Wikipedia Data Analysis using Hadoop

Mentor: Dr. James Abello

Presenters: Ashish Jindal, Yikun Xian and Sanjivi Muttena

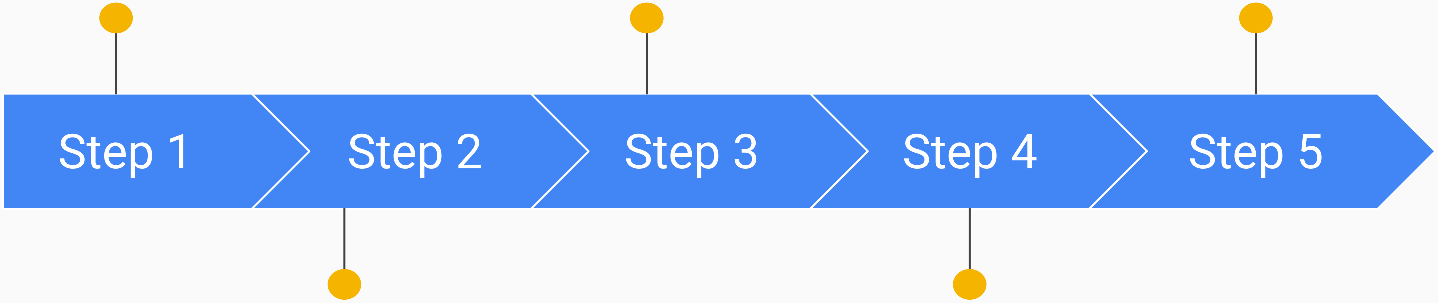
Outline

- Project Goals
- Project description
- Infrastructure Overview
- Findings and Results
- Stumbling blocks
- References
- Ideas for Extension

Project Goals

- Set Up an infrastructure for hadoop based data-analytics.
- Implement a trend estimation algorithm for Wikipedia data.
- Calculate page ranks for various Wikipedia Pages.
- Create an interactive web app to visualize the calculated data.

Project Description



Step 1

Set up a map-reduce development environment on local system.

Step 2

Write map-reduce jobs for trend estimation and page rank calculation and test them on local system using a small subset of data.

Step 3

Setup AWS based infrastructure (EMR and S3) to analyze bigger dataset.

Step 4

Dump the output from Amazon EMR to an interactive database system (MongoDB).

Step 5

Create a Web application that fetches data from MongoDB and visualizes result using a JS library (Highcharts).

The Team

Ashish Jindal

1. Setup Hadoop infrastructure using EMR and S3.
2. Implemented Page-Rank calculation using map-reduce paradigm.
3. Implemented a simple baseline algorithm for calculating trend factor.

Yikun Xian

1. Created the web-application using Spring MVC.
2. Implemented visualization of data using JS library (Highcharts).
3. Implemented data cleaning jobs for Wikipedia data using map-reduce.

Sanjivi Muttena

1. Implemented aggregation map-reduce jobs for accumulating hourly Wikipedia data.
2. Setup Interactive database system (MongoDB) on an Amazon EC2 server.
3. Setup database interactivity in web-application using Spring Data API.

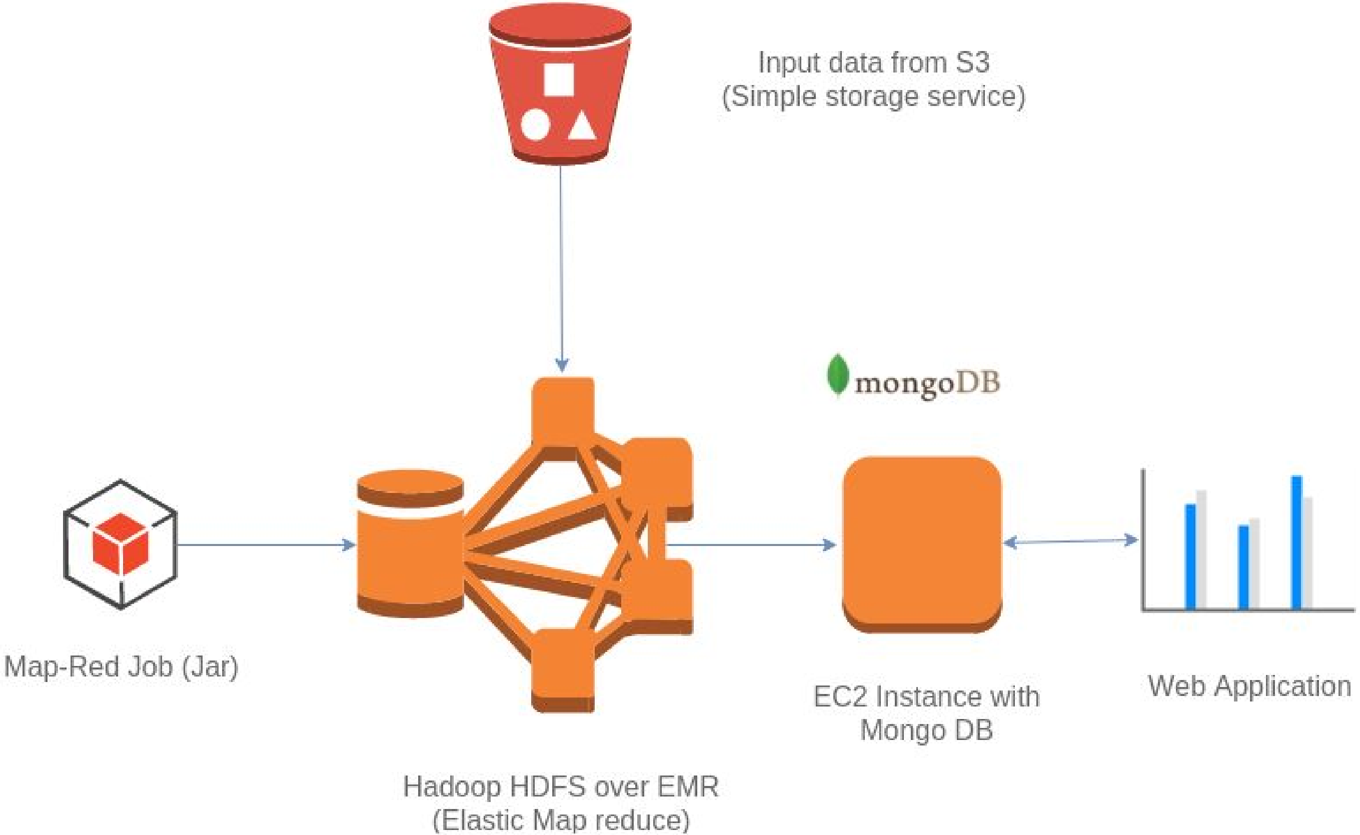
Hadoop

What is it?
Why do we need it?

The Apache Hadoop is a framework for the distributed processing of large data sets across clusters of computers using simple programming models.

- Commodity inexpensive hardware.
- Efficient and simple fault tolerant mechanism .
- Scalability.
- Accepts all data formats. No predefined schema.

