Explorations in Data Science-Summer2020

Reflections paper

Team Members:

- ✓ Likhitha vanga
- ✓ Shashank Shekhar

Our project twitter sentiment analysis had to deal with a lot of data and analyzing them.

The process and hurdles during the project:

- we were trying to incorporate Pandas into the data pipeline to create reports out of it.
- There are difficulties in following areas:
    - Selecting specific columns from Data frame.
    - Grouping
- The closest possible thing we were looking for is converting the data frames to dictionaries.
- There were also issues installing pandas and making it compatible to python version.
- Installed anaconda and did the following things:
    - Created a virtual environment.
    - Installed TensorFlow and keras.
    - Installed pandas.
    - It was finally made compatible.
- We have used jupyter to run all our code.
- Our main challenge is to understand how nltk is able to analyze tweets from a person and categorize them into positive, negative and neutral. Throughout the process, we got a chance to really dive into the tweets and understand the specific words it uses to classify.
- Data is present indifferent forms around us. Choosing the data format took a really long time for us.
- After choosing the data format, the next challenge was to be able to read the data.
- We have cleaned the data removing the non-readable fields and empty columns which took a really long time because the data files are large.
- NLTK and Text Blob are two algorithms used to do the twitter sentiment analysis. We have tested the tweets with both algorithms.

- The main idea behind the project is being able to relate the polarity of the tweet to the stock value. But how? We were actually able to categorize the twitter data and also able to understand the stock prices data. Manually we are able to relate the twitter sentiment to rise or fall in stock prices.
- A lot of research went after the idea of correlating the both data. We have decided to calculate the percentage change in stock values and calculate the polarity on each day.
- So, we have the polarity of tweets and polarity of stock value and as both are a value, we were able to relate them.
- We have a lot of prediction algorithms in machine learning we learnt in Spring 2020.
- We have gone through few algorithms where and tested our data on Linear Regression, Random Forest and MLP.
- Finding which algorithm will be good for the specific theory we want to prove was hard.
- Each algorithm has its positives and drawbacks.
- It took a week to analyze each algorithm and prove which works better.

Major Takeaway from the Course:

- Going through Calling B.S videos and discussing them with the professor was the major takeaway.
- We all know that many of the algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning process.
- We understood that It's time to do something, and as educators, one constructive thing we know how to do is to teach people.
- Calling B.S taught how to think critically about the data and models that constitute evidence in the social and natural sciences.
- It is not always easy to get results.
- Trying again n again made us understand how all the ML models are built.
- The final presentations are very thoughtful and made us think heartfully.
- We also understood that there is a pattern under every data module.
- We just have to research and find how this pattern can be analyzed and understood.

What I would like to do differently next time:

- We would like to think about the project from the pain point.
- Beginning the project with a problem statement would be more sensible and useful.

- Listening to the peer presentations made us realize that there are many underlying problems we usually ignore, which shouldn't be.

References:

- https://medium.com/bhavaniravi/whats-wrong-with-python-pandas-32ba5bb2b658
- https://machinelearningmastery.com/machine-learning-in-python-step-by-step/
- https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed