Team Information

Team Member 1: Rahul Varma
Person #: 50208353
UBIT Name: rvarma

Team Member 2: Sunil, Kunjappan Vasu
Person #: 50205673
UBIT Name: sunilkun

Environment Details

For this lab we used Pyspark set up with Jupyter Notebook (Anaconda-Python 3.6) on our local system.

Please note that using pyspark with Jupyter doesn't require explicit creation of the SparkContext (sc).
Therefore we've have commented out that line in our notebook submissions.

Submission Details

The submission contains two folders : one for the Vignette and the other for the Featured Activity

Vignette

The Vignette folder contains a Vignette.ipynb notebook and the titanic.csv data.

Featured Activity

The FeaturedActivity folder includes FeaturedActivity.ipynb, new_lemmatizer.csv, a sample_input folder, a sample_output folder and PerformanceEvaluation.pdf.

sample_input: Contains two folders. One for 2-grams and the other for 3-grams. Both folders contain only one Latin Document as the input file due to excessively large output files with an increase in input size. In case of 3-grams only a small part of the original Latin document is taken.

sample_input: Contains the output for the corresponding sample_input in two folders - 2-grams and 3-grams.

Both notebooks contain the full paths to the input and output files. Please change these paths as required.

PerformanceEvaluation.pdf: This file contains the results of the performance evaluation of our Spark program on our system for 2-grams and 3-grams.