: 100  n estimators: 80 | Table  Description automatically generated |
| Random Forrest Classifier | Max depth : 100 | Table  Description automatically generated |
| Keras Sequential (NN) | Epochs: 30  Optimizer: sgd | A screenshot of a computer  Description automatically generated with low confidence |

Chart, bar chart

Description automatically generated