Infrastructure



Input data from S3
(Simple storage service)

mongoDB

Map-Red Job (Jar)

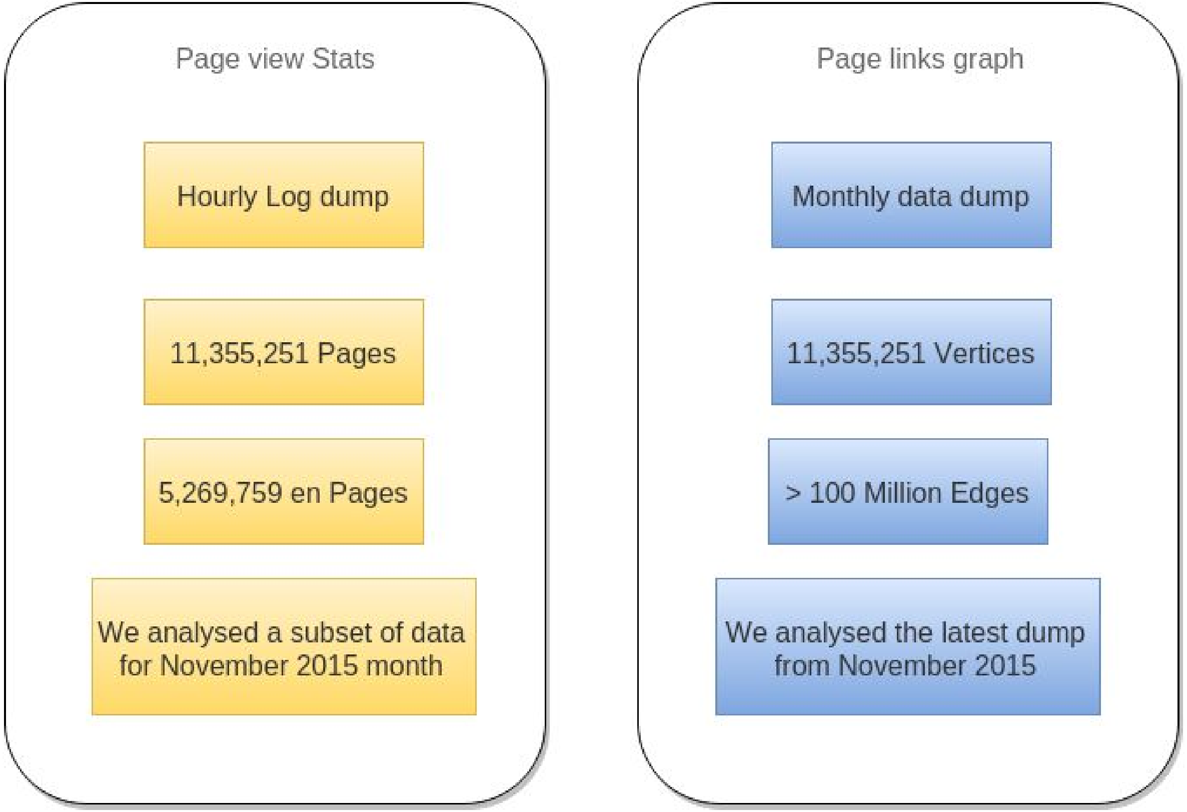
EC2 Instance with
Mongo DB

Web Application

Hadoop HDFS over EMR
(Elastic Map reduce)

Dataset

WIKIPEDIA
The Free Encyclopedia



Page view Stats

Hourly Log dump

11,355,251 Pages

5,269,759 en Pages

We analysed a subset of data for November 2015 month

Page links graph

Monthly data dump

11,355,251 Vertices

> 100 Million Edges

We analysed the latest dump from November 2015

Page Rank

What is page rank?

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.

$$PageRank\ of\ site = \sum \frac{PageRank\ of\ inbound\ link}{Number\ of\ links\ on\ that\ page}$$

OR

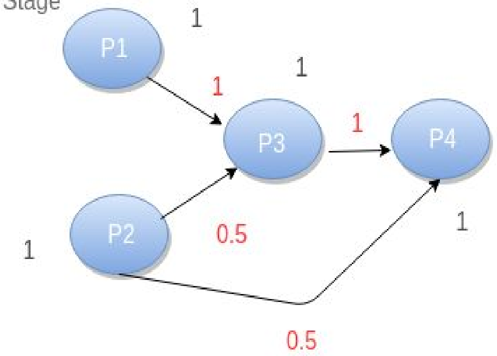
$$PR(u) = (1 - d) + d \times \sum \frac{PR(v)}{N(v)}$$

The underlying assumption is that more important websites are likely to receive more links from other websites

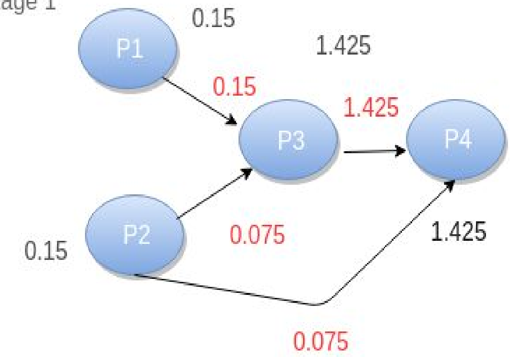
Page Rank

Calculation

Initial Stage



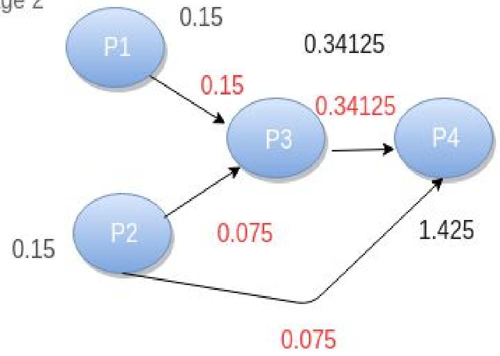
Stage 1



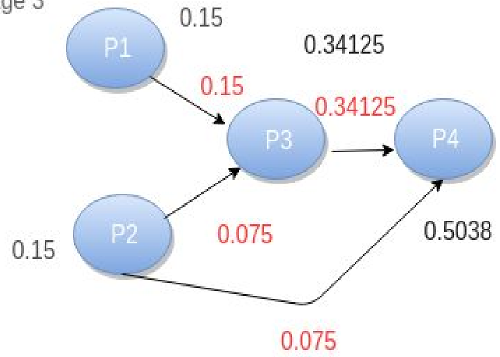
PR = (1 - DF) + DF * (Total PR contribution from inbound links)

