Wikipedia Peak Detection

Capstone Project Update

November 17th, 2016

Tasks achieved in Data Preprocessing

- The Wikipedia Hourly PageViews files from January 1st 2016 to October 31st 2016 have been downloaded and stored in the the project S3 bucket.
- The MapReduce Scripts have been coded and tested.
- The script to Autodownload the 24 files from a day (Hourly Pageviews) is fully functional.
- The script to start an AWS EMR cluster and run the MapReduce jobs over the Hourly PageViews is fully functional.

With the Autodownload and EMR cluster scripts coded and read, the preprocessing tasks can be fully automatized through schedules or triggered by certain actions.

With these tasks, the integration between Storage, Computing and Analytics services has been achieved.

Peak Detection and Scoring

For each day, we find all the pages that reach the peak at this day. The peaks are simply the most page views in a 15 days interval (the day we chose + former 7 days + later 7 days).
After finding all the pages that reach peaks at a day we chose. We implement a score function below to give each page a score, and then rank these pages by their scores.

HTML Dashboard

A HTML Dashboard will be implemented to visualize in a friendly way these peaks.
The Dashboard will allow the user to:

- Select a specific date from the year.
- Obtain the Top-N peak articles for the selected date.
- View the daily views for one of the Top-N peak articles in a specific time window.
- Compare the daily views of that Top-N peak article with its other language versions.

The Dashboard will be hosted in the AWS, so it can be consulted by any person.

Next Steps

- Apply Score Function to the English Wikipedia articles for all days available.
- Storing Scores in DB.
- Implementation of Dashboard Interface.
- Connection of DB to Dashboard (for user queries).
- Deployment of System.

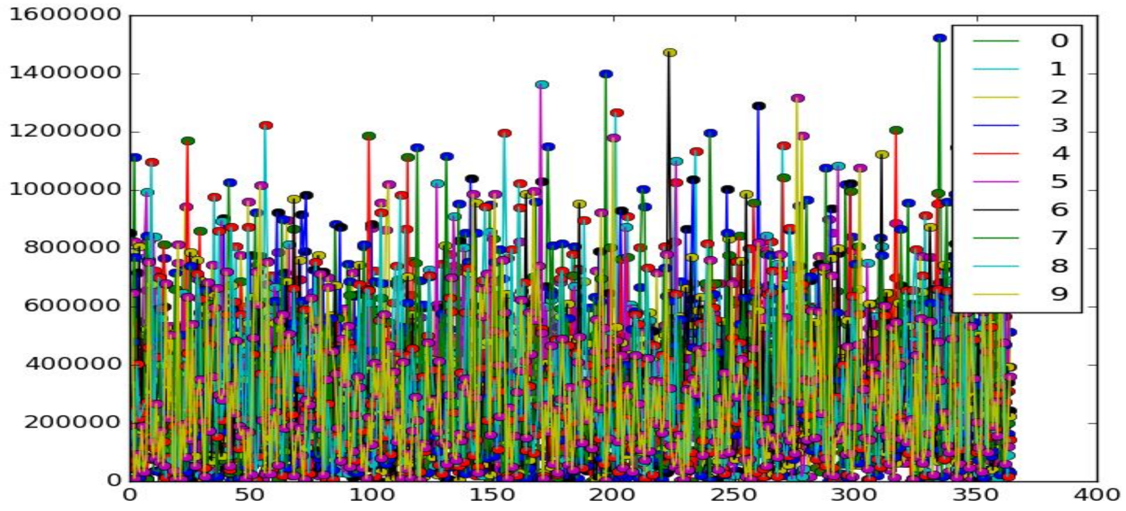
Images and Plots:

Scoring Formulas

$$personalized\_page\_views = \frac{peak\_page\_views - average\_page\_views}{peak\_page\_views}$$

