Final Project Proposal

Youtube Videos Network Analysis

For Advanced Data Sciences at Northeastern University

Under the guidance of Prof Srikanth Krishnamurthy

Team 1 – Manasi Dalvi & Vishal Satam

Problem Statement

Youtube is a web based application that provides user's the facility to upload, view and rate videos. The major income source for youtube comes from the number of views that each video generates and they maximize their revenue by recommending related videos to user's to increase the likelihood of user clicking on these links which eventually increase the number of views. Our aim is to analyze the networking graphs and generate insights and analytics based on that. In the dataset that we have chosen, we have videos which we will consider as nodes. If there is a video a in b's related videos's list, then there is a directed edge from a->b. We will derive a weight for the edges based on the related videos and analyze the network structure of the videos to develop machine learning algorithms which can predict an estimate of the number of views a new video can generate. We will also analyze network characteristics to be presented to business users. We will also perform clustering to identify which clusters to which a video belongs to and identify the related videos for the new video.

Dataset

Data from Youtube has been scraped and is available at the following link.

http://netsg.cs.sfu.ca/youtubedata/

Tables and associated columns

There are 3 tables that are a part of this dataset.

Videos

video ID	an 11-digit string, which is unique
uploader	a string of the video uploader's username
age	an integer number of days between the date when the video was uploaded and Feb.15, 2007 (YouTube's establishment)
category	a string of the video category chosen by the uploader
length	an integer number of the video length
views	an integer number of the views
rate	a float number of the video rate
ratings	an integer number of the ratings
comments	an integer number of the comments
related IDs	up to 20 strings of the related video IDs

Video File size and bitrate

video IDs	Id of the video – 11 digit string
Video length	Length of the video in minutes, numeric
Video file size	Size of the video, numeric
Bitrate	Bitrate of the video, numeric

Users

Userid	String
Number of uploaded videos	Number of videos uploaded by the user, number
Number of videos watched	Number of videos that have been watched by the user, number
Number of friends	Number of friends for this user, number

Stakeholders

Regular Users

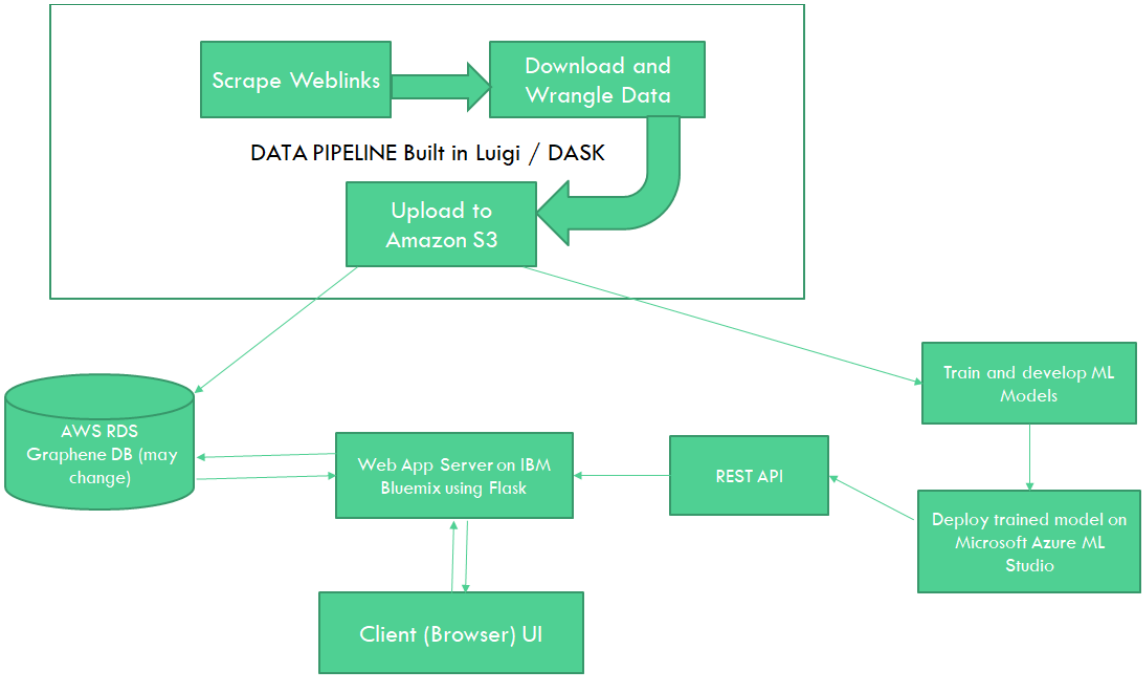
The main stakeholder of this project is the user himself. On logging into the web application, he will be able to see certain analytics about his profile such as the number of videos that he has uploaded, number of subscribers he has, his most popular videos according to views, total size and length of videos that he has uploaded.