DF = 0.85 in our application

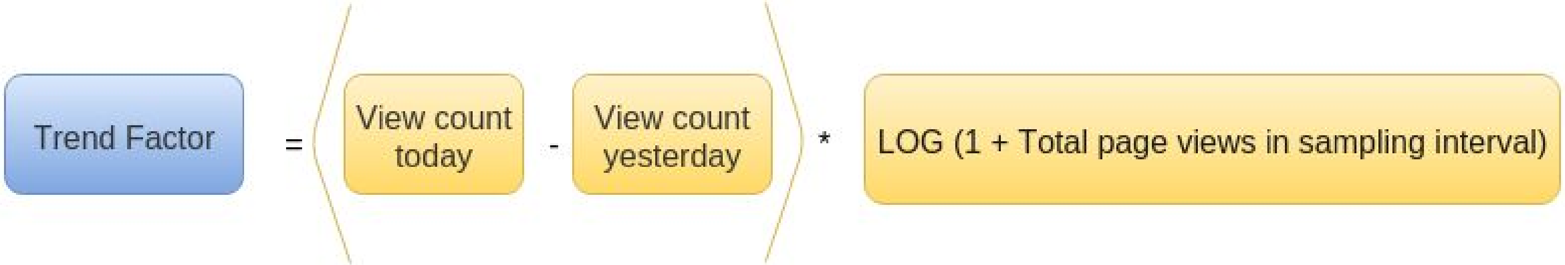
Stage 2



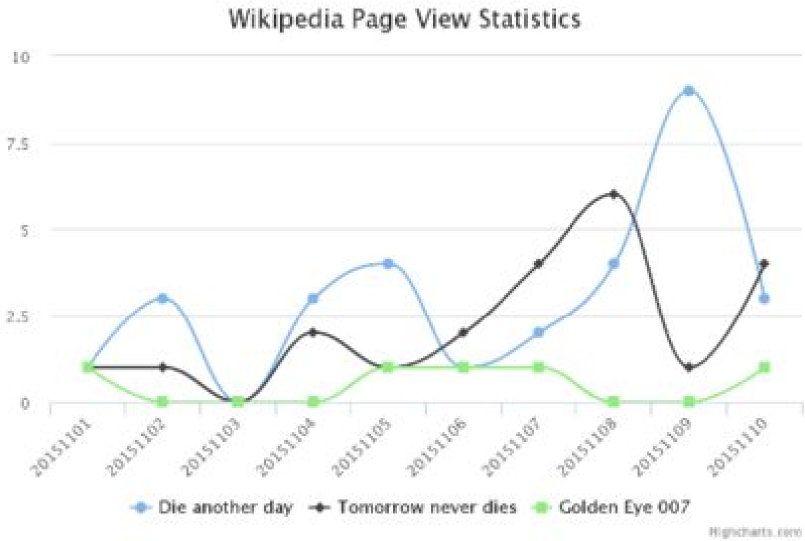
Stage 3



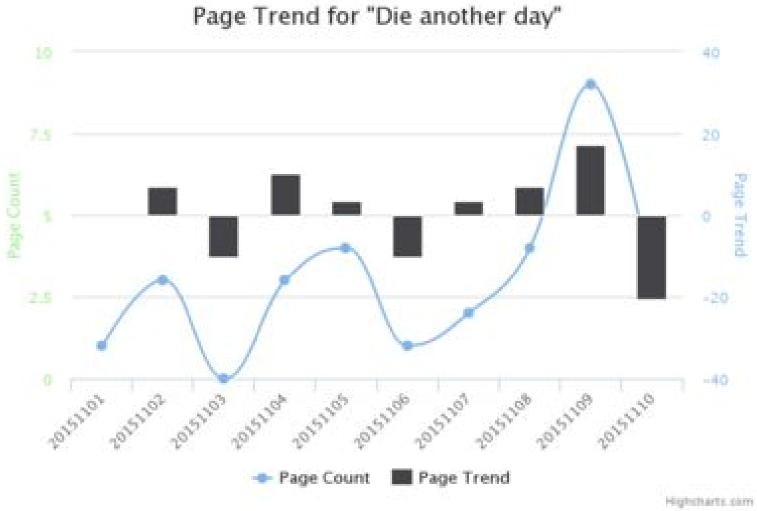
Trend Factor Calculation



Trend Factor = (View count today - View count yesterday) * LOG (1 + Total page views in sampling interval)

