ASSIGNMENT 2 OVERVIEW

Feature Engineering & Exploratory Data Analysis

By Sri Krishnamurthy

In the first part of this exercise, you scraped a few reports related to Fintech and collated key terms that are associated with Fintech. You also scraped jobs from two of the assigned largest banks in the US and evaluated hiring trends in these bacnks. The goal of this exercise is to build a dataset with features that can be used to analyze hiring trends in Fintech in the 24 largest banks by market cap in the United States. Note that 12 teams have scraped data from two companies each. Your primary tasks include:

1. Aggregation of multiple datasets into a single dataset
2. Build a list of 100 keywords in Fintech and assign Fintech categories to each keyword
3. Tag if a job is a fintech related job or not
4. Assign the categories as features to each job posting if it is a fintech related job
5. Analyze the final dataset and document insights.

PART 1 DATA PREP AND PRE-PROCESSING

Gathering and forming a single dataset with the help of data every other team has gathered from the banks they were assigned with.

- Process the datasets into a single file. Add a column to indicate the company associated with the job
- Handle the duplication/missing values of columns(features) in the data-set.

PART 2 FORMING CLUSTERS(Categories) FOR DIFFERENT AREAS IN FINTECH

- You have list of the words is based on the three methods from Assignment 1:
  - Word count
  - TF/IDF
  - Text Rank
- S**ummarize a single list of 100 keywords** by manually reviewing the words/bigrams/trigrams
- Build a word cloud.
- Choose 8-10 clusters that these words could be bucketed into.
- You can choose any methodology to create clusters. (Manual/Any NLP technique). We are just reducing dimensions here.
- Each keyword can be associated with one/many clusters

PART 3 FEATURE ENGINEERING

- For each job, our goal is to:

1. Determine if the job is related to fintech or not.

2. Assign each job to a fintech category/categories

Note: How would you do this? Research methodologies that can be used to do these tasks(1,2). Document your chosen methodology. (You can use algorithms/rule-based methods). Review results and comment on how good your methodology is.

PART 4 ANALYZING THE DATA AND GAINING INSIGHTS

- Summarize your findings in a claat document to answer questions like:
    - How are the top 24 banks hiring**?**
    - How are the fintech related job hiring trends?
    - Which companies have the most fintech related jobs and which ones least?
    - Rank the jobs in fintech categories. Which categories have the most jobs and which ones least?
    - If you have a job seeker, which areas would you recommend the job seeker to focus on based on available jobs?
- Use the tools you used in Assignment 1 for charts/graphs

PART 5 BUILDING A PIPELINE AND AUTOMATING IT

- We want to automate Part 3. Build a pipeline that can automatically take data from the 24 companies and compiles a single dataset with added features.
- Every team is **expected to generate a pipeline carrying out all the above-mentioned tasks efficiently**. (According to their allocation with LUIGI, AIRFLOW or DASK)
- Dockerize the pipeline and test it.
- Publish your Docker image to Dockerhub

Deliverables:
Deadline: 27th Noon
Links to:

1. **Data**
    a. Final dataset of jobs
    b. Final list of keywords used
2. Exploratory data analysis report with key findings (claat document link)
3. Docker file, Docker hub link and instructions to reproduce the pipeline.