Summer Internship Report

Sanket Shahane

Lucidworks, San Francisco, CA

Unity Id: svshahan

I am currently interning at **Lucidworks** for the period **15 May, 2017 to 15 August, 2017** as **Data Engineering Intern**. Lucidworks is an enterprise search company based on Apache Solr. Their product enables organizations to provide search capabilities and features on the huge amounts of data. They also provide out of the box machine learning capabilities like document clustering, product recommendations, and learning to rank search results and a few others. My internship had two major roles; Machine Learning, and Engineering.

My first task at Lucidworks was to learn and understand their product “Fusion”, “Searchhub” and “www.lucidworks.com” website and how things are engineered to work together. Simultaneously I had to learn **Scala** & **Apache Spark** which would be required for the Machine Learning tasks.

As a first deliverable, I had to improve the search experience on the lucidworks.com website itself. Starting by adding advanced search features like ***Auto Correction, and Typeahead.*** This required learning about Apache Solr and various functions available in it. One of the biggest challenge in this project was speed of execution. We needed things to be very fast at runtime. This is where the concepts learned at CSC 505 Design and Analysis of algorithms has helped me to understand the tradeoffs between memory requirements and speed and I was able to combine Natural Language Processing techniques along with engineering approaches to successfully implement those features in their website. I also had the opportunity to coordinate with the User Interface design team to bring out these features into the production website. Another good experience working with cross functional teams.

The major second part of my internship was focused a lot on the machine learning features that the product Fusion has to offer. Focusing on further improving the search experience, I worked on to bring in the “people also search for: _ ” **similar query recommendation.** Apache Spark ALS Recommendations were used. I had a chance to work on Scala and Apache Spark which is currently state of the art, scalable tool used by many organizations involved in Machine Learning and Artificial Intelligence. I used the Natural Language processing techniques to implement the query similarity feature.

The last project I worked on was **Apache email lists recommendations**. Lucidworks has indexed all of the Apache foundation emails lists and discussions. The goal of the project was to

be able to recommend emails to visiting users based on their search query and the current email being read. I had an opportunity to research about semantic similarity between documents to be able to recommend not just textually similar documents but relevant ones. I also got the opportunity to work from scratch; from considering the use cases to address, researching about NLP semantic similarities, prototyping in python, to implementing a scalable Spark job in Scala that would run on almost a Terabyte of data. During the course of this project I learnt about Approximate Nearest Neighbors using Locality Sensitive Hashing and how it can be useful to get similar documents using very less amount of memory and being computationally feasible at big data scales. During this project I had a chance to follow the Data Science principles, approaches, and techniques learnt at school and evaluate the results and make statistically sound inferences from the results of the experiments. Some of the challenges I tackled during this project were lack of training data in the right format to train the models that I wanted to. As an alternative we came up with solutions to collect the required data in a smart fashion using other machine learning techniques. Prototypes on 22431 emails required 4GB of memory and 2 mins on a 8 cores machine for the standard cosine similarity approach. Using Locality Sensitive Hashing (LSH) we were able to reduce it to The LSH in scala-spark required 520 MB memory and less than 1 min using 2 cores, I also had a chance to fine tuned the models, achieving a precision of 96%.

Coursework at NC State that helped me during the internship:

1. CSC 505 - Design and analysis of Algorithms:
   All the engineering and data pipeline management work I did at Lucidworks was keeping in the mind the time and space complexity and how to balance a tradeoff between Memory requirements and Runtime.
2. CSC 591 - Foundations of data science:
   Statistical inferences and evaluation metrics like *precision* learned during this course were super helpful in evaluating my machine learning models.
3. CSC 591 - Machine Learning for User Adaptive Systems:
   Recommendation engines, NLP, deep learning and Markov models were the techniques I learnt during this course and was able to apply them during my internship for the Apache mailing list recommendations project.
4. CSC 522 - Automated learning and Data Analysis:
   The course project experience I had during CSC 522 project and the guidance received from the Professor and TA's on how to approach any given problem in data science helped me to handle the projects at Lucidworks very efficiently.

To summarise, the internship experience at Lucidworks was satisfying and up to my expectations. It has given me the opportunity to work in the field of Machine Learning and NLP and has further nurtured my interest to work in this domain. The location of the office also gave me the opportunity to participate in many meetups and have conversations with people from different organizations which helped me grow my professional network.