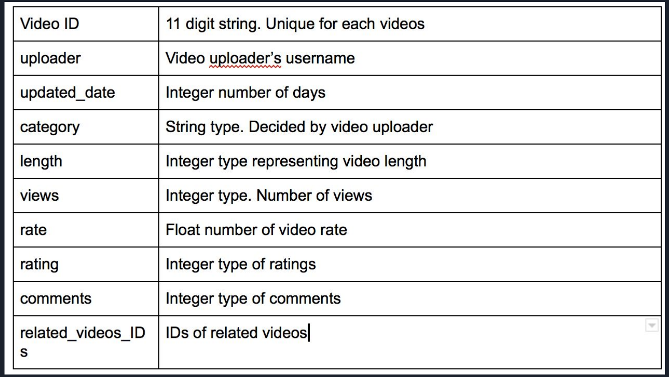
YouTube Analyzer

Insun Lee
Kim Nguyen
Chao Zheng

Background and Motivation

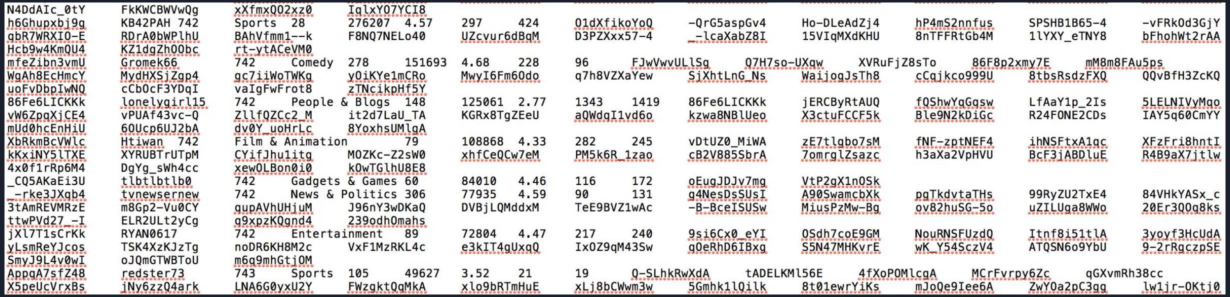
- Extract meaningful information from a large set of dataset
- Project could be applicable on other big data problem
    - Ex. Amazon's shopping list
        - Top k queries: Find the top k categories/items in which sold to customers; top k most popular item
        - Range queries : find all items in cateogries X with price within range [t1,t2];
        - User identification in recommendation patterns
        - Using PageRank algorithms. With high PageRank score, that means that item is related to other items in graph, thus has a high influence.

Problem Formulation



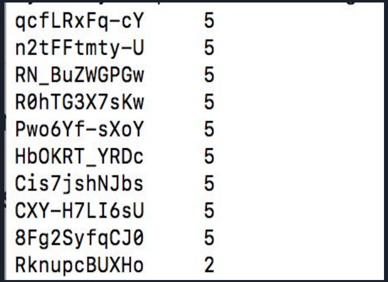
Video ID	11 digit string. Unique for each videos
uploader	Video uploader's username
updated_date	Integer number of days
category	String type. Decided by video uploader
length	Integer type representing video length
views	Integer type. Number of views
rate	Float number of video rate
rating	Integer type of ratings
comments	Integer type of comments
related_videos_IDs	IDs of related videos

