ASSIGNMENT 2 OVERVIEW

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.**
- Processing the data-set to make it in a certain format usable for the further parts.
- Handling the duplication of columns(features) in the data-set.

PART 2 FORMING CLUSTERS FOR DIFFERENT AREAS IN FINTECH

- The CSV of the words which every team created based on the three methods given:

  - Word count
  - TF/IDF
  - Text Rank

  Will be needed to be **summarized into one single list of words** by adding the words **gathered by all the other teams.**
- After forming one single list of all the words in it, next task would be to form clusters.

What is a cluster?
- Cluster will be similar to a bucket having its own list of words to compare with, when any given job description needs to be categorized into that cluster.

- Cluster will be formed based on the key areas in fintech. **And each cluster will have its own list of words associated with it**.
- **Every team is expected to form the list for a particular cluster based on the single word list formed**.

PART 3 FEATURE ENGINEERING

- After forming the clusters, feature engineering needs to be done.
- **Every team is expected to add new features in the data-**set which can support their classification and also analysis.

PART 4 ANALYZING THE DATA AND GAINING INSIGHTS

- Case study summary should focus on aspects like:

  - **Analysis based on each cluster formed.**
  - **Similarities and Dissimilarities between clusters and accordingly the jobs which fall under them.**
  - **Is a particular job opening related to Fintech or not?**
  - **Key hiring trends observed after the new compiled data-set. (Can be based on clusters also.)**

PART 5 BUILDING A PIPELINE AND AUTOMATING IT

- 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)
- Every team is also expected to **Dockerize the pipeline.**