Duplicate Question Identification System

1. Software Requirement Specification

The identification of similar questions to the ones entered by the user. The system will display a list of questions that matches the context of the user query. The system could also determine if two questions entered by the user are similar or not.

2. Dataset Description

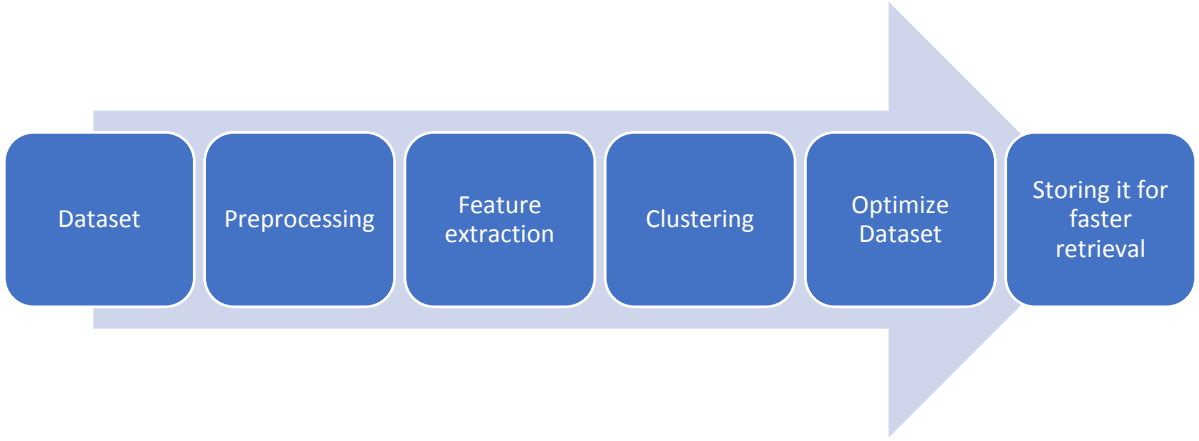
The dataset comprises of comma separated values with over 4 lakh question pairs, each question identified by a unique ID and is_duplicate column which indicates if the pair of questions in reference are similar or not. The questions are mainly from the Q&A based website quora where users tend to ask a lot of questions. The train/test split has been created for benchmarking.

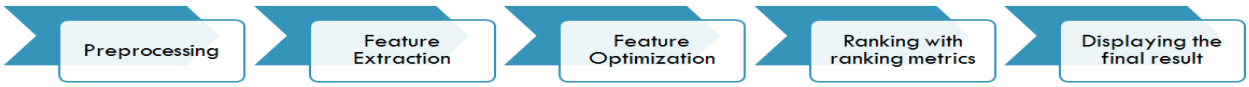
3. Proposed Approach

The model consists of two modules in which the output of first module is fed as input for the second. The module 1 follows a clustering approach using NLP available with NLTK. This gives an intermediate result with a set of 10 questions which is fed as input for second module which follows a Deep learning approach and uses LSTM model to get the similarity scores.

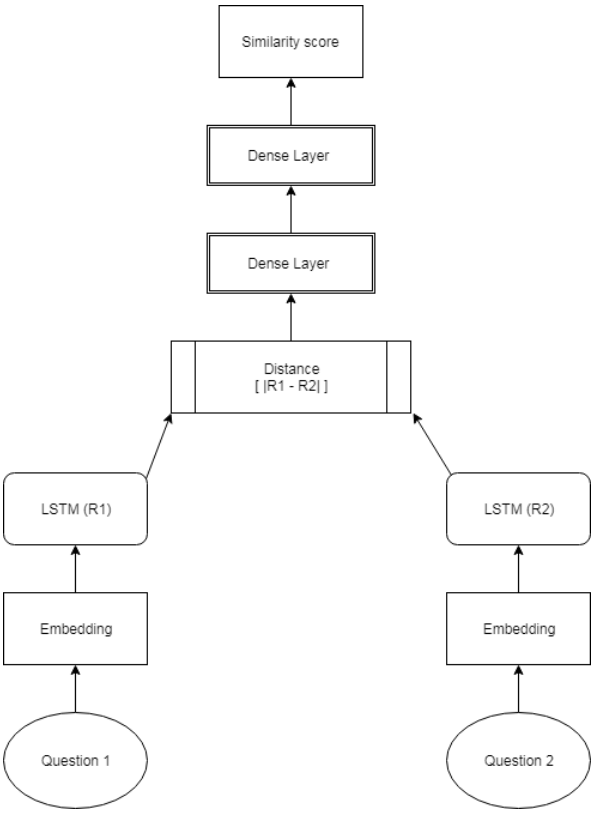
a. Design

Module 1





Module 2



b. Data preprocessing

- Convert to lower case

    i. What's the alternative to Machine-Learning ?

ii. what's the alternative to machine-learning?

- Removing vague characters

  i. what's the alternative to machine-learning?

  ii. what the alternative to machine learning

- Tokenization

  i. Splitting sentence into words

  ii. Generating ID for unique tokens

- Padding

  Make the length of all the questions same by adding empty spaces.

c. Technologies used

- NLTK

  Natural language toolkit used extensively in natural language processing to accomplish tasks such as cleaning and preprocessing of textual data, extraction and tagging of features which can be understood by a computer.

- Tensor Flow

  TensorFlow is an open-source software library for dataflow programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google, often replacing its closed-source predecessor, DistBelief.

- Keras

  Keras is an open source neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit or Theano. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible. It was developed as part of the research effort of

project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System), and its primary author and maintainer is François Chollet, a Google engineer.

- Flask

  Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

4. Test Plan

The dataset is subjected to a 75-25 split where 25 % of the dataset is used for validation. The resulting accuracy was approximately 84%.

a. Unit testing

Each module is fed with input and checked for the accuracy of the output. The module 1 uses jaccard similarity to produce results.

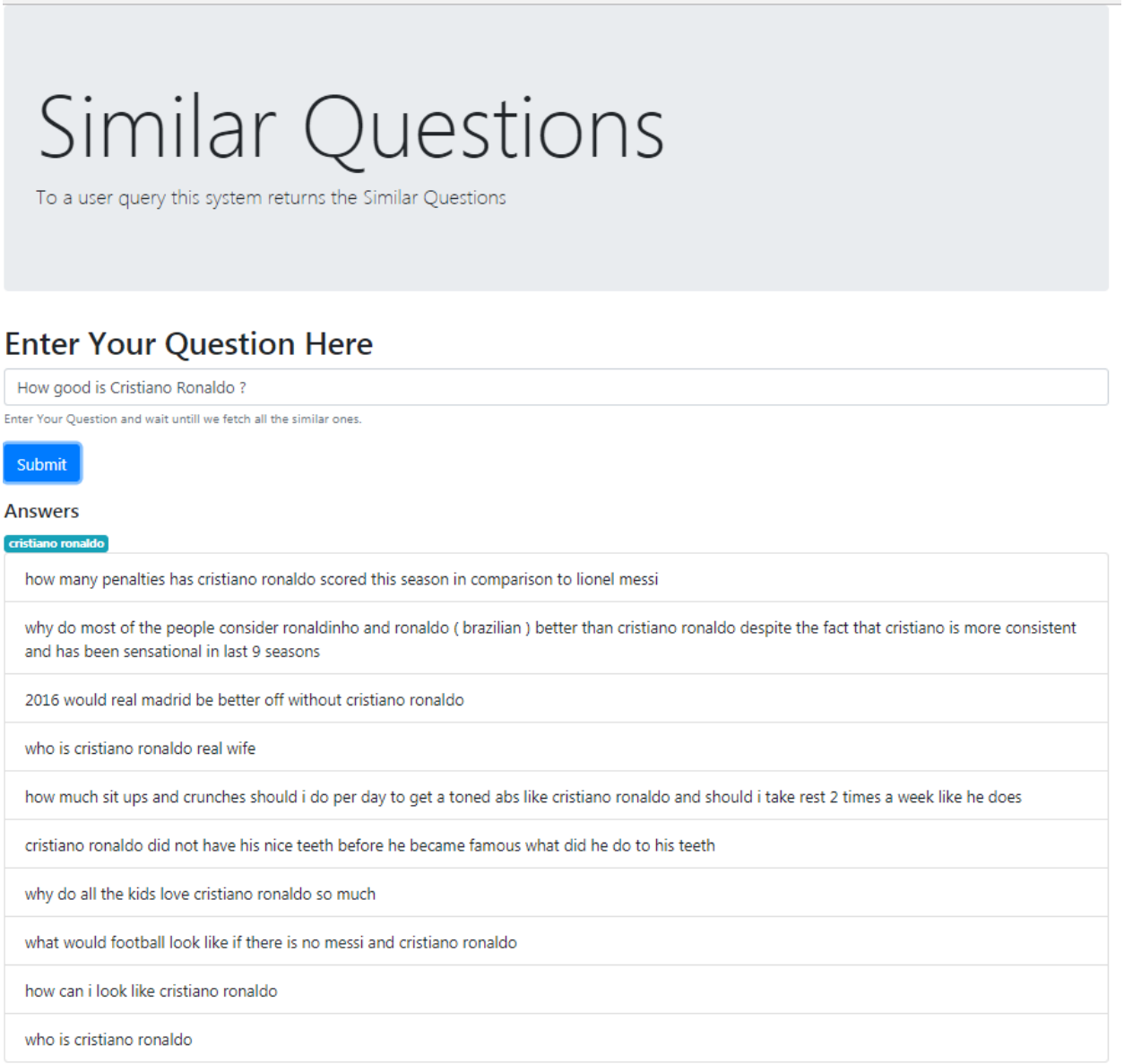
b. Integration testing

The system is integrated along with a graphical user interface. Also the input for Module 2 is fetched from the output of module 1.

5. Final Solution

Module 1 Results

The module one selects the best suited 10 questions contained in the dataset on the basis of noun phrases, nouns, adverbs and adjectives and stores the results in a list.



Module 2 Results

The output of module 1 is fed as the input for module 2. The module two compares each of the 10 questions in the list with th2 given user input and calculates the similarity scores which if beyond a threshold value is considered as similar.

6. Conclusion and Future work

The system was able to successfully fetch the questions based on the noun phrases, nouns adverbs and adjectives. Also, the fetched list was checked for similarity using deep learning and the accuracy was found to be fetching the approximate similarity scores. The system can be further improved by customizing POS tagging and improved context sensitive search.