***CS 636/436 Project Proposal***

***Project title: Influence Analysis of YouTube Data***

*Submitted By:*

*Sowmya Kudva [G]*

*Evan Grill [UG]*

***1.1 Project Background and Related Work:***

YouTube, a Google company, has millions of unique visitors every day. These visitors upload or view millions of videos per minute. With that amount of activity comes an equally large amount of data about viewing patterns. Processing this data is of huge importance to YouTube. Done right, analysis of this data could improve visitor engagement, thus leading to an increase in ad revenue for creators and YouTube alike.

Just as Google’s search algorithms rank websites to improve search results, YouTube videos can be ranked to improve engagement. In this project we will be analyzing a sample YouTube dataset provided by researchers at Simon Fraser University and was obtained through scraping in the early days of YouTube, 2007 and 2008. This dataset includes quantitative data about each video as well as a list of related videos. Through analysis it should be possible to improve recommendations both at the home screen (highly ranked videos overall) and at the video player screen (highest ranked related video). Further work can be done to analyze other aspects of video meta-data such as popular categories.

Similar research is being performed on datasets as diverse as the Internet Movie Database, airline route data, or weather analysis. In all of these cases, improvements to distributed processing from MapReduce, Hadoop, Spark, etc. has advanced the field of data science tremendously.

***1.2 Project Objective and Approach:***

Our analysis should be able to determine the following:

* Highest ranked related video (Improve “next video” recommendations)
* Highest ranked videos overall (Improve “home page” recommendations)
* Top video categories
* Performance metrics for top videos.

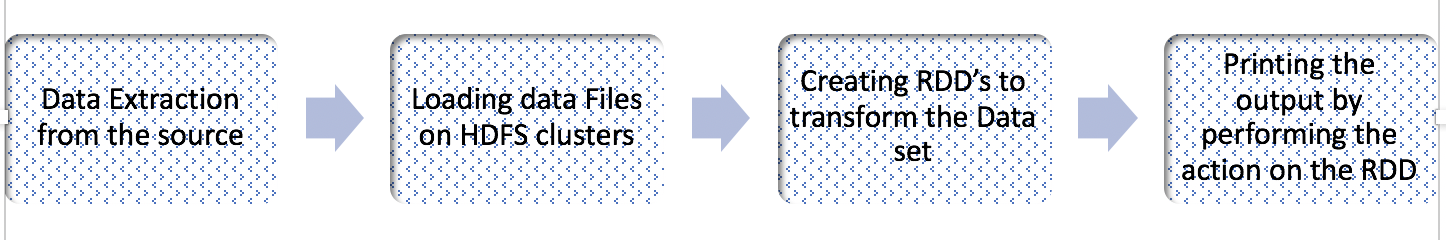


Figure 1: A flow chart outlining the data pipeline for analysis.

***1.3 Challenges and/or Motivation:***

At present, the team has installed PySpark on our personal machines. The current challenge is determining the MapReduce workflow to generate results. Future work will include moving the analysis over to a cloud-based service such as Amazon’s Elastic MapReduce (EMR). Past experience with Amazon Web Services by the team have indicated that their learning curve is high.

***1.4 Timeline , Milestones*** ***and Task ownership:***

The team has split up the timeline into weekly segments, with 6 major milestones. Figure below outlines this timeline.

A screenshot of a cell phone

Description automatically generated

Figure 2: A Gantt chart outlining the timeline for the project.

***1.5 Tools***

The team debated which MapReduce implementation would be the best choice. The team decided on Spark both due to its improved IO performance and because there exists a high-quality implementation in Python, PySpark. Using Python allows the team the opportunity to use data visualization tools such as MatPlotLib as well.

***1.6 Responsibilities:***

Since Evan has the most experience in Python, his primary job will be to do initial work developing the MapReduce workflows. Sowmya will be experimenting on the data distribution on HDFS clusters. Evan will work with Sowmya to develop the interface for the tools as well as create the visualizations for the final report and presentation.