- Input & Output
    - Input : .txt format



```
N4DdAIc_0tY  FkKWCBWVwQg  xXfmxQO2xz0  IqlxYO7YCI8
h6Ghupxbj9g  KB42PAH 742  Sports  28   276207  4.57  297     424   O1dXfikoYoQ  -QrG5aspGv4  Ho-DLeAdZj4  hP4mS2nnfus  SPSHB1B65-4  -vFRkOd3GjY
qbR7WRXIO-E  RDrA0bWPlhU  BAhVfmm1--k  F8NQ7NELo40  UZcvur6dBqM  D3PZXxx57-4  _-lcaXabZ8I  15VIqMXdKHU  8nTFFRtGb4M  1lYXY_eTNY8  bFhohWt2rAA
Hcb9w4KmQU4  KZ1dgZhOObc  rt-ytACeVM0
mfeZibn3vmU  Gromek66     742    Comedy  278    151693  4.68  228   96     FJwVwvULlSg  Q7H7so-UXqw  XVRuFjZ8sTo  86F8p2xmy7E  mM8m8FAu5ps
WgAh8EcHmcY  MydHXSjZgp4  gc7iiWoTWKg  yOiKYe1mCRo  MwyI6Fm6Odo  q7h8VZXaYew  SjXhtLnG_Ns  WaijogJsTh8  cCqjkco999U  8tbsRsdzFXQ  QQvBfH3ZcKQ
uoFvDbpIwNQ  cCbOcF3YDqI  vaIgFwFrot8  zTNcikpHf5Y
86Fe6LICKKk  lonelygirl15 742    People & Blogs  148  125061  2.77  1343    1419  86Fe6LICKKk  jERCByRtAUQ  fQShwYqGqsw  LfAaY1p_2Is  5LELNIVyMqo
vW6ZpqXjCE4  vPUAf43vc-Q  ZllfQZCc2_M  it2d7LaU_TA  KGRx8TgZEeU  aQWdqI1vd6o  kzwa8NBlUeo  X3ctuFCCF5k  Ble9N2kDiGc  R24FONE2CDs  IAY5q60CmYY
mUd0hcEnHiU  6OUcp6UJ2bA  dv0Y_uoHrLc  8YoxhsUMlgA
XbRkmBcVWlc  Htiwan  742  Film & Animation       79   108868  4.33  282     245   vDtUZ0_MiWA  zE7tlgbo7sM  fNF-zptNEF4  ihNSFtxA1qc  XFzFri8hntI
kKxiNY5lTXE  XYRUBTrUTpM  CYifJhu1itg  MOZKc-Z2sW0  xhfCeQCw7eM  PM5k6R_1zao  cB2V885SbrA  7omrglZsazc  h3aXa2VpHVU  BcF3jABDluE  R4B9aX7jtlw
4x0f1rRp6M4  DgYg_sWh4cc  xewOLBon0i0  kOwTGlhU8E8
_CQ5AKaEi3U  tlbtlbtlb0   742    Gadgets & Games 60   84010   4.46  116     172   oEugJDJv7mg  VtP2gX1nOSk
_-rke3JXqb4  tvnewsernew  742    News & Politics 306  77935   4.59  90      131   g4NesDsSUsI  A90SwamcbXk  pgTkdvtaTHs  99RyZU2TxE4  84VHkYASx_c
3tAmREVMRzE  m8Gp2-Vu0CY  qupAVhUHjuM  J96nY3wDKaQ  DVBjLQMddxM  TeE9BVZ1wAc  -B-BceISUSw  MiusPzMw-Bg  ov82huSG-5o  uZILUga8WWo  20Er3QOq8ks
ttwPVd27_-I  ELR2ULt2yCg  q9xpzKQgnd4  239odhOmahs
jXl7T1sCrKk  RYAN0617     742    Entertainment   89   72804   4.47  217     240   9si6Cx0_eYI  OSdh7coE9GM  NouRNSFUzdQ  Itnf8i51tlA  3yoyf3HcUdA
yLsmReYJcos  TSK4XzKJzTg  noDR6KH8M2c  VxF1MzRKL4c  e3kIT4gUxqQ  IxOZ9qM43Sw  gOeRhD6IBxg  S5N47MHKvrE  wK_Y54SczV4  ATQSN6o9YbU  9-2rRgczpSE
SmyJ9L4v0wI  oJQmGTWBToU  m6q9mhGtjOM
AppqA7sfZ48  redster73    743    Sports  105    49627   3.52  21    19     Q-SLhkRwXdA  tADELKMl56E  4fXoPOMlcgA  MCrFvrpy6Zc  qGXvmRh38cc
X5peUcVrxBs  jNy6zzQ4ark  LNA6G0yxU2Y  FWzgktQqMkA  xlo9bRTmHuE  xLj8bCWwm3w  5Gmhk1lQilk  8t01ewrYiKs  mJoQe9Iee6A  ZwYOa2pC3gg  lw1jr-OKtj0
```

