StackOverflow Data Visualization

E6893 Big Data Analytics

Goal & Novelty

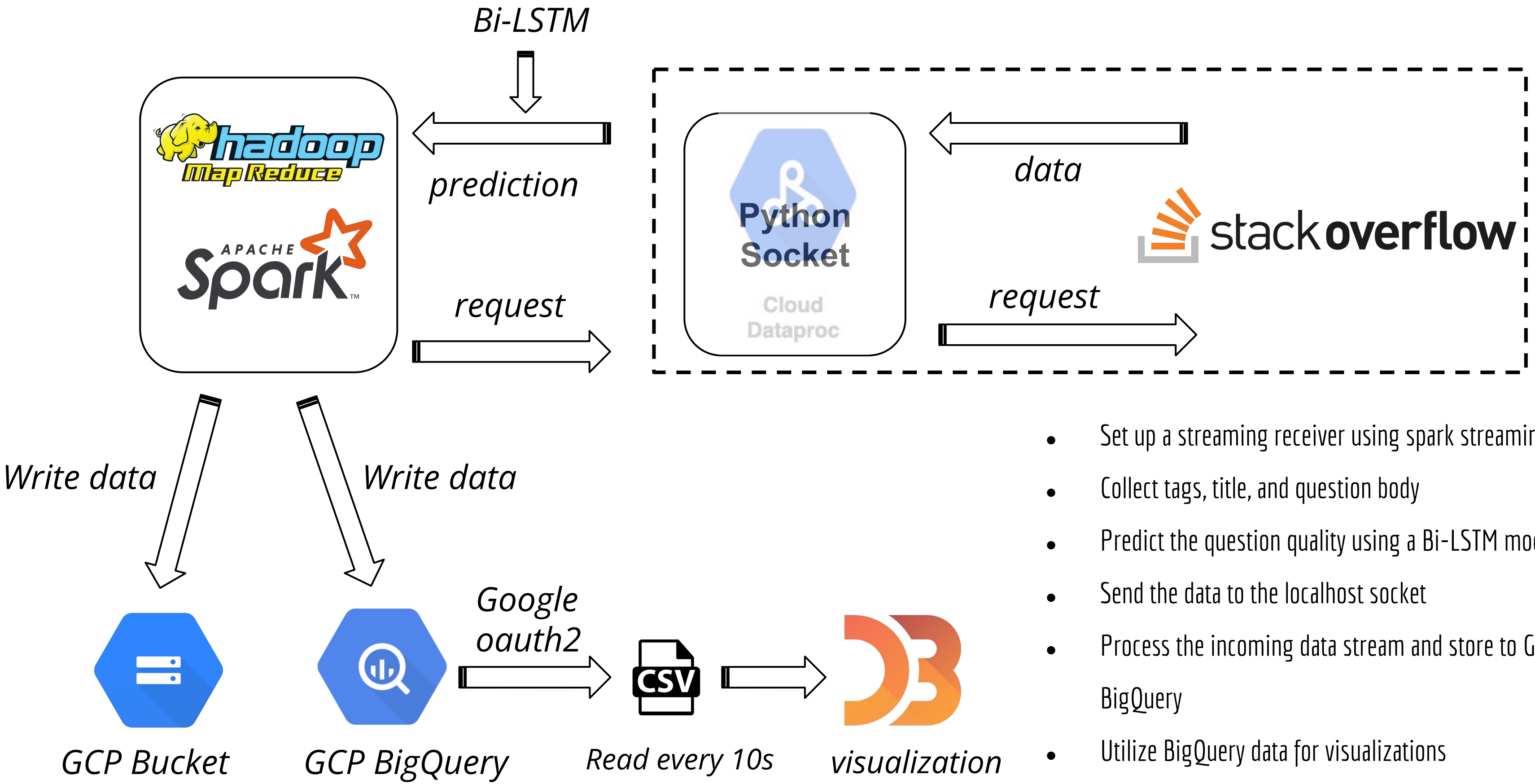
The goal of our project is to get streaming data from StackOverFlow.com and obtain a real-time representation of what topics and questions are being asked by users around the world. This can used to visualize and analyze the hot trend in programming and even identify common problems in a topic/framework. The proposed project should be considered novel because it creates an intuitive way for the users to visualize and interact with the live trend on stackoverflow.

