Project 3 Presentation

Team 2

Overview

Intro to Big Data
Project 3

Analyze Twitter's data using Apache Spark.

- Spark 2.1.0 has APIs to support different languages. For this project PySpark, the python API, is used.
- Three different analytical tasks are implemented. Common theme: **How do trends behave?**
- Built with a Graphical User Interface to provide a better user experience.

The Big Question

On twitter, trending is an important way for ideas to stand out from a sea of tweets. So we ask the question: How do trends behave?

- **Task 1**: Are users with lots of friends and followers important vectors of hashtag spread?
- **Task 2**: Are retweets important vectors of hashtag spread? How do they compare to regular tweets?
- **Task 3**: How do trends change over time? What about the top 3 trends?

Note: Task 1 & Task 2 conclusions are made from data collected over a 15 minute period, containing ~46,000 tweets. Task 3 is based on data provided via class. They are used to demonstrate the analysis, but should not be generalized for all tweets.

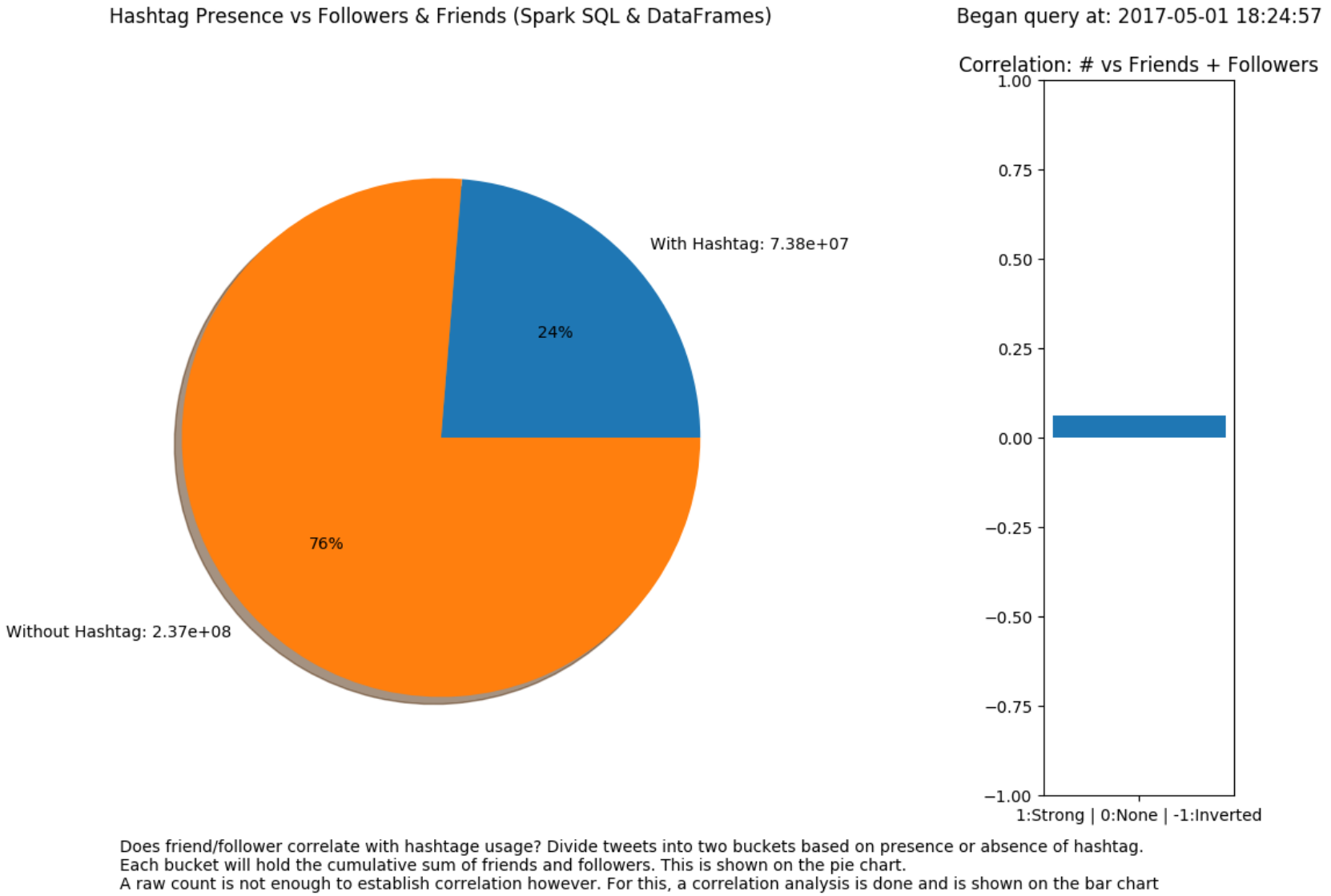
Task 1

space

Tweets without hashtags have three times as many friends and followers.