- Output : .txt format



```
qcfLRxFq-cY    5
n2tFFtmty-U    5
RN_BuZWGPGw    5
R0hTG3X7sKw    5
Pwo6Yf-sXoY    5
HbOKRT_YRDc    5
Cis7jshNJbs    5
CXY-H7LI6sU    5
8Fg2SyfqCJ0    5
RknupcBUXHo    2
```

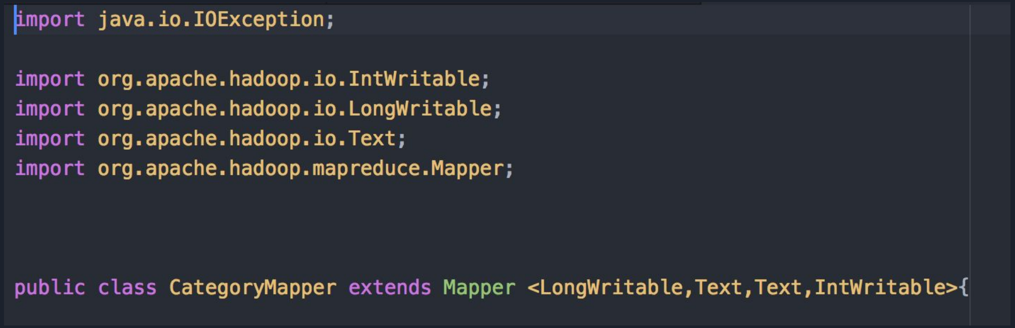
Top 10 Rated Videos

Algorithm

MapReduce (Hadoop)

- MapReduce is the heart of Hadoop
- Hadoop is a highly scalable storage platform designed to process very large data sets across hundreds to thousands of computing nodes that operate in parallel. It provides a cost-effective storage solution for large data volumes with no format requirements.
- Map - converting data set to another set, broken down into tuples (key/values pairs)
- Reduce - Taking the output of Map as input and combines data tuples into smaller set of tuples.

Algorithm

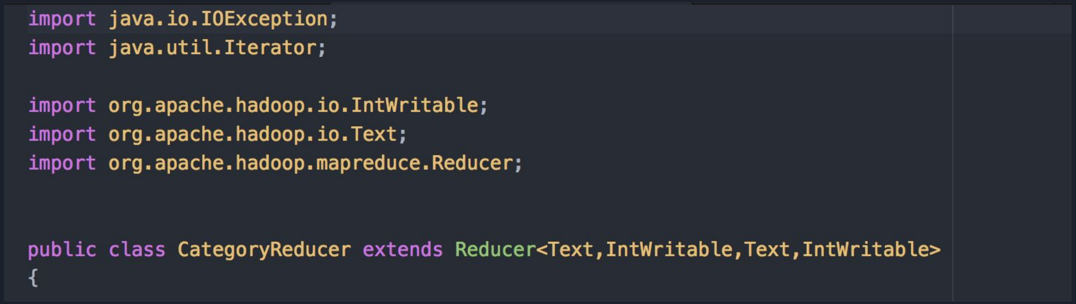


```java
import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;



public class CategoryMapper extends Mapper <LongWritable,Text,Text,IntWritable>{
```