Global analytics such as trending videos and top uploaders list. This will be calculated based on the in degree of videos and the number of views that a video has received and the preferences he has selected at the time of sign up.

Business Users

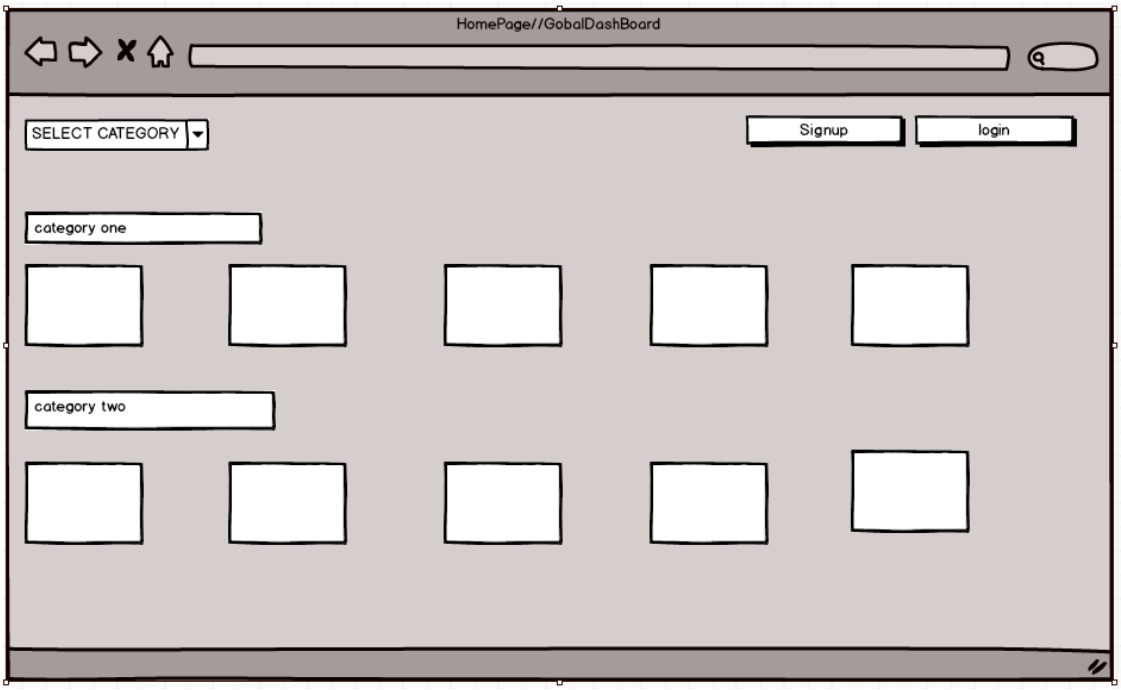
Get top users, videos and a graph of the network of videos to find out most popular videos to validate if the most related links have a certain degree of nodes.