Despite this, the correlation between hashtag and follower + friends is weak.

Users with many followers and friends are not important vectors of hashtag spread in this set of data.



Task 1 Methodology

From a tweet, four attributes are retrieved:

- user id, hashtag, followers, friends
  - hashtag converted to zero or one, based on absence/presence

For analysis:

- Users that appear multiple times and sometimes do and sometimes don't use hashtags in their tweets are removed.
- Users that appear multiple times but consistently do/don't use hashtags are only counted once.
- Friends + Followers are summed cumulatively by absence/presence of hashtag and displayed in pie chart
- Correlation is done between two lists: the first signifying if tweet has hashtag, and the second holding the sum of friends and followers for that tweet

Task 2

Task 2: Methodology

From a tweet, three attributes are retrieved:

- hashtag, retweet hashtag, retweet's retweet count
  - hashtag and retweet hashtag converted to zero or one
  - RT will stand for retweet, for readability

For analysis:

- Although collected tweets have a retweet count of zero, if they were a RT, then there is a RT Count on the RT, or RT RT Count
- If a tweet is a RT that contains a hashtag, see how many times the original RT has been RTed. This tells us how many times the hashtag has been spread via retweeting.
- Aggregate that into the first category, RTs with #. Compare that with retweet volume that does not contain a hashtag.
- Finally compare it to new tweet volume (not a RT) with a hashtag.

