Assignment 1

Cloud Computing

Topics: Hadoop Distributed File System

Analysis of Pokec Social Network data

Instructions to be read Carefully

1. Marks of every logic will be awarded in case of adapting map reduce programming Paradigm.
2. It is a group Task. For this assignment, any group consist of 2-3 members.
3. Assignment should be done on Python or C++.
4. Submit the source code file and document file with the screenshots of input/output with the code snipping
5. Plagiarism is strictly prohibited; assignment will be marked zero if plagiarized.
6. Late Submission is simply not Allowed.
7. Marks will be uploaded after conducted Viva Exams.
8. Evaluated out of 200 Marks.
9. Viva Rubrics will be shared with you soon.

Pokec is the most popular on-line social network in Slovakia. The popularity of network has not changed even after the coming of Facebook. Pokec has been provided for more than 10 years and connects more than 1.6 million people. Datasets contains anonymized data of the whole network. Profile data contains gender, age, hobbies, interest, education etc. Profile data are in Slovak language. Friendships in Pokec are oriented.

(Dataset is available on: https://snap.stanford.edu/data/soc-Pokec.html)

Dataset statistics	
Nodes	1632803
Edges	30622564
Nodes in largest WCC	1632803 (1.000)
Edges in largest WCC	30622564 (1.000)
Nodes in largest SCC	1304537 (0.799)
Edges in largest SCC	29183655 (0.953)
Average clustering coefficient	0.1094
Number of triangles	32557458

Fraction of closed triangles	0.01611
Diameter (longest shortest path)	11
90-percentile effective diameter	5.2

The dataset contains:

File	Description
soc-pokec-relationships.txt.gz	User relationship data
soc-pokec-profiles.txt.gz	User profile data
soc-pokec-readme.txt	Description of files

About this assignment, only use User profile data.

Objective:

To Develop a Map-Reduce based classifier to predict the `completion_percentage` of user profiles based on their demographic, behavioral, and preference-based attributes. Also, to Develop a Map-Reduce based clustering of user data.

Develop a system for the following set of functions using Map reduce approach only.

1. Analyze the user features (`AGE`, `gender`, `region`) and generate a report on it.
2. Investigate correlations between `completion_percentage` and other features.
3. Visualize relationships (e.g., boxplots for `completion_percentage` vs. `favourite_color`, heatmaps for numerical correlations).
4. Group users into segments using clustering (e.g., K-means) on the basis of their age and analyze how completion rates vary across clusters.
5. Handle sparsity by addressing outliers in the datasets (provide justification for your approach).
6. Encode `gender`, `region`, `eye_color`, and other categorical variables (use one-hot encoding, target encoding, or embeddings).
7. Process multi-label columns like `hobbies` and `spoken_languages` (e.g., split into binary flags or use count encoding using Map reduce word frequency count).
8. Implement the function with and without map reduce (Derive `days_since_registration` from `registration` and `last_login`).
9. Normalize/standardize numerical features using map reduce based clustering method (e.g., `AGE`).

Model Building (25%) (HDFS based only)

Choose any two classifiers (Linear Regression, Random Forest, Gradient Boosting) based on your choice and mapped the complete logic into HDFS.

Split data into training and testing and validation sets. Drop non-predictive columns (e.g., `user_id`).

Bonus Challenges (Optional)

Deployment on Cloud: Bonus marks will be awarded to someone using Virtual Machines on any public cloud.

Deliverables:

1. Report: A PDF document explaining your approach, visualizations, and insights.
2. Code: A Jupyter Notebook or Python script with reproducible steps.
3. Presentation: A 5-minute summary of key findings (if submitting for a course/team).

Many from you have only couple of months to learn.

Good Luck!