CS 6240: Project Proposal

Team members:

1. Paras Chauhan
2. Phani Dharmavarapu

Data Set:

We are using Linkedin user profile data. (7072 MB)

Data Source:

http://www.reddit.com/r/dataisbeautiful/comments/25qjpz/how_many_employees_are_moving_between_companies_oc/chjvd0g

Data description:

The Linkedin dataset has records in JSON format with the common keys being “skills”, “positions”, “public-profile-url”, “location”, “first-name”, “num-connections”, “last-name”, “industry”. The “position” tag value has an internal JSON with tags “summary”, “title”, “start-date”, “end-date”/”is-current” and “company-name”. Each JSON record corresponds to a single user profile.

Below JSON record is an example record having all the different attributes associated with a user profile. However, some of these attributes might be missing in other records. This is due to the fact that some users might be “lazy” to completely fill in their details.

Data example:

Below is a list of all the fields associated with each record the dataset:

{skills":["Regulatory Affairs","Regulatory Requirements","FDA","Regulatory Intelligence","Pharmaceutical Industry","NDA","GMP","CAPA","Regulatory Filings","ANDA","Generic Programming","Sop"],
"positions":[{"summary":"Drug Listing for Pharmaceuticals, web publishing Company Core Data Sheets, electronic redlining package inserts, packaging labeling, edit, proof regulatory documents related to labeling, Annual Report coordinator, submission management, create document packages with changes to regulatory documents, specifications and bill of materials.",
"title":"Sr. Regulatory Affairs Associate","start-date":"1978-02-01","is-current":true,"company-name":"Baxter Healthcare"}],
"public-profile-url":"/pub/vos-l/33/b91/754","location":"Greater Chicago Area","first-name":"Vos","num-connections":"2","last-name":"L","industry":"Pharmaceuticals"}.

Development Work and Analysis:

The project tasks have been aligned as follows: -

Main tasks:

1. **Create a graph (Directed):** A directed graph will be created as follows for analysis
   1.1. Nodes: Companies
   1.2. Edges: Employees
2. **Decision Trees using ID3 algorithm with Random Forests:** We will build a decision tree to predict if an employee will leave or stay in the company in a specific year. (Binary classification).
3. **Analysis:** For all years/cumulative - Each of the following analysis will be done on a per year basis as well as per the entire year range of the dataset.
   A. **Number of connections vs Attrition rate:** Analysing the correlation between number of connections users have and their tendency of leaving/switching jobs.
   B. **"Hot" and "Cold" companies per sector:** Hot companies are those companies where more employees are joining. Cold companies are those which have higher attrition rates.
   C. **High and low attrition rates per sector:** Analysing which industry sectors are more bound to have higher employee attrition rates.
   D. **Dominant sector per region(country/state):** Analysing which industry sector is more prevalent in a specific region.
   E. **"Hottest" regions per year:** Analysing which regions attracted more employees.
   F. **Top 10 recruiting companies per school:** Analysing which are the top 10 recruiting companies for each school.
   G. **Top 10 schools per sector:** Analysing top 10 schools which have most number of students per sector.
   H. **On-demand sector/company analysis:** Per month analysis of a single sector/company across years.

Extension:

Further features like "Predicting the most hot/cold sector/company per month/year based on a new model", "A small GUI for running on-demand sector/company month/year analysis", etc can be implemented provided we have time.

Sub tasks:

1. **ETL:** Extracting meaningful data, building data model and loading.
2. **Clustering companies to sectors:** As part of the ETL, tagging each company a user has worked for, with one or more associated "sectors". As that information is not available in our raw data. For example, 'IBM' can be associated with two tags, viz: 'Software' and 'Hardware'.
3. Building feature vector for each employee that will be considered during the building of the decision tree.
4. **HBASE**: Saving the learned model into HBASE tables for quick access during prediction of test data.
5. Evaluating the strength of decision tree using ROC curve and AUC.
6. Analysis on speedup and scaleup.

7. Explore how different design patterns and Hadoop parameter settings affect the performance of a MapReduce program.

Anticipated Issues/ Concerns:

1. Tagging each company with a sector will be a challenging task.
2. Selecting the features and building feature vector for each employee. Selecting wrong features can ruin the efficiency of the entire model.