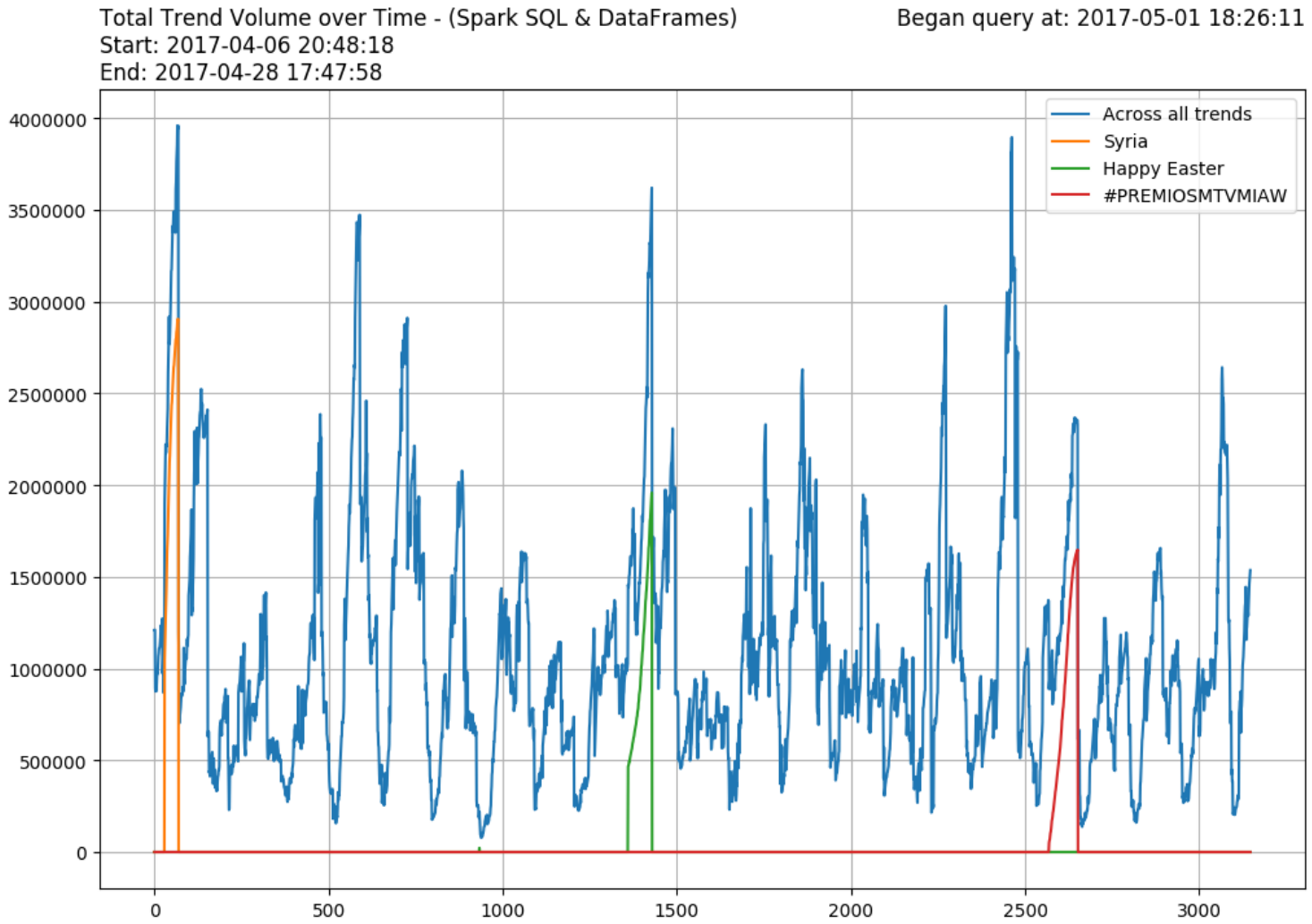
Task 3

space

Cumulative trend volume, across all trends, varies erratically.

The top three trends, although brief, account for significant portions of surges in trend volume.

Trends are brief, but can flood the Twitter quickly.



How do trends behave over time? The main plot shows cumulative volume across all trends over time.