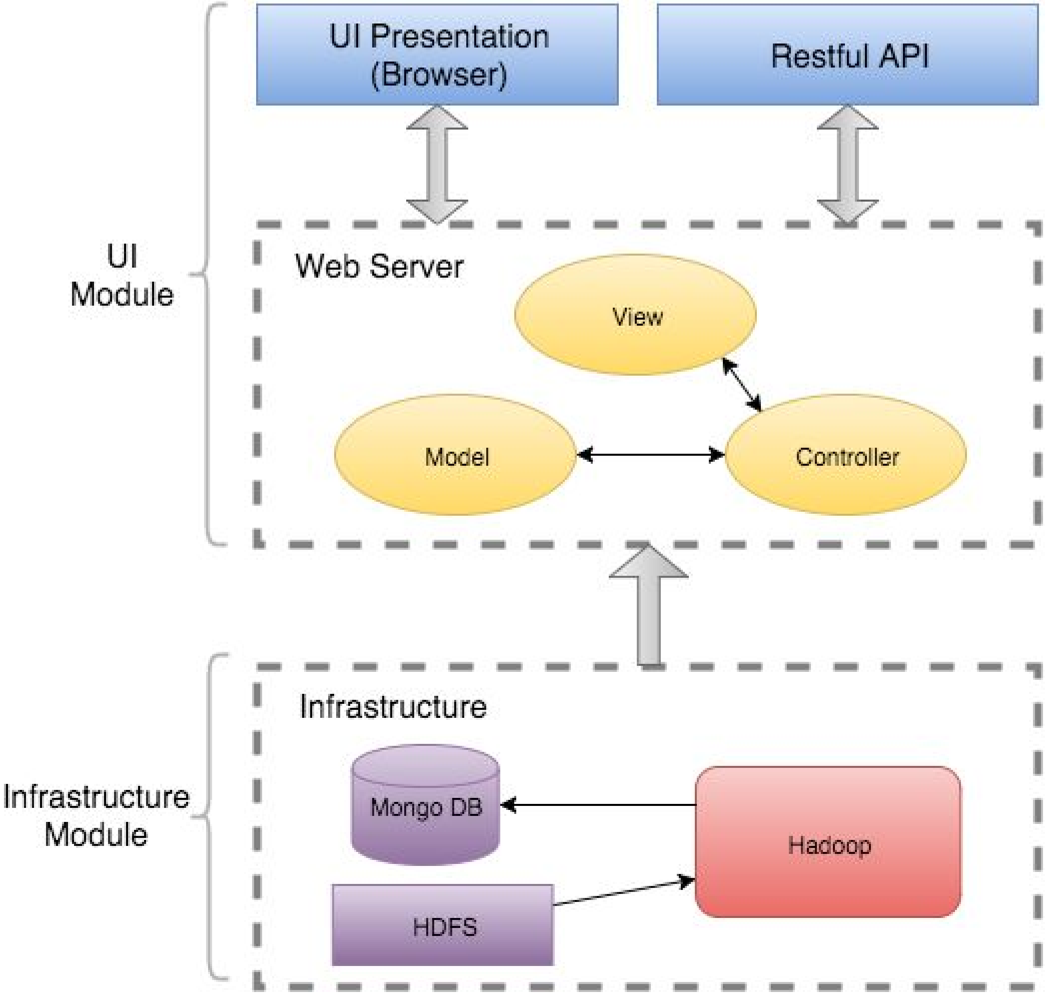


Wikipedia Page View Statistics

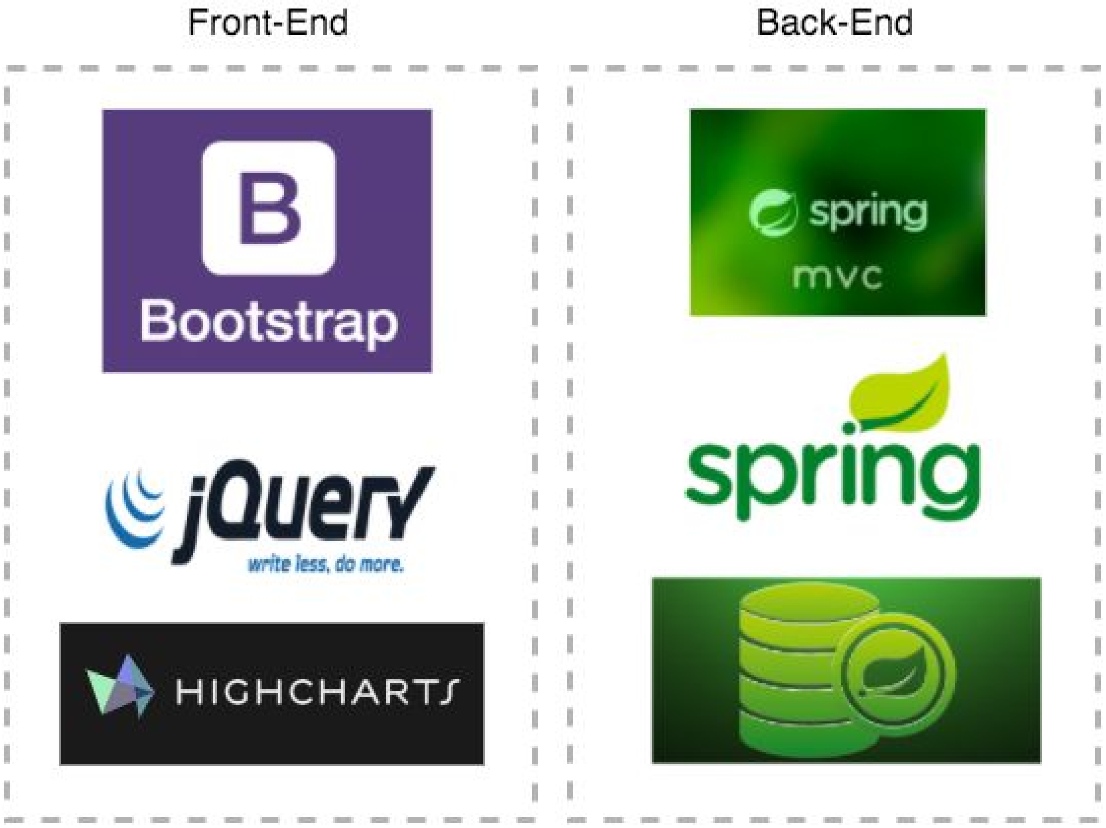


Page Trend for "Die another day"

Web App Architecture



Web App Technology



Front-End

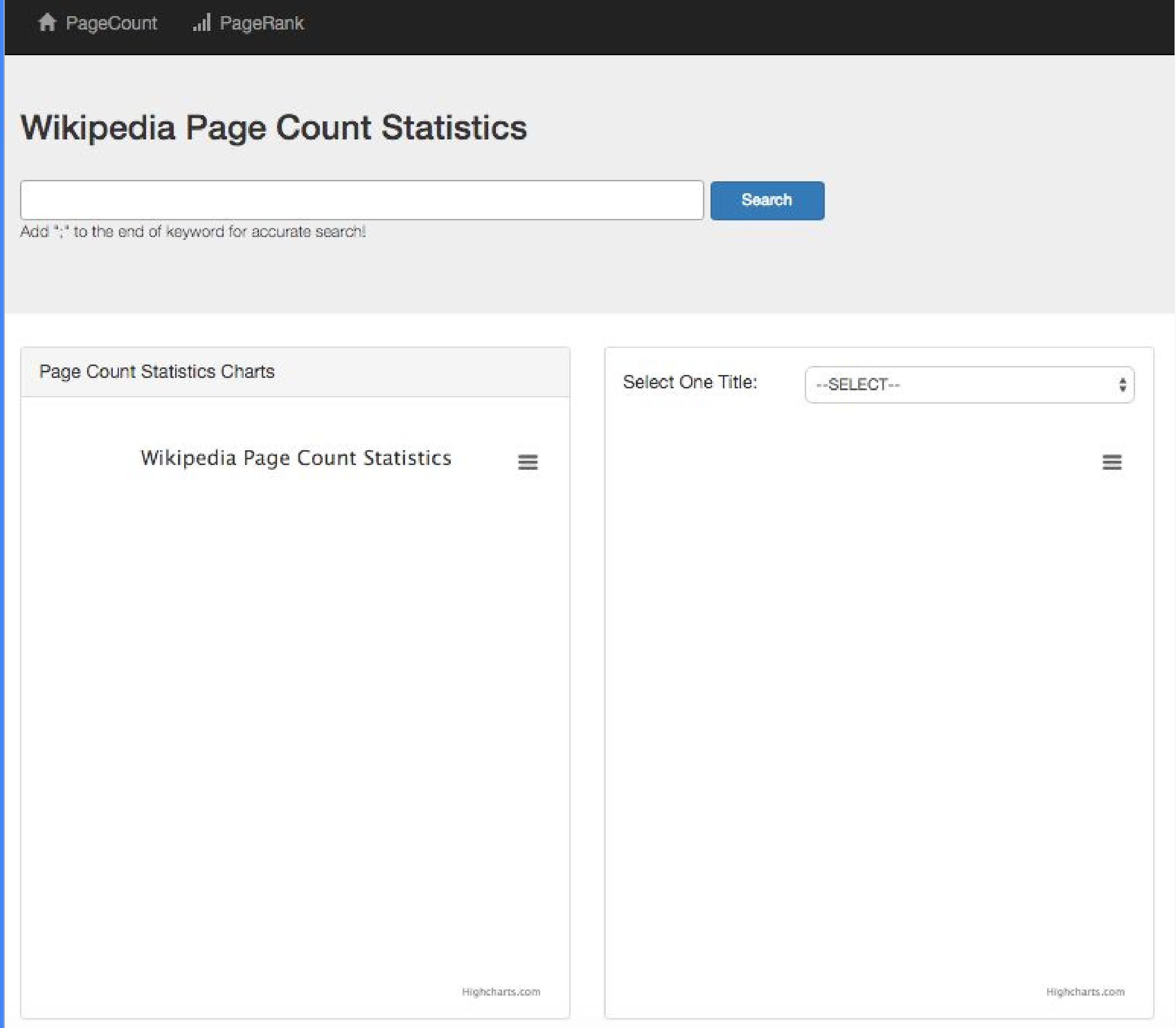
Back-End

Findings & Results

- Page View Count
- Page Trending
- Input/Output
- Page Rank Index
- Interesting Findings

Page View Count

Layout



Wikipedia Page Count Statistics

Add ";" to the end of keyword for accurate search!