Data

- The Stack Exchange API enables users to retrieve answers, comments, badges, events, questions, revisions, suggested edits, user information, and tags from a Stack Exchange based website.
- The API uses REST calls issued in JSON and JSONP.
- StackAPI is a simple Python wrapper for the Stack Exchange API.
- Stack Exchange API requests are limited to 300 per day without an API key - can be increased to 10000 per day with an API key. We have registered and received an API key for this project
- **Real time streaming data obtained** from StackAPI every 10 seconds and recording all new posts

Volume	Velocity	Variety
21 million question asked	13.6s between each question	4000+ potential question categories

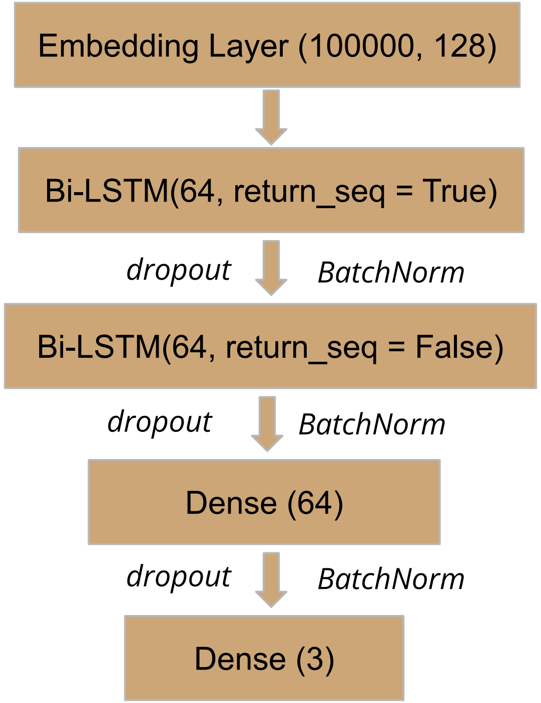
System - Data Gathering, Deep Learning & Data Processing (MapReduce/Storage)



- Set up a streaming receiver using spark streaming context
- Collect tags, title, and question body
- Predict the question quality using a Bi-LSTM model
- Send the data to the localhost socket
- Process the incoming data stream and store to Google BigQuery
- Utilize BigQuery data for visualizations

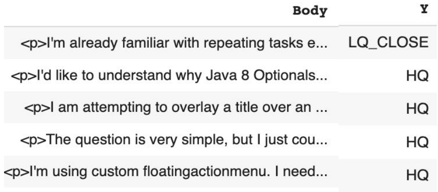
Technologies - Bi-LSTM Model

- Kaggle's "60k Stack Overflow Questions with Quality Rating" dataset, which contains 60,000 question samples collected from the Stack Overflow website from 2016 to 2020, was utilized for the data analysis.
- Irrelevant features dropped to reduce bias
- Question body scraped and tokenized for training, .
- Bi-LSTM model trained for 100 epochs with batch size of 32 and achieved 83% validation accuracy.



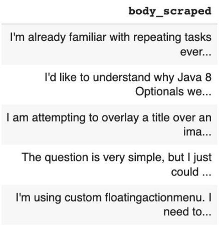
Embedding Layer (100000, 128)
Bi-LSTM(64, return_seq = True)
dropout
BatchNorm
Bi-LSTM(64, return_seq = False)
dropout
BatchNorm
Dense (64)
dropout
BatchNorm
Dense (3)

kaggle



Body	Y
<p>I'm already familiar with repeating tasks e...	LQ_CLOSE
<p>I'd like to understand why Java 8 Optionals...	HQ
<p>I am attempting to overlay a title over an ...	HQ
<p>The question is very simple, but I just cou...	HQ
<p>I'm using custom floatingactionmenu. I need...	HQ