System Architecture Diagram

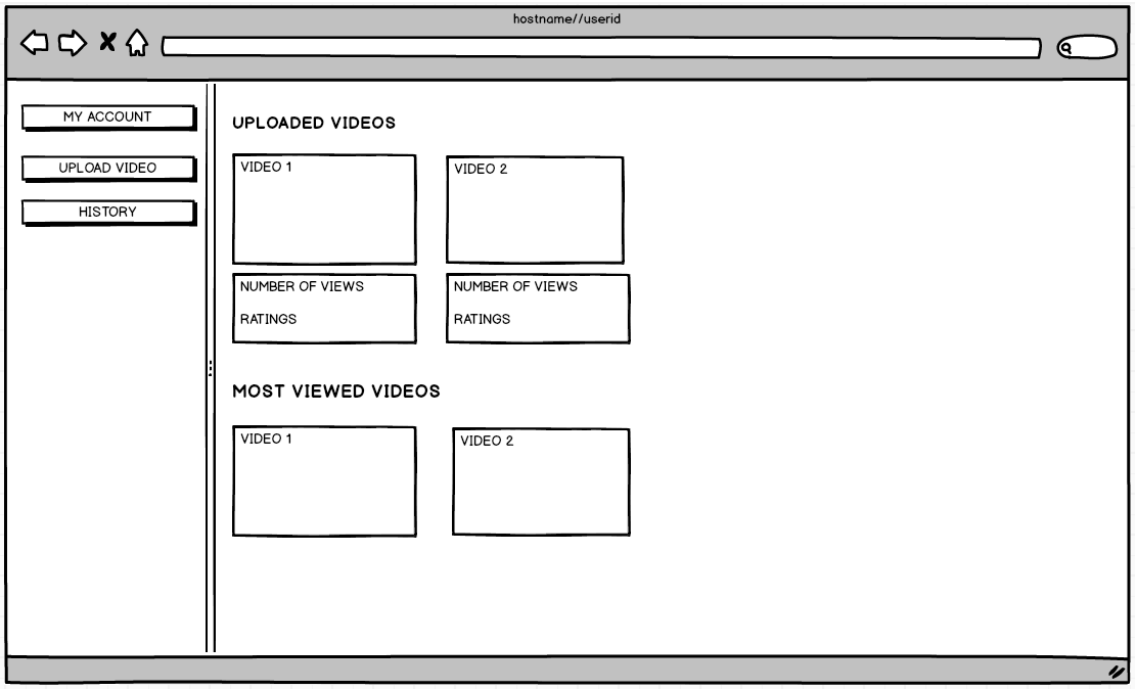


Mock UI

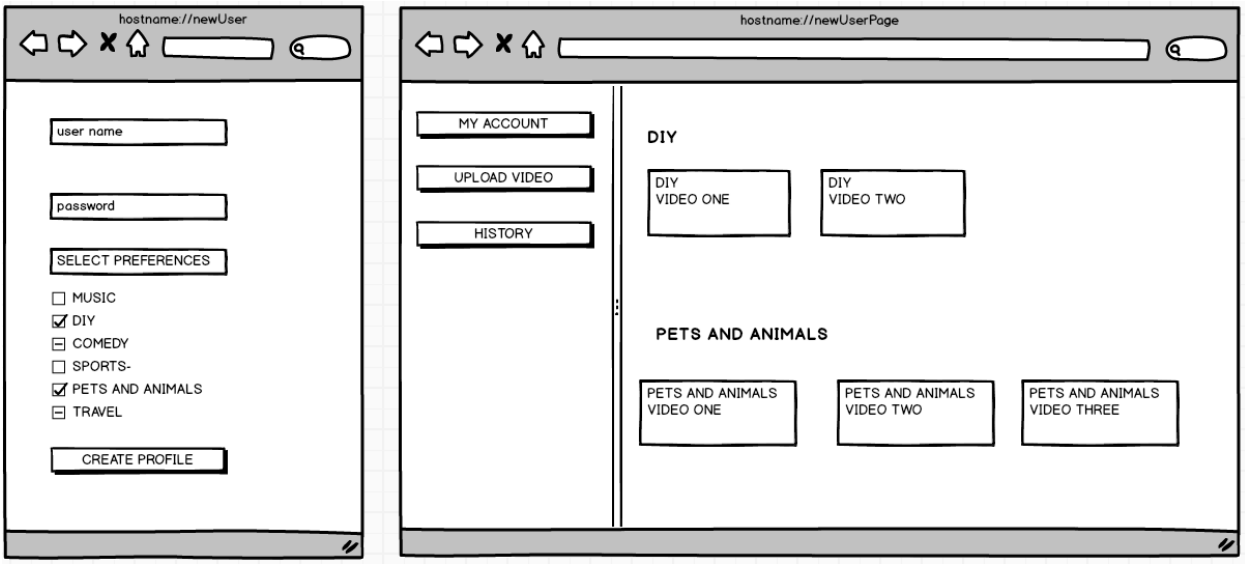
HomePage



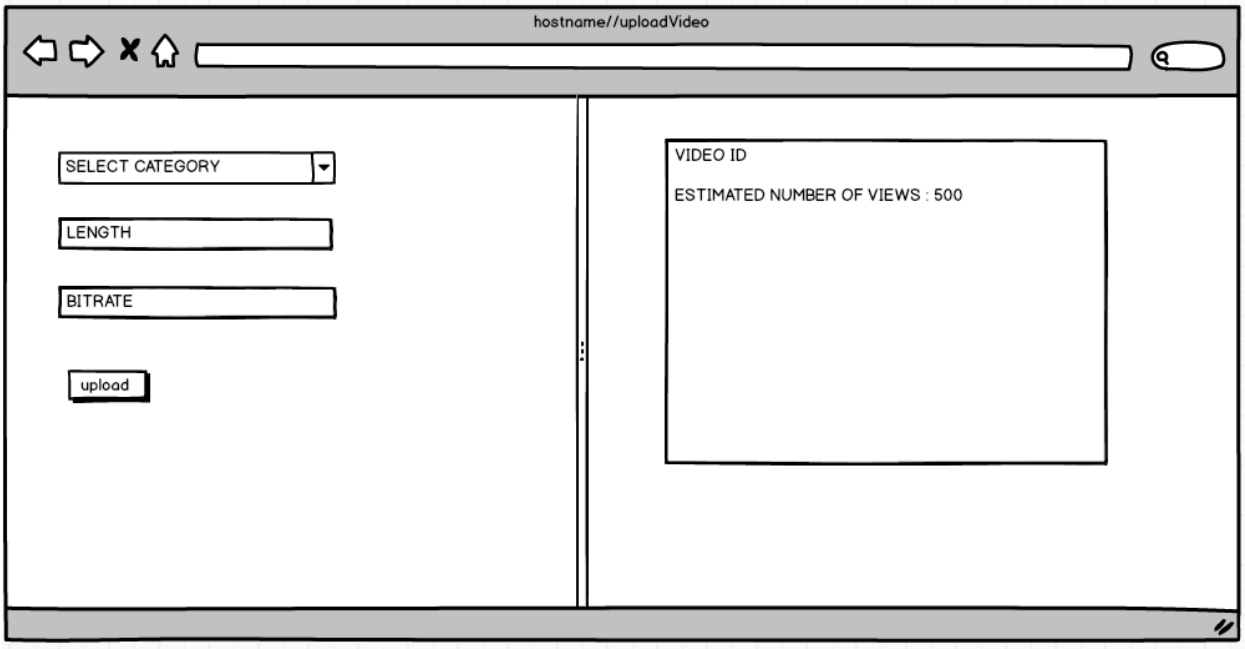
Existing user profile



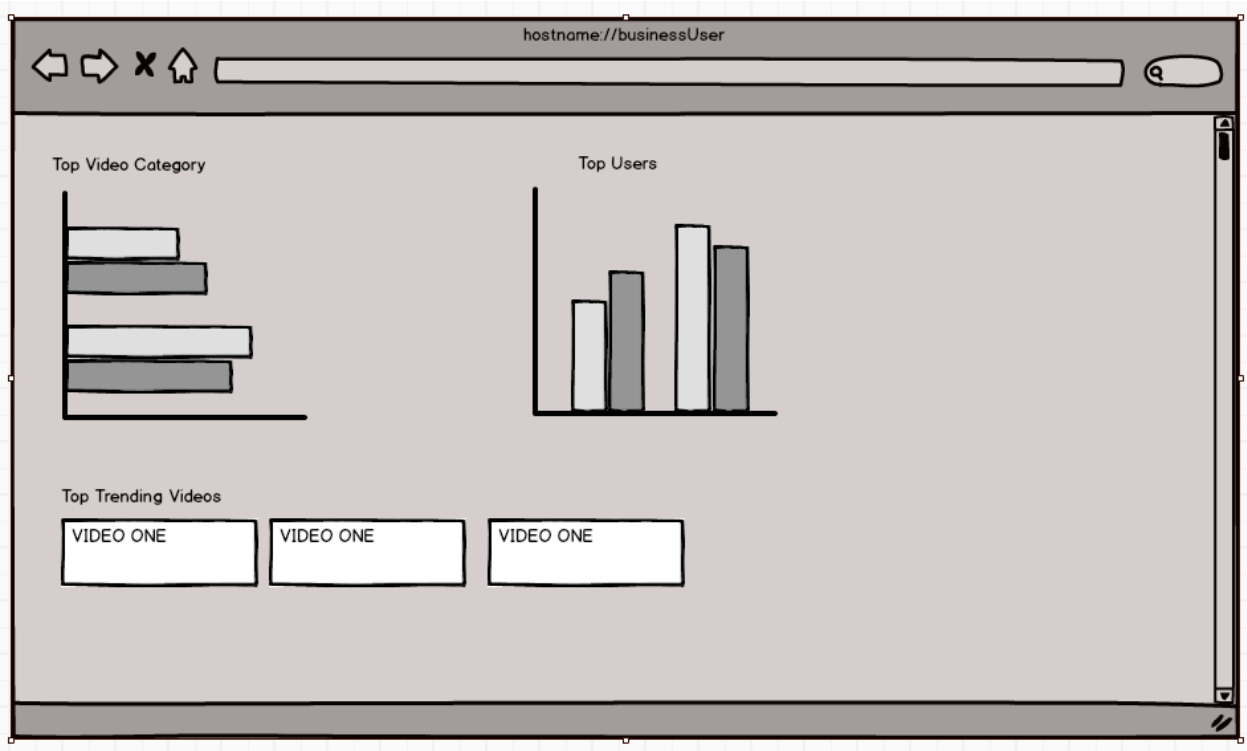
New User Login and Profile



Estimated Views Prediction for New Video



Business User Dashboard



Deployment details

1. Docker will be used to deploy code for pipelining and wrangling
2. Luigi will be used for pipelining
3. AWS will be used to store data files.
4. AWS will be used to train and test data to build Machine Learning models.
5. Data will be hosted on AWS to be provided as a service
6. Flask will be used to build the Web Application
7. IBM Bluemix to be used for deploying web application.
8. Microsoft Azure Machine Learning Studio will be used to deploy ML models.