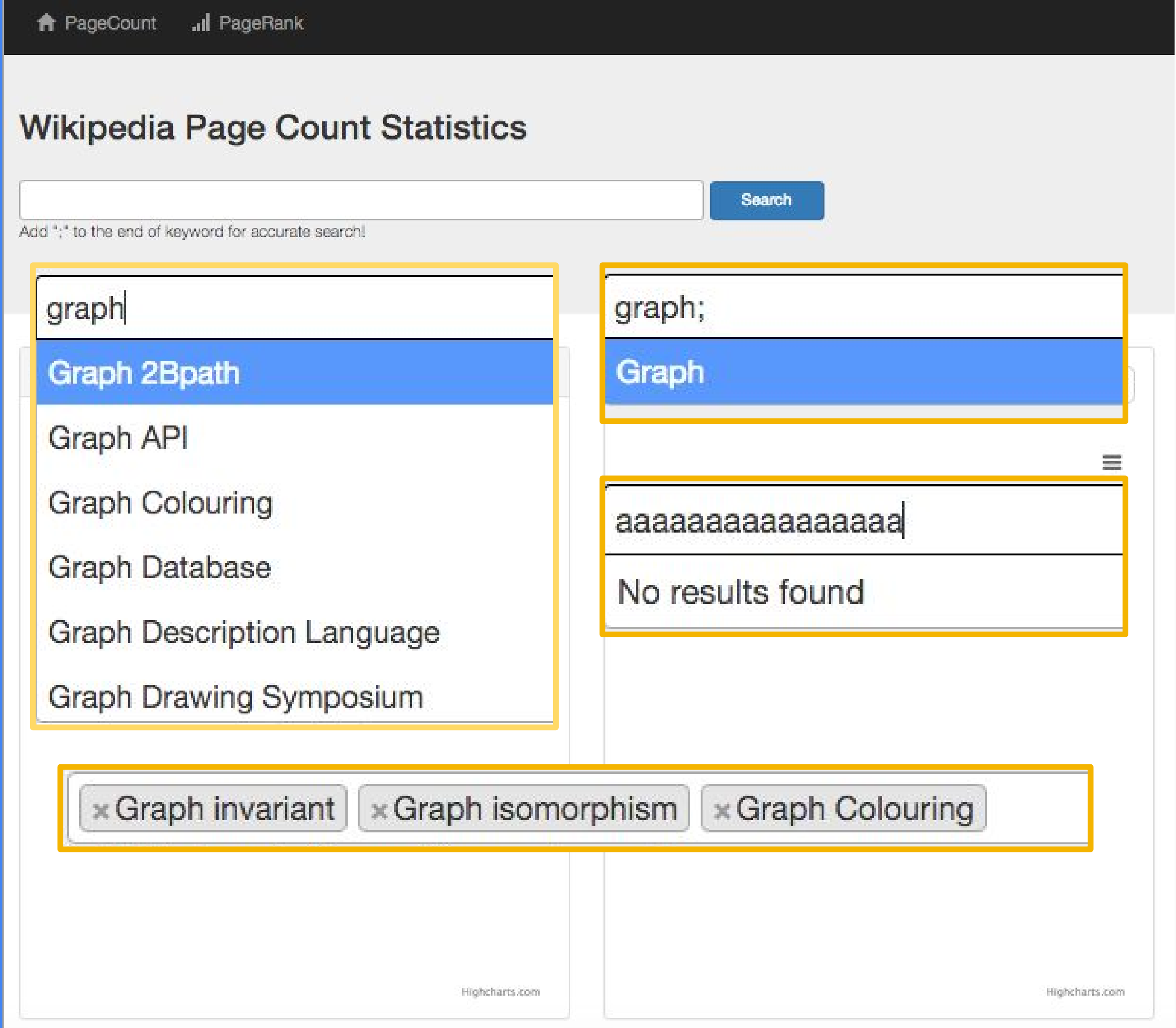
Page Count Statistics Charts

Wikipedia Page Count Statistics

Select One Title:

Page View Count

Input



Wikipedia Page Count Statistics

Add ";" to the end of keyword for accurate search!

graph

Graph 2Bpath
Graph API
Graph Colouring
Graph Database
Graph Description Language
Graph Drawing Symposium

graph;