In addition, the top three trends are also shown over time. The top three are identified by finding cumulative volume across all time by trend. The top 3 are then plotted, showing behavior.

Task 3: Methodology

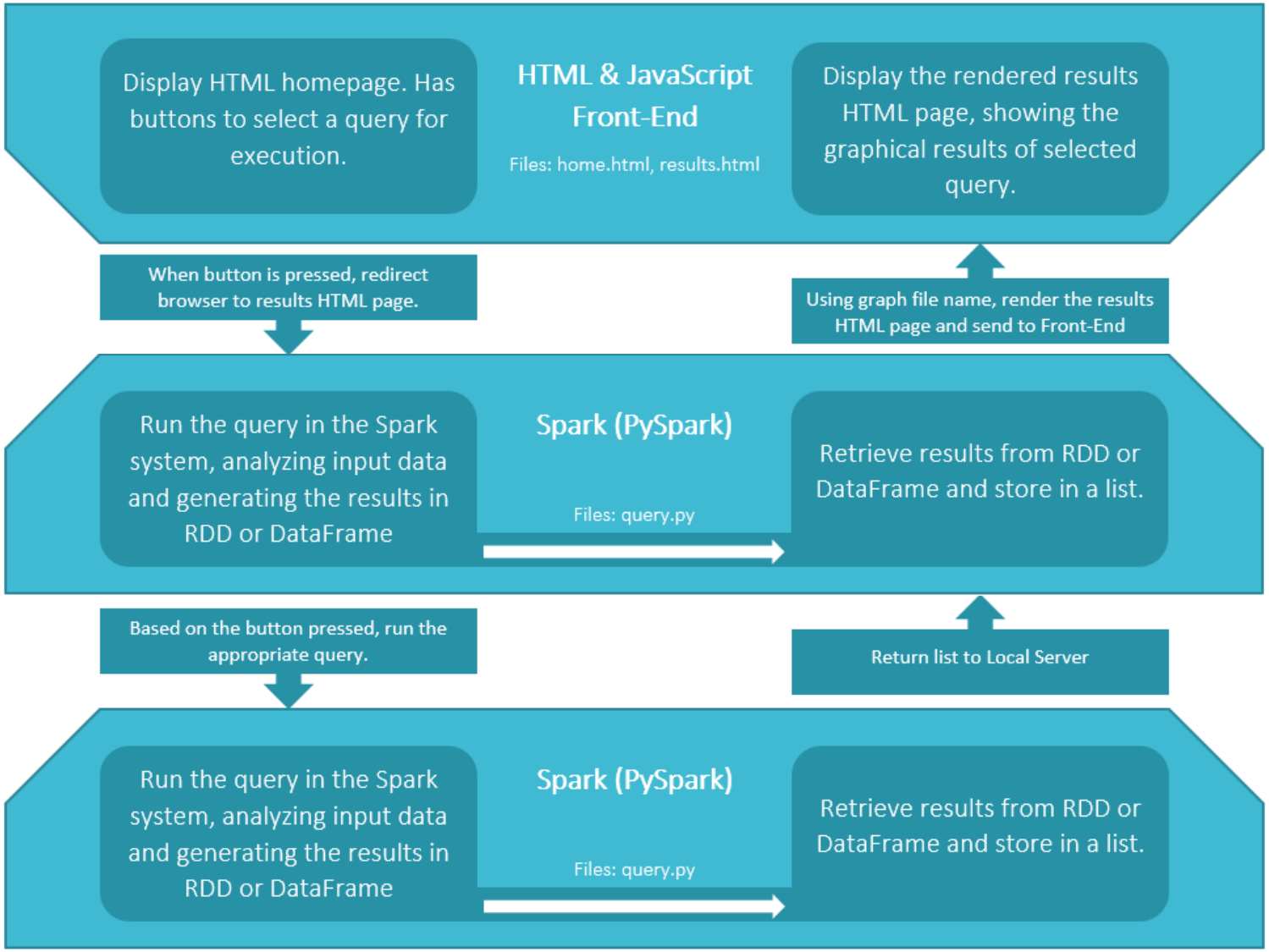
From a query, three attributes are retrieved:

