Logo

Description automatically generated**Mid-Term Project**

Using the text data attached, your task is to perform text clustering using **PySpark** as well as some other processing. The list of tasks are as the following:

1. What pre-processing steps do you think is necessary? Explain what you have included and why (explain this in the form of comments on the Jupyter notebook)
2. Implement your text pre-processing
3. Split your data randomly into train and test
4. Tune the hyper-parameters of the k-means algorithm (the value K and the type of distance metric – i.e. Euclidean and Cosine). Use only the training dataset to do the tuning
5. Apply the k-means algorithm using the winning hyper-parameter values to the test data
6. Propose and implement an approach to evaluate the performance of your algorithm on the test data. If your approach is completely manual explain that in the comments of the Jupyter notebook.
7. For each cluster, report the top frequent words

Few notes regarding the submission:

* You will submit this as a group, however, each individual in the group needs to submit the notebook separately (I know, there will be redundancy)
* Please submit only the jupyter notebook (word documents, pdfs, zip files, screenshots and basically anything other than the jupyter notebook will simply be ignored)
* Please submit on-time, there will be no extensions.
* There will be no presentations, I will grade you based on the notebook