CS 267 Topics in Database Systems – Project Proposal

Project Name: Yelp Insights and Recommendation System
Team Members: Piyush Bajaj, Akhilesh Jichkar

Background:

Yelp is providing its dataset for students to conduct research and analysis as a part of its Yelp Dataset Challenge. The challenge encourages students to use Yelp's dataset in innovative ways and to conduct research in fields such as Image Recognition, Natural Language Processing, and Graph Mining. We are using this opportunity to build an Insights and Recommendation System using Big Data technologies such as Hadoop and Hive.

Objectives:

- Define and utilize relevant features from the dataset to provide users with meaningful recommendations.
- Perform Sentiment Analysis on use reviews and determine user influence.
- Create dashboards to summarize and predict business trends.
- Create a web interface to enable user to input their preferences.
- Learn various Big Data tools and technologies and their best practices.

Significance:

This application will provide the user with a web interface to enter their preferences and to find the most relevant results. Research suggests that user input is necessary in creating a useful recommendation system. Another research has found that user ratings and reviews play a crucial role in the success of a business. Sentiment Analysis would be performed on the dataset to help the businesses as well as in user recommendation. Another useful feature would be to rate the credibility of a user, since a lot of information nowadays is not reliable, and user influence is also an important factor. The application will be useful to both users and business owners using Yelp. Finally, users and businesses need a summarized analysis of their performance, hence dashboards will be created to help analyze business performance.

Scope:

The result of the project would be a recommendation system for users which will be transparent to both users and businesses. Initially data cleaning and feature set selection would be performed in parallel with user interface creation. Once completed, big data analysis would be performed using Hive and the dashboards would be created using Tableau. The next step would be sentiment analysis, followed by User influence determination and information verification.

Project Plan

As we are a team of two, we plan to distribute implementation of various domains, such as setup, preprocessing, data crunching & cleansing, SQL queries, web interface design, data visualization, sentiment analysis, and user influence calculation.

The following chart gives an approximate timeline for the completion of the project:

Phases Timeline	Description of Work	Start and End Dates
Phase 1	Preprocessing of Data-sets / Creation of User Interface	03/01/18 – 03/25/2018
Phase 2	Hive design process/Dashboard Creation	03/26/18 – 04/10/2018
Phase 3	Sentiment Analysis / NLP	04/11/18 – 05/01/2018
Phase 4	User Influence / Information Verification	05/02/2018-05/25/2018

Resources:

Programming Language: Java.

Technologies: Hadoop, HDFS, Hive, AWS EC2, Ubuntu.

Tools/IDE: Eclipse, Putty, Cloudera VM, Tableau.