CS 6350

ASSIGNMENT \_\_\_\_\_3\_\_\_\_\_\_\_

Names of students in your group:

Aditya Kulkarni (axk230069@utdallas.edu)

Number of free late days used: \_\_\_\_\_\_\_\_0\_\_\_\_\_\_\_\_\_\_\_   
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

Solution:

Part 1: Report is from page 2.

Solution Link: <https://github.com/adityavkulkarni/6350_assignment3/tree/master/Part1>

Part 2:

Solution Link:

Google Colab link:

<https://colab.research.google.com/drive/1yP-7jeaYKonvGEgznyWzRck4rDjr39nS?usp=sharing>

Github link:

<https://github.com/adityavkulkarni/6350_assignment3/tree/master/Part2>

Dataset used:

<https://snap.stanford.edu/data/wiki-Vote.html>

Dataset is present in the GitHub repository: <https://github.com/adityavkulkarni/6350_assignment3/tree/master/Part2/input>

Assignment 3 Part 1

Spark Streaming with Real Time Data and Kafka

1. **Problem Statement**

**In this part, you will create a Spark Streaming application that will continuously read text data from a real time source, analyze the text for named entities, and send their counts to Apache Kafka. A pipeline using Elasticsearch and Kibana will read the data from Kafka and analyze it visually.**

1. **Execution steps**

**Detailed steps are present in** [README.md](https://github.com/adityavkulkarni/6350_assignment3/blob/master/Part1/README.md)

**A screenshot of a computer program

Description automatically generated**

**Kafka, Spark and ELK need to be downloaded and setup for running the above steps. Libraries required to run are mentioned in:** [requirements.txt](https://github.com/adityavkulkarni/6350_assignment3/blob/master/Part1/requirements.txt)

1. **Execution steps**