Graph

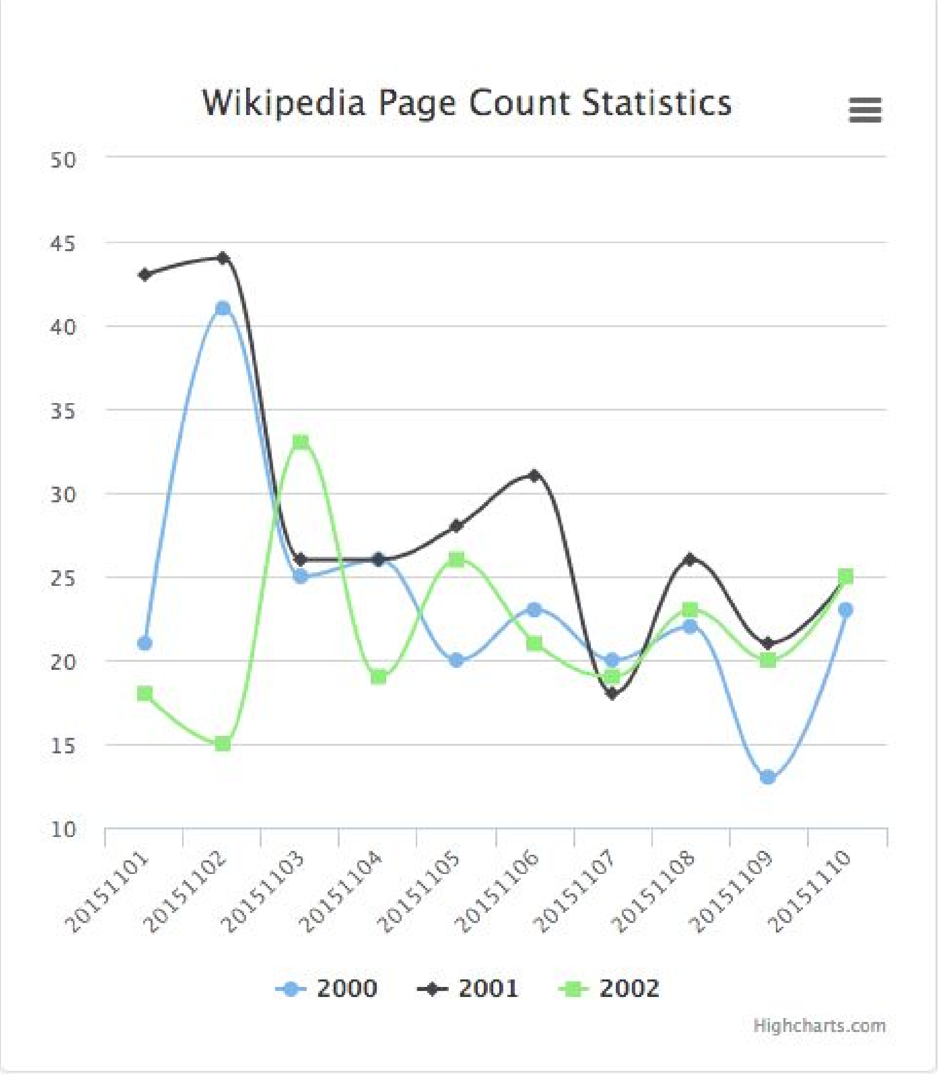
aaaaaaaaaaaaaaa

No results found

× Graph invariant × Graph isomorphism × Graph Colouring

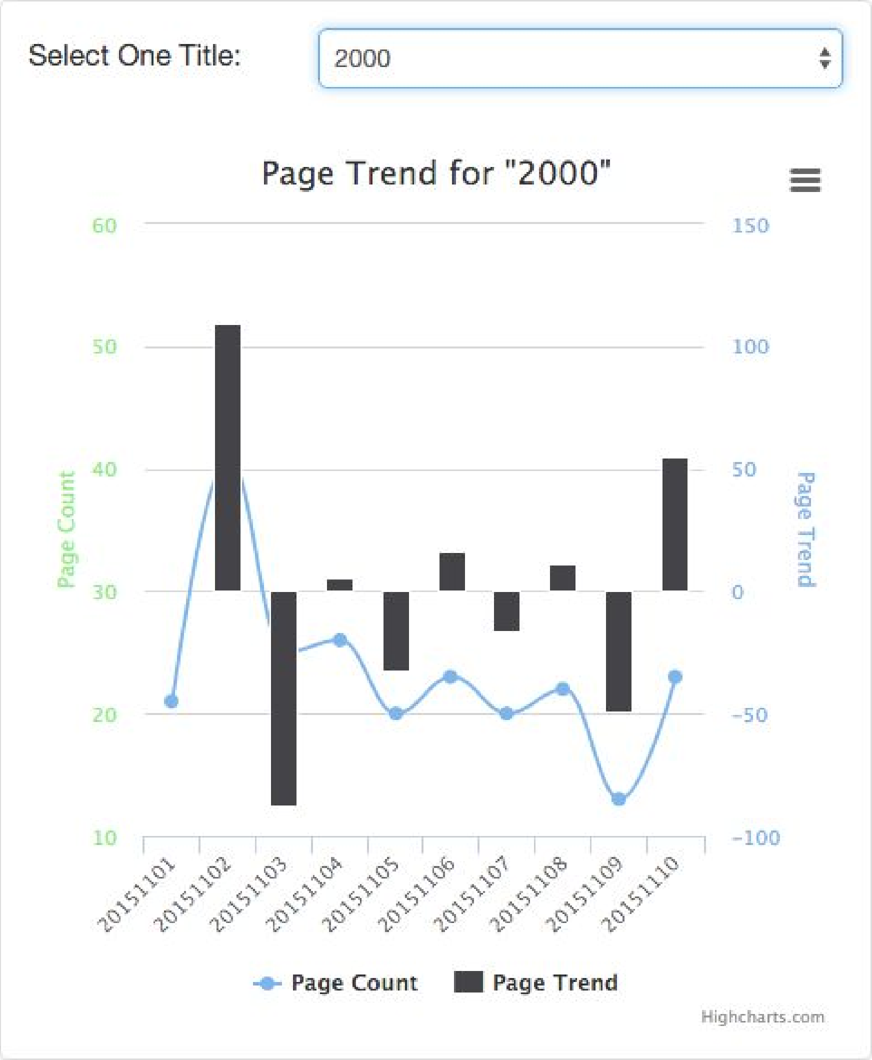
Page View Count

Page View Count



Page View Count

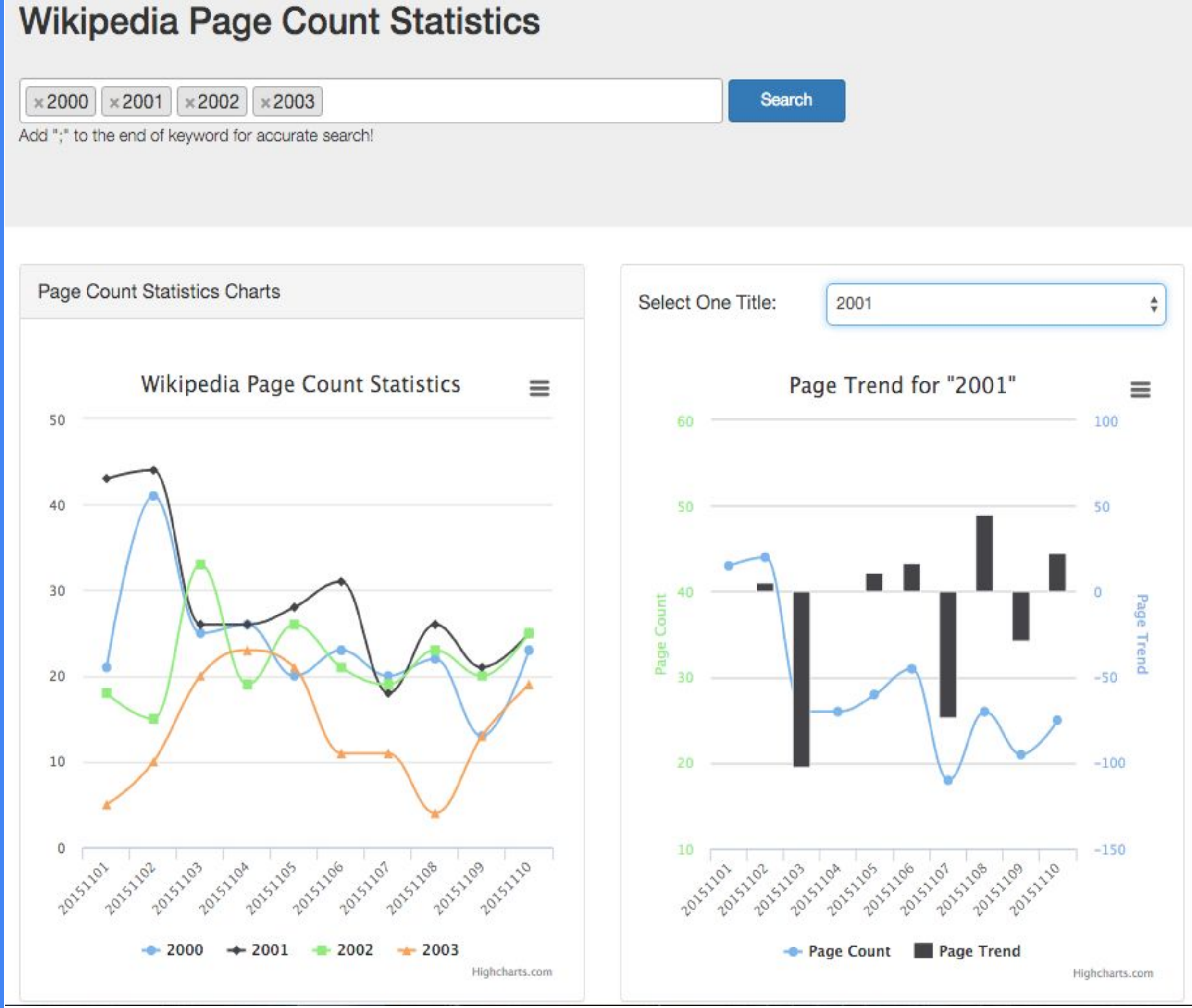
Page View Count



Select One Title: 2000