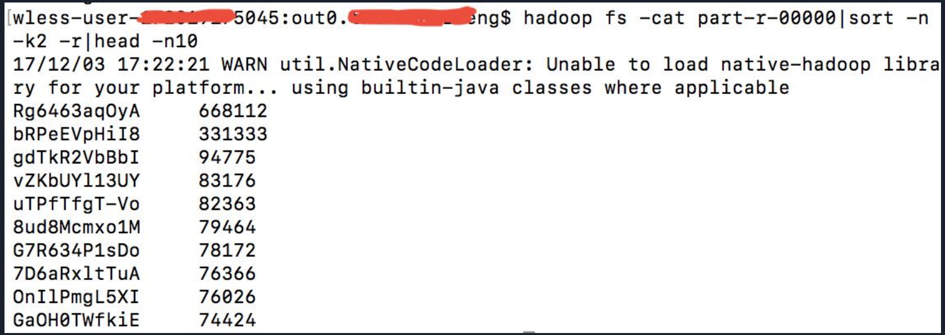
```java
import java.io.IOException;
import java.util.Iterator;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;


public class CategoryReducer extends Reducer<Text,IntWritable,Text,IntWritable>
{
```

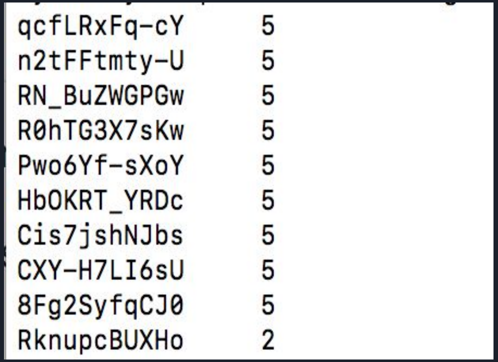
Experimental Study output

Top 10 Viewed Videos:



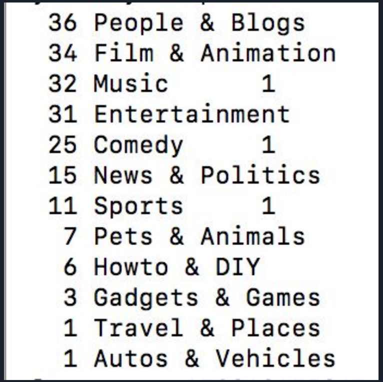
```
wless-user-[redacted]5045:out0.[redacted]eng$ hadoop fs -cat part-r-00000|sort -n
-k2 -r|head -n10
17/12/03 17:22:21 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Rg6463aqOyA     668112
bRPeEVpHiI8     331333
gdTkR2VbBbI     94775
vZKbUYl13UY     83176
uTPfTfgT-Vo     82363
8ud8Mcmxo1M     79464
G7R634P1sDo     78172
7D6aRxltTuA     76366
OnIlPmgL5XI     76026
GaOH0TWfkiE     74424
```

Top 10 Rated Videos:



```
qcfLRxFq-cY     5
n2tFFtmty-U     5
RN_BuZWGPGw     5
R0hTG3X7sKw     5
Pwo6Yf-sXoY     5
HbOKRT_YRDc     5
Cis7jshNJbs     5
CXY-H7LI6sU     5
8Fg2SyfqCJ0     5
RknupcBUXHo     2
```

Top 10 Categories:

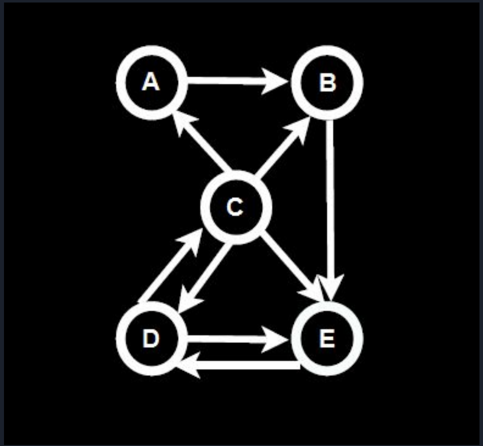


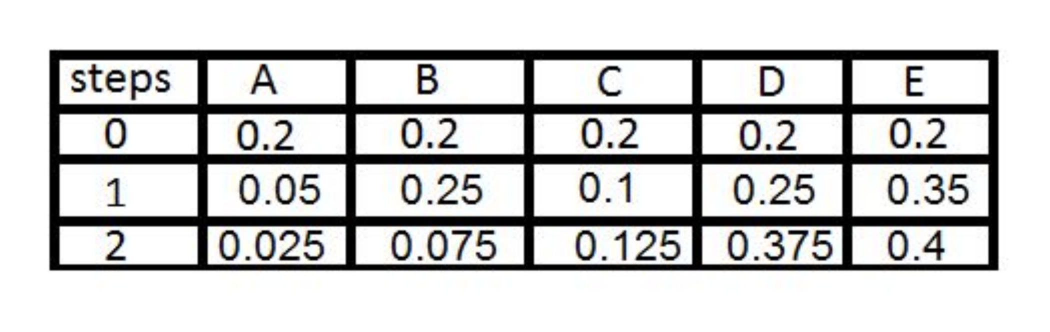
```
36 People & Blogs
34 Film & Animation
32 Music       1
31 Entertainment
25 Comedy      1
15 News & Politics
11 Sports      1
 7 Pets & Animals
 6 Howto & DIY
 3 Gadgets & Games
 1 Travel & Places
 1 Autos & Vehicles
```

PAGERANK ALGORITHM

PageRank

- rank websites in their search engine results
- Measure the importance of website pages





steps	A	B	C	D	E
0	0.2	0.2	0.2	0.2	0.2
1	0.05	0.25	0.1	0.25	0.35
2	0.025	0.075	0.125	0.375	0.4

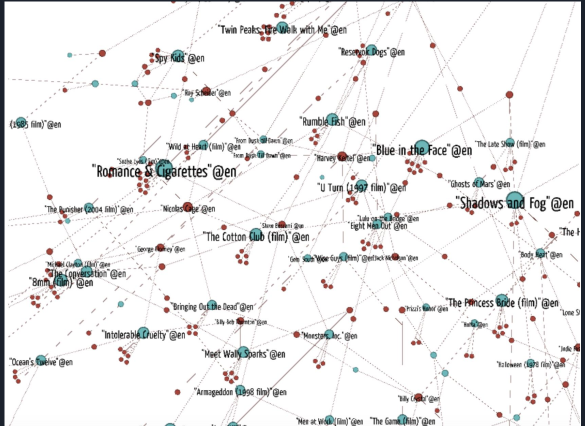
PAGERANK ALGORITHM

Top	Video ID	PageRank Score
1	rkvEuAtErwQ	0.15230242746613892
2	skn2nNHH8co	0.15168760879025242
3	wvPOOXdlO8U	0.151531893028099
4	5c9OBqgmjzE	0.15147045107345972
5	CCz1kmfqL7g	0.1514101936466493
6	U4bk0UH1poQ	0.15136395778938208
7	uAIIMI0BUQ8	0.15134156041956306
8	OIkUJnXLU74	0.15130953057948002
9	nMl6B9m1rDw	0.15122525021758051
10	RG2kMfInkFA	0.15121963744246494

Sample Recommendation

- Generate Graph based on
    - Video ID, Category, and rate

- Steps
    - Initially select the category
    - Clean the dataset based on selection
    - Select top 10 video of that category and set it as the main
    - Related video will be sub-node
    - Distance will be the rate of each videos



Conclusion & Future work

This project enables us to analyze very large data set on trending topics and interests of people through social networking forum.

We have done data analysis on a single YouTube dataset. Future work can be done by doing the same analysis on weekly gathered data to identify the current trending topics and area of interest of the mass.

Future work also includes extracting more important information using other attributes in dataset.

Related Work

Youtube Graph Network Model and Analysis

- figure out which videos the viewer can potentially watch
    - Probability computation, PageRank algorithm

Understanding User Behavior in Large-Scale Video-on-Demand Systems

- Make Analysis on User’s behavior with national media company
    - User Access Pattern over time (Daily, Hourly, Weekly)
        - Used Poisson Distribution
    - Changes of user interest over time
        - When new videos were uploaded, changes of user interest on existing videos

QUESTION?