Project Proposal - Curated FAQs

CS410

Student Details:

Name	Fnu Kishan Borule Rahul
NetId	fnuk2@illinois.edu
Team members	Working Alone
Captain	Fnu Kishan Borule Rahul
Team Name	NFSMW
Project Name	Curated FAQs

Task:

Build a system to retrieve Frequently Asked Questions (FAQs) that users have about certain consumer products (Like Pampers -Diaper Brand, Tide - Detergent etc) from twitter and rank them to provide the relevant resolution to the problem.

We often see that consumer product goods (CPG) companies are active on their social media sites (like twitter) for answering the queries from customers. Building a system to automatically retrieve the nearest problem that people have in mind with the resolution to the problem will have multiple points of impact. We can see that customers can effectively resolve their query with no wait-time from the company's representative to answer the query and moreover, the agent will have less questions to answer and henceforth, be more effective in their job of serving the customers better.

I will describe the system's implementation in the next section. I plan to use python as the programming language for implementation. Further, I assume that I will require 40 hrs of quality time to complete this project.

Current assumption on the tools required:

- Twitter Scraper - https://github.com/taspinar/twitterscraper
- Sentence Embedder - https://github.com/UKPLab/sentence-transformers
- Indexer (To create data index) - Faiss or Annoy
- Display the working prototype on streamlit or command line

Project Description:

Part 1 - Data Set Creation:

Identify the twitter handles where company's representatives are most active and scrape the data from these handles. We will need to further clean this data with some text mining heuristics to retain a good diversity of questions for our data set. We will split this data into two groups of train and test set (90% and 10% split respectively). The 10% of test set data is purely used for evaluation purposes.

Part 2 - Exploration of Vector Space Model

We need to explore and identify the deep learning based pretrained sentence embedding model available here - https://github.com/UKPLab/sentence-transformers and in the repository of hugging face. I will optimise the trade-off between the speed of model evaluation on cpu with the sentence accuracy mentioned for these models and choose a model that works the best in this scenario. Further, I will choose 3 more models that are optimised for high accuracy and reserve them for evaluation. As the evaluation is done offline, I don't have to worry about the inference time.

Part 3 - Indexing

Once the sentence embedding model is chosen, I will index all the questions and store the relevant mapping to answers in an open-source indexer like annoy or faiss. This will help me retrieve the relevant questions quickly. Further I will create an API for this indexer so that it can work with any system.

Part 4 - Interface

I will create an interface for the users to interact with the whole system. Steam-lit provides a clean web-interface for users to interact with the system.

Evaluation:

As mentioned in the part 2 of the project description, we will select three independent sentence embedding models and evaluate the similarity score with 10% of customer queries (held out test set) . We will report average similarity scores (with geometric mean average) retrieved from these independent models for the test set.