Page View Count

Overview



Wikipedia Page Count Statistics

×2000 ×2001 ×2002 ×2003 Search

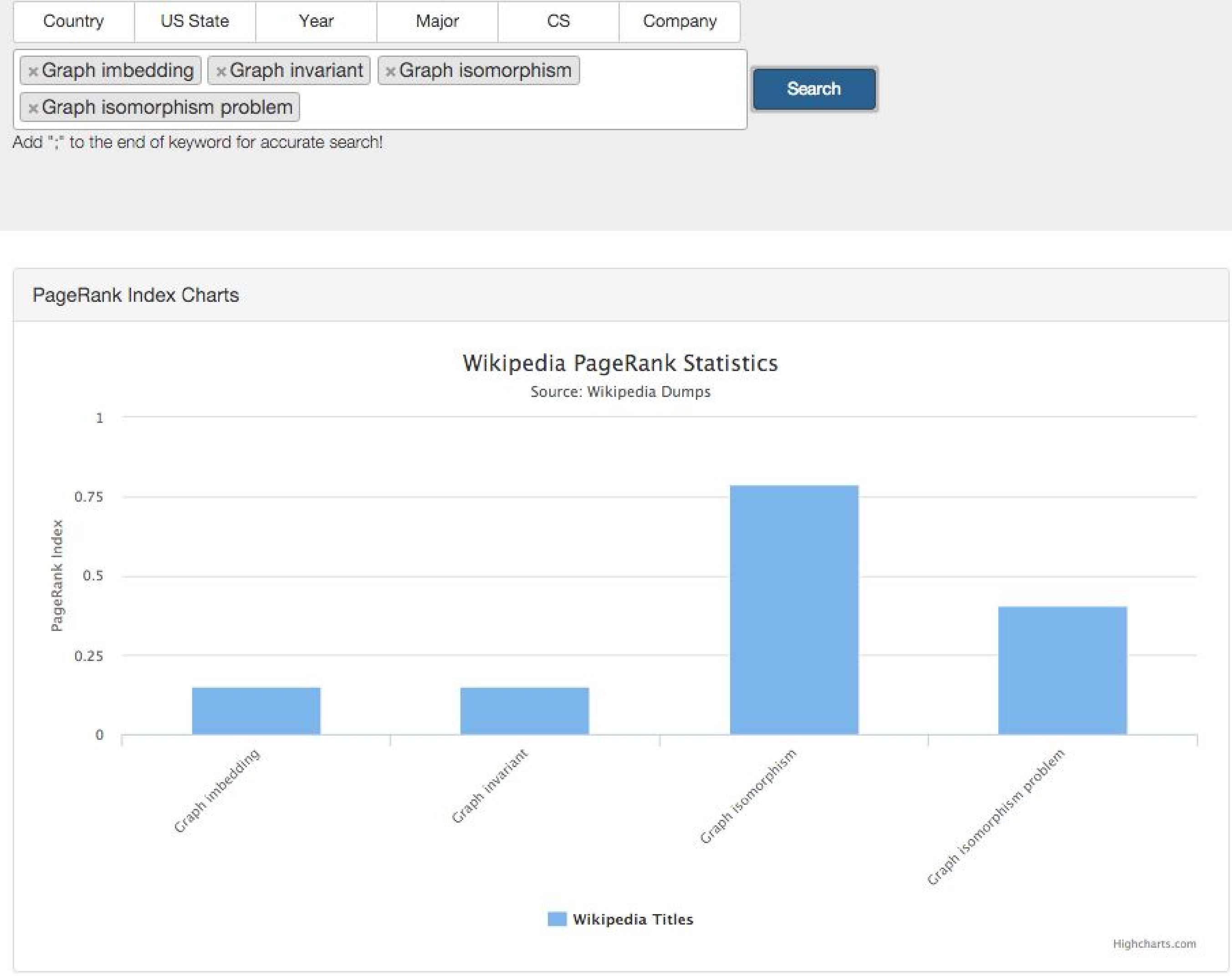
Add ";" to the end of keyword for accurate search!

Page Count Statistics Charts

Select One Title: 2001

Page Rank

Search



Add ";" to the end of keyword for accurate search!

