CS – 425

Algorithms for Web-Scale Data

Project Proposal

Kerem Ayöz – 21501569

Muhammed Safa Aşkın – 21502860

Mert Epsileli – 21502933

Simon Arda Yuvarlak – 21501402

List of implementation project ideas:

1. **LSH and Convolution Techniques (Similar to #2 in project document)**

In this project we will try to make image analysis by using LSH and convolution techniques. We will extract the features of image with convolution and after clarifying the features of images we will apply LSH algorithm to find similar ones. The aim of this project is mainly finding similar images from huge number of images. Python programming language will be used for implementation.

Possible datasets are listed below:

- Rijsk Museum Image Dataset: http://rijksmuseum.github.io/
- ImageNet: http://imagenet.org/imagenet_data/urls/imagenet_fall11_urls.tgz

2. **Collaborative Filtering and LSH (Similar to #4 in project document)**

In this project our main goal is implementing a movie recommendation system by using different collaborative filtering algorithms (1-) user-user, 2-) item-item, 3-) latent-factor) and LSH. For example, in user-user algorithm we will try to make our recommendations by using other users' choices with similar movie preferences. Furthermore, at the end of the project we will try to validate the correctness of our collaborative filtering algorithms result by applying LSH on movies' features. Python programming language will be used for implementation.

- Netflix Movie Dataset: https://www.kaggle.com/netflix-inc/netflix-prize-data

3. **MapReduce (Similar to #3 in project document)**

In this project we will try to analyze and extract meaningful information from huge datasets by using MapReduce algorithm. For this project We're planning to use Amazon Web Services or Google Cloud Platform with Map-reduce support. We will implement one of the fundamental graph algorithms such as Dijkstra's Shortest Path, Breadth First Search etc. Additionally, we may implement the algorithm for finding the tf-idf score of a given word from given documents. For this task, we will collect academic articles about specific subject and try to find the tf-idf score of the words that are related with that subject. Python programming language will be used for implementation.

- Twitter Graph: http://an.kaist.ac.kr/traces/WWW2010.html