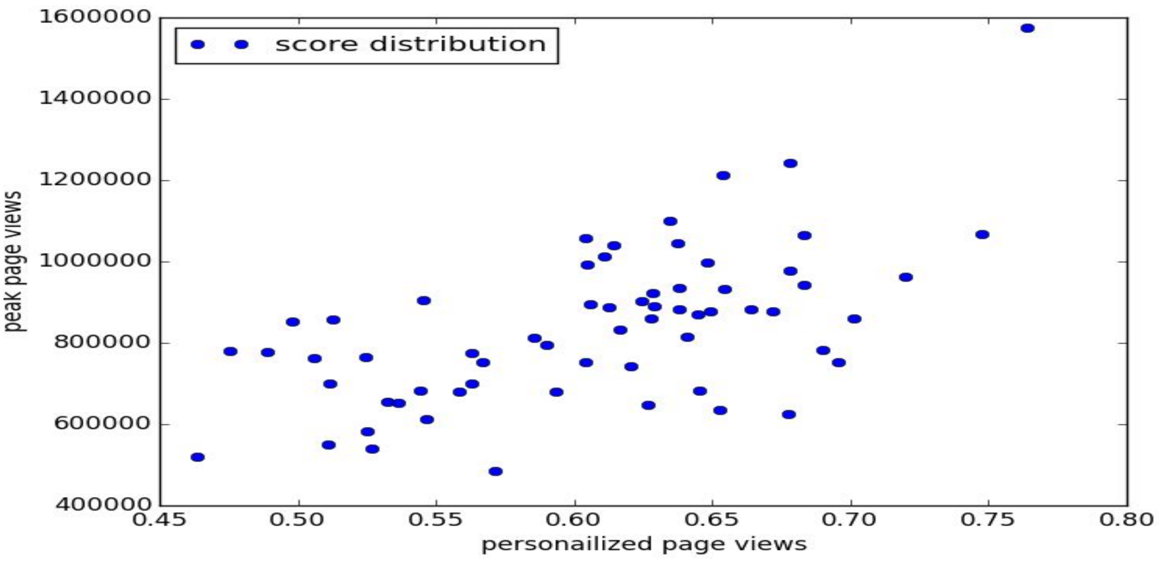
$$Score = \lambda_1 \times peak\_page\_views + \lambda_2 \times personalized\_page\_views$$

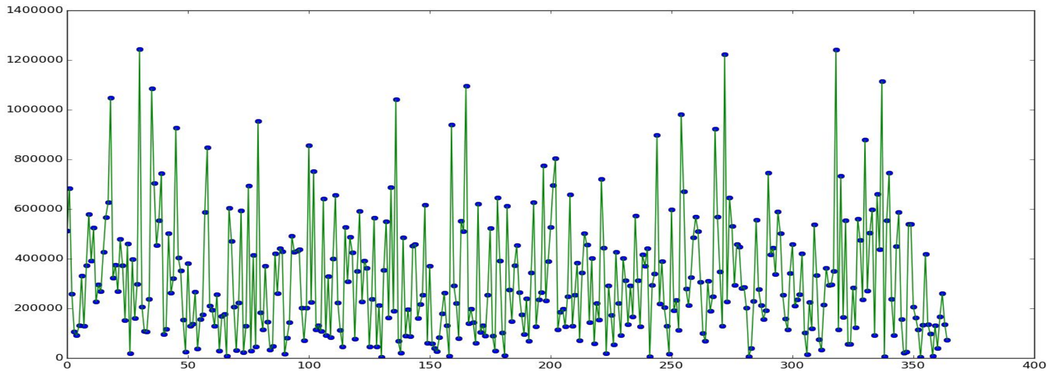
Ten Pages' year curves

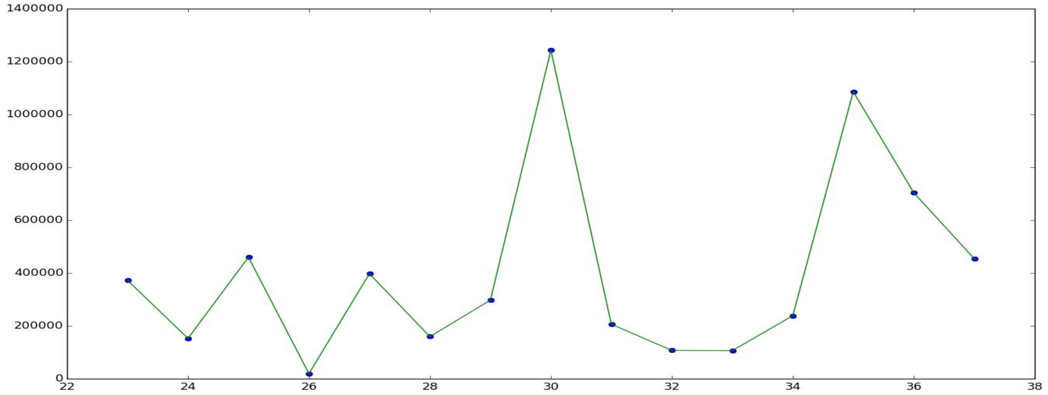


Ten pages' year curves

Score Distribution


● ● score distribution
peak page views
personailized page views


1400000
1200000
1000000
800000
600000
400000
200000
0
0
50
100
150
200
250
300
350
400


1400000
1200000
1000000
800000
600000
400000
200000
0
22
24
26
28
30
32
34
36
38