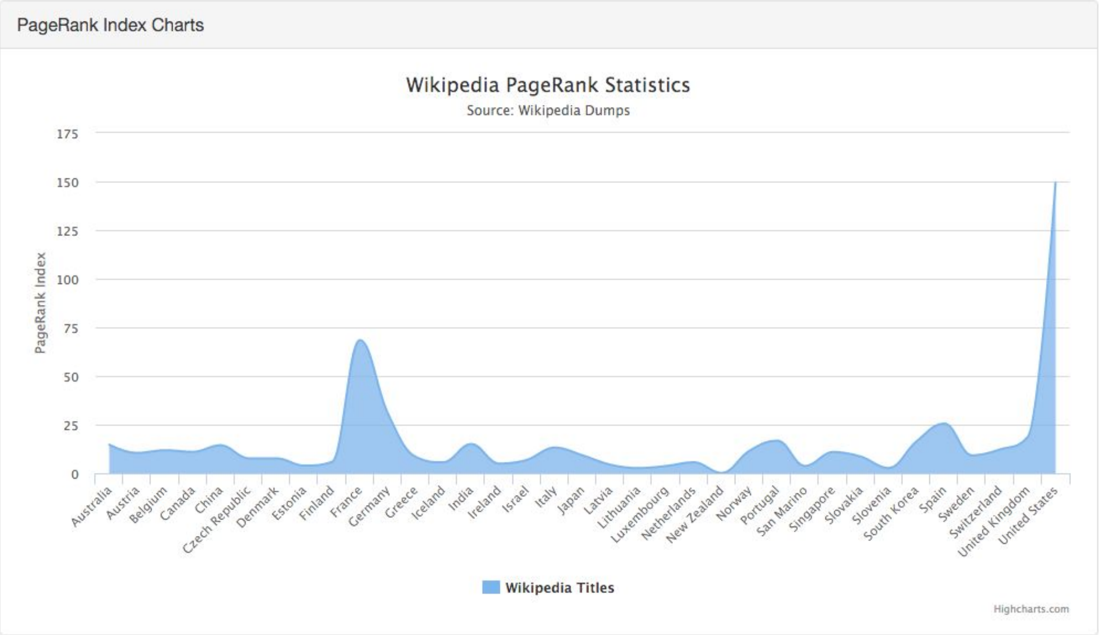
PageRank Index Charts

Wikipedia PageRank Statistics

Source: Wikipedia Dumps

Page Rank

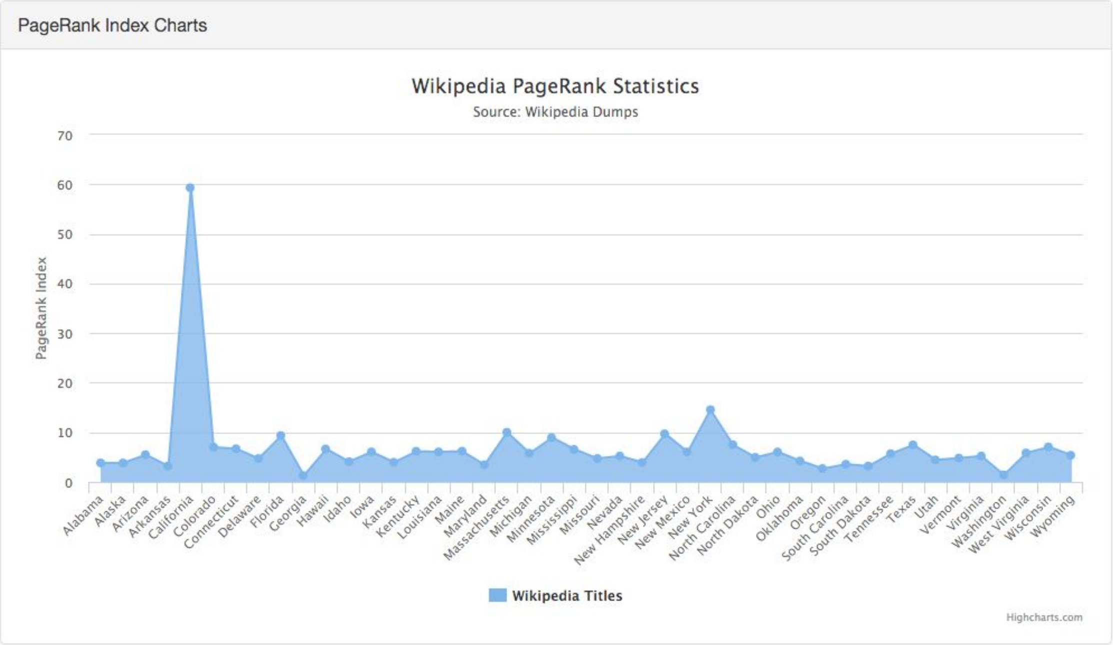
Countries



PageRank Index Charts

Page Rank

US States



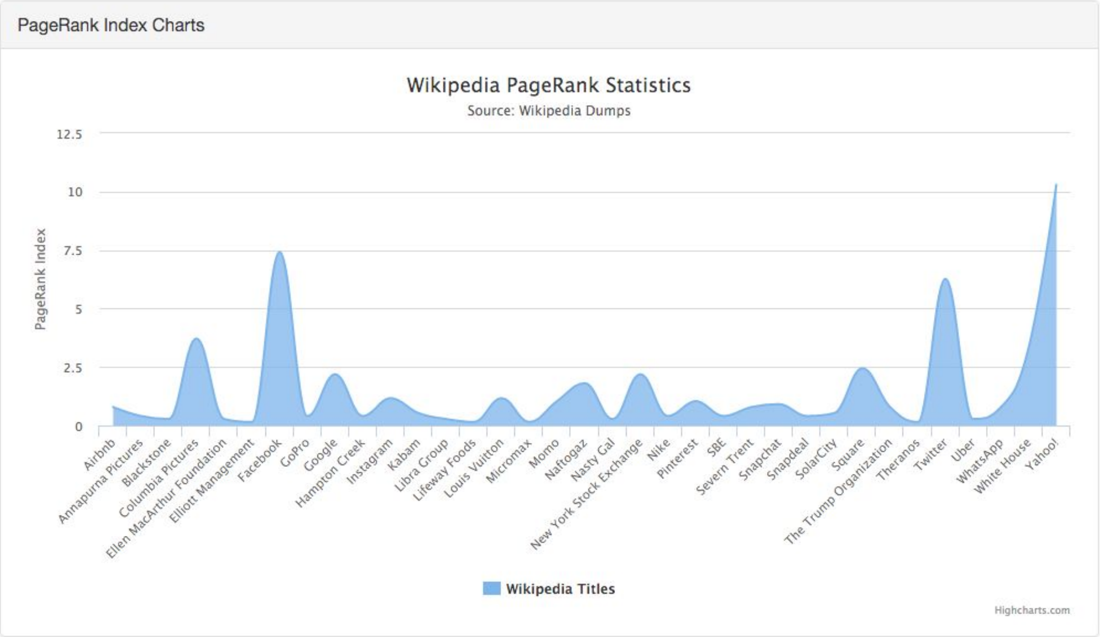
PageRank Index Charts

Wikipedia PageRank Statistics

Source: Wikipedia Dumps

Page Rank

Companies



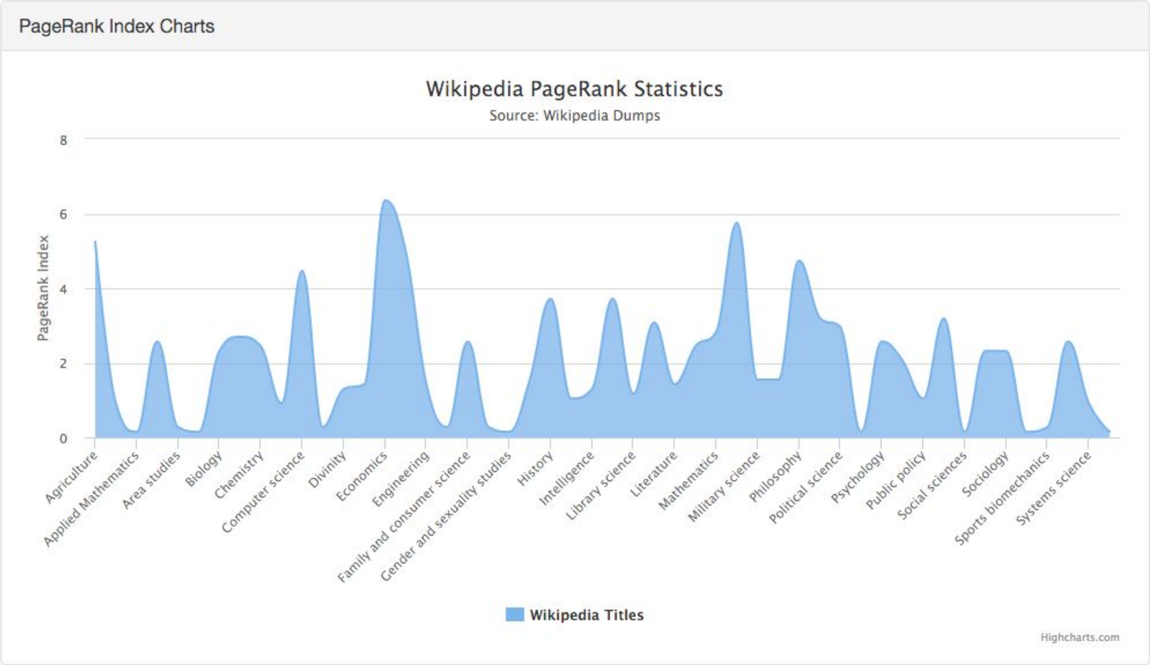
PageRank Index Charts

Wikipedia PageRank Statistics

Source: Wikipedia Dumps

Page Rank

Various Majors



