Technical Competency Analyzer

Task

Using the provided stackoverflow dataset and github dataset, you are required to carry out the preprocessing,data transformation and clustering for each of those dataset using X-means clustering (I selected X-means as the exact no. of clusters is not predetermined, this can be done using other methods such as K-means as well if its more appropriate).

Once the clustering is done, metrics of each of those clusters such as the average net_user_upvotes, average reputation etc. needs to be provided for each cluster as well (This is to get an idea about what that cluster means).

Project Description (What the clusters generated by the task would be used for)

This project is part of a candidate shortlisting/filtering system for a HR application. This system would take a candidate's github username and stackoverflow id which would be the input to the system. Using these, it would use the github and stackoverflow API to retrieve specific data regarding the candidate on these platforms. Based on the specific attribute values, the system would determine which specific github cluster and stack overflow cluster that candidate belongs to and would display that information along with the specific meaning of that cluster.

Links

- Google Drive Link: https://drive.google.com/drive/folders/1uD4IHtuObM6pEo0xQLcFXR364_oV-JBc?usp=sharing

- StackOverflow Dataset
    - The dataset which i have extracted and partially cleaned can be found on the drive link: (see stackoverflow.csv https://drive.google.com/drive/folders/1wKVLOPP-EZuDhdagifWaoMaxa7RIfrbv?usp=sharing)
    - The dataset was extracted from the Stackoverflow dataset hosted on BigQuery (kaggle link: https://www.kaggle.com/datasets/stackoverflow/stackoverflow) where I have extracted 1000 records to generate a small dataset with my

required attributes (https://www.kaggle.com/code/aselgunaratne/stackoverflow-data-cleaning). In case you believe that this could be done better please let me know.

- Attributes not used for the clustering
  - id (not for clustering but for reference purpose)
  - reputation (not used for clustering but can be used to do external validation of the clusters)
- Attributes that could be used for the clustering
  - upvotes, downvotes, and net_upvotes (used for clustering, can either remove net upvotes or both upvotes & downvotes instead based on which is more relevant)
  - question_count
  - average_question_score
  - average_question_view_count
  - answer_count
  - accepted_answer_count
  - average_answer_score
  - no. of badges, gold_badge_count,silver_badge_count,bronze_badge_count, badge_score (badge score was calculated by ((gold_badge_count*10)+(silvder_badge_count*4) + bronze_badge_count). Have to determine which will be used for the clustering)

- Github Dataset
  - A github dataset that I extracted and partially cleaned can be found in (github.csv in https://drive.google.com/drive/folders/1wKVLOPP-EZuDhdagifWaoMaxa7RIfrbv?usp=sharing).
  - This dataset was obtained from https://www.kaggle.com/datasets/johntukey/github-dataset
  - Attributes not used for clustering
    - id
    - name
    - login
  - Attributes used for clustering
    - public_repos
    - followers
    - public_gists
    - commits

    From the repo_list attribute
    - list of languages used in the repos list
    - count of list of languages used in the repos list

(not used for this clustering, rather used for a later clustering work)

- count of java repositories in the repo list
- count of javascript repos in the repo list
- count of C# repos in the repo list
- count of python repos in the repo list
- count of ASP .net repos in the repo list

From the commit_list attribute (referencing to the repo_list attribute)
(not used for this clustering, rather used for a later clustering work)

- count of commits made to java repositories
- count of commits made to javascript repositories
- count of commits made to C# repositories
- count of commits made to python repositories
- count of commits made to ASP .net repositories

Deliverables

- X-means/ K-means clustered models of the stackoverflow data and the github data.
- Source code of the data preprocessing, data transformation, feature engineering and the clustering
- Analysis of the clusters and cluster validation (internal, external and relative validation)
- Visualization of the clusters
- Preprocessed and transformed dataset prior to clustering

Time span

I am relatively flexible with the timespan as long as the project can be completed properly. I was hoping this can be done by 24th April..