- created_at, name, tweet_volume

For analysis:

- For the main plot, cumulative tweet volume, aggregate the tweet volume across all trends, by time.
- To identify the top 3 trends, aggregate the tweet volume across all time, by trend. The trends with the most tweet volume are identified as the top 3 trends.
- Retrieve tweet volume from the main plot that corresponds to the top 3 trends.
- Plot the top 3 trends over time. Where the top trend is not present, plot a zero.

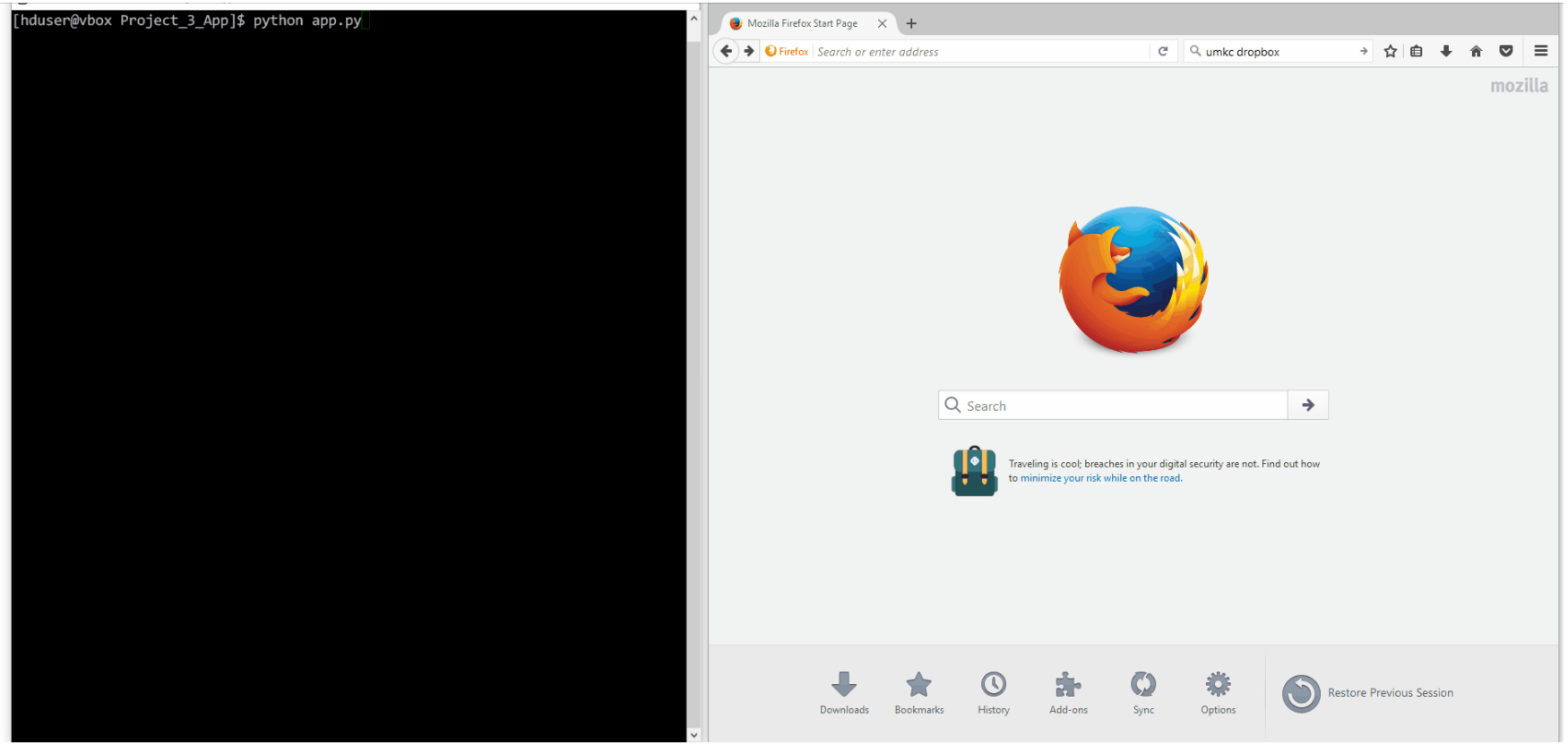
Extra Req.
Software Stack



Demo

Server console on left
GUI on right

Flask Server and Task 1



Demo

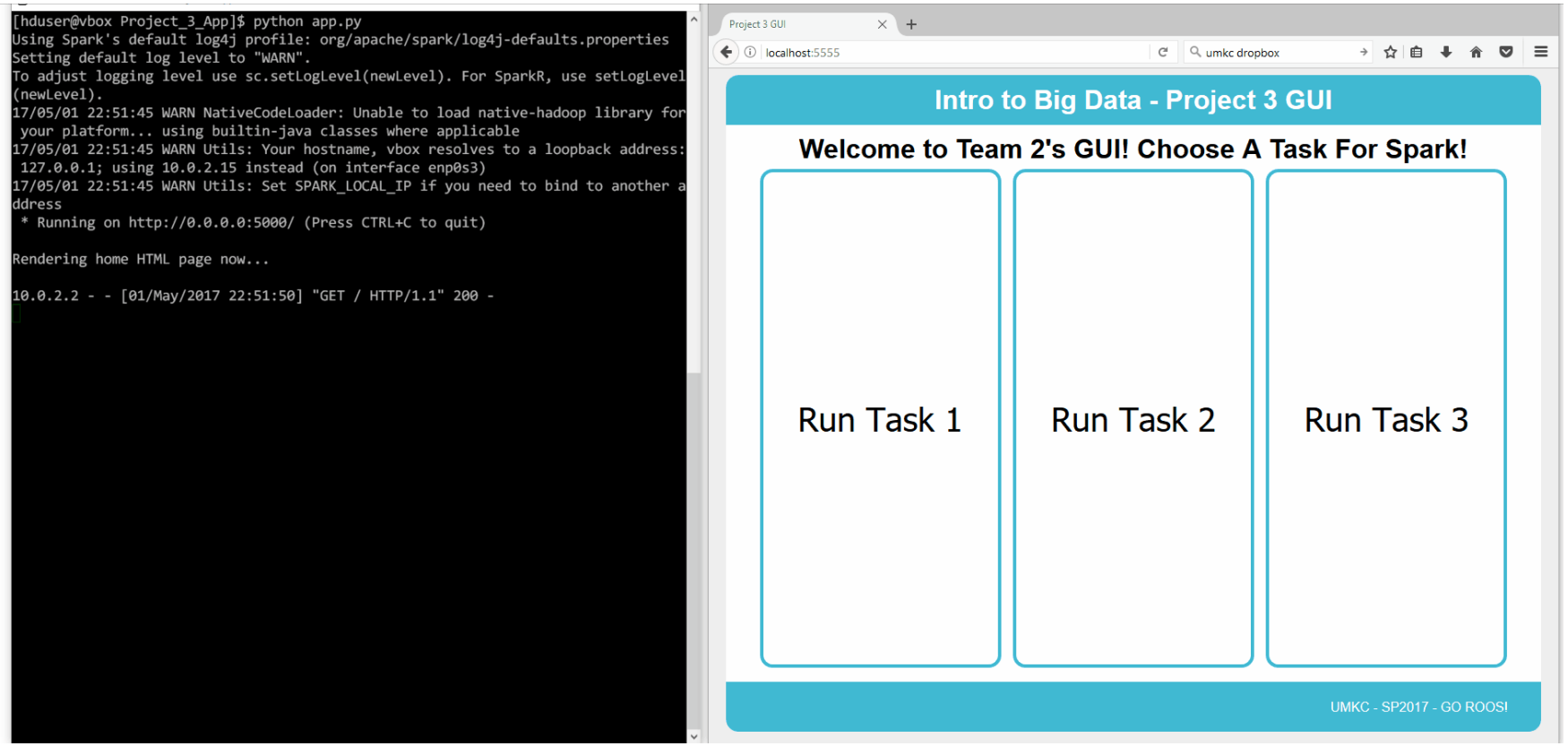
Server console on left
GUI on right

Task 2

Demo

Server console on left
GUI on right

Task 3



Questions?