BeautifulSoup

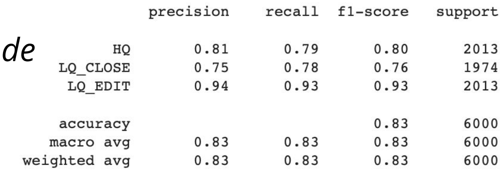


body_scraped
I'm already familiar with repeating tasks ever...
I'd like to understand why Java 8 Optionals we...
I am attempting to overlay a title over an ima...
The question is very simple, but I just could ...
I'm using custom floatingactionmenu. I need to...

tokenize

One-hot encode

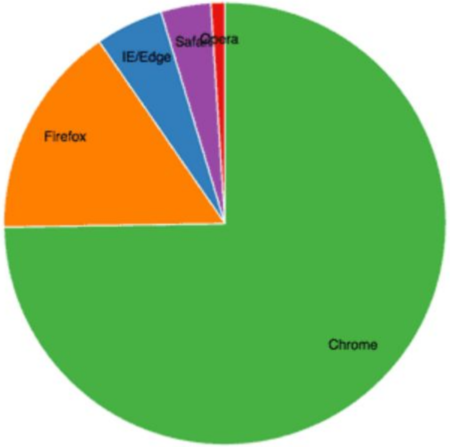
train

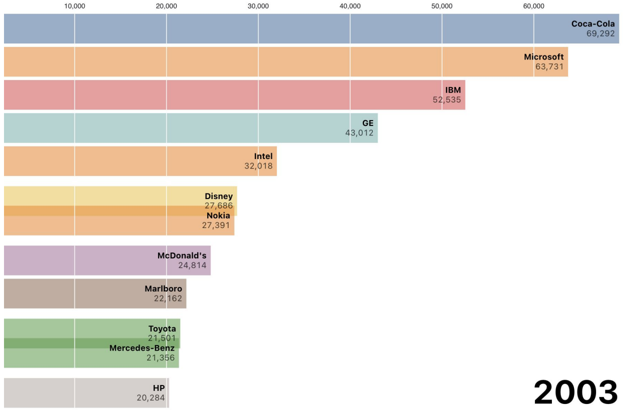


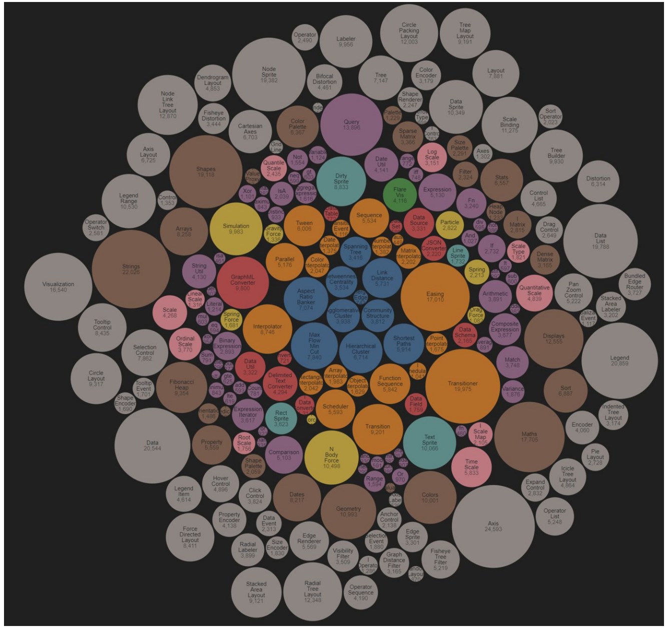
```
              precision    recall  f1-score   support

          HQ       0.81      0.79      0.80      2013
    LQ_CLOSE       0.75      0.78      0.76      1974
     LQ_EDIT       0.94      0.93      0.93      2013

    accuracy                           0.83      6000
   macro avg       0.83      0.83      0.83      6000
weighted avg       0.83      0.83      0.83      6000
```

Technologies - D3 Visualizations

- Bubble chart to visualize the hot topics
- A bar chart race showing trending question categories
- Pie-chart showing the quality of incoming questions based on an LSTM model







Questions ?