LAB 5 : PROGRAMMING THE DATA FLOW FOR BIG DATA ANALYTICS USING APACHE SPARK

Team Members:
Anushree Gupta (agupta38)
Sumedha Nashte (sumedhas)

Activity 1 : Understand Apache Spark with Titanic data analysis

Environment: PySpark run on Jupyter

.ipynb noteook to run Apache Spark vignette.

Activity 2: Analysis of Latin documents for word-co-occurrence

Environment: PySpark run on Jupyter

Instructions on running the file:

The program is coded using Pyspark library and is run on Jupyter as a .ipynb notebook.

The input and output directory path can be specified in the code.

The code can be run as a normal Python notebook.

All libraries required for the program are imported at the beginning of the code.

Required output is generated and stored in the specified output folder. The output folder consists one .txt file for every input file.

The code also consists of importing the lemma file, whose path needs to specified.

Format of Output:

[((Word Pair/Trigram 1), ['<Location>']), ([Lemma Pair 1], ['<Location>']), ([Lemma Pair 2], ['<Location>']), (['Lemma Pair 3], ['<Location>']), ([Lemma Pair 4], ['<Location>'])]

[((Word Pair/Trigram 2), ['<Location>']), ([Lemma Pair 1], ['<Location>']), ([Lemma Pair 2], ['<Location>']), (['Lemma Pair 3], ['<Location>']), ([Lemma Pair 4], ['<Location>'])]

[((Word Pair/Trigram 3), ['<Location>']), ([Lemma Pair 1], ['<Location>']), ([Lemma Pair 2], ['<Location>']), (['Lemma Pair 3], ['<Location>']), ([Lemma Pair 4], ['<Location>'])]

Graph of No. of Files vs Time: