Quora Question Pairs

Sukanya Behera [160714733008]
Chandrabhatta Sriram [160713733042]
Abul Faiz Mohammed Abdul Hadi Mouzzam [160714733108]

Project Guide: E. Shailaja [Asst. Prof. Dept. of CSE]

Abstract

- Quora is a place to gain and share knowledge—about anything. It’s a platform to ask questions and connect with people who contribute unique insights and quality answers.
- Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question(s), and make writers feel they need to answer multiple versions of the same question.
- The goal is to predict which of the provided pairs of questions contain two questions with the same meaning.

Introduction [Problem Definition]

- More formally, the duplicate question detection problem can be defined as follows: given a pair of questions *q1* and *q2*, train a model that learns the function:
  - *f(q1, q2) → 0 or 1*
  - where *1* represents that *q1* and *q2* have the same intent and *0* otherwise.

Let’s Talk Data, Shall We?

- The Dataset being analysed/solved is taken from kaggle.com (a popular website for hosting data science/mining competitions), which was provided by Quora.
- It contains more than 400,000 data points.
- File Type: .CSV [Comma Separated Value].
- File Size: 130 Mega Bytes [Uncompressed].

Dataset Snapshot

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

Existing System

- The Existing System had a production model for solving this problem as a Random Forest Model with tens of handcrafted features, including:
  - the number of common words,
  - the number of common topics labelled on the questions,
  - and the part-of-speech tags of the words.

Proposed System

- The Proposed System will be a first cut solution to predict whether the given questions are similar or not, using ML & AI techniques like:
  - Exploratory Data Analysis [Summarizing the main characteristics of data, using scatter plots].
  - Dimensionality Reduction [Pre-processing the questions by stop word removal, stemming, lemmatization].
  - Data Classification & Regression techniques to classify our data.

Algorithms & Technologies Used

- Some of the Machine Learning Algorithms used are:
  - Linear Regression,
  - Bag of Words [BoW], NLP, Document Extraction from Document Corpus,
  - k – Nearest Neighbours, etc.
- Technologies used:
  - Python 3.6.x with Anaconda Package Manager.
  - Code will be written and documented in Jupyter Notebooks [formerly known as “iPython Notebooks”].
  - Machine Learning Packages used are:
    - numpy, matplotlib, pandas, scipy, nltk, etc.

Thank You!

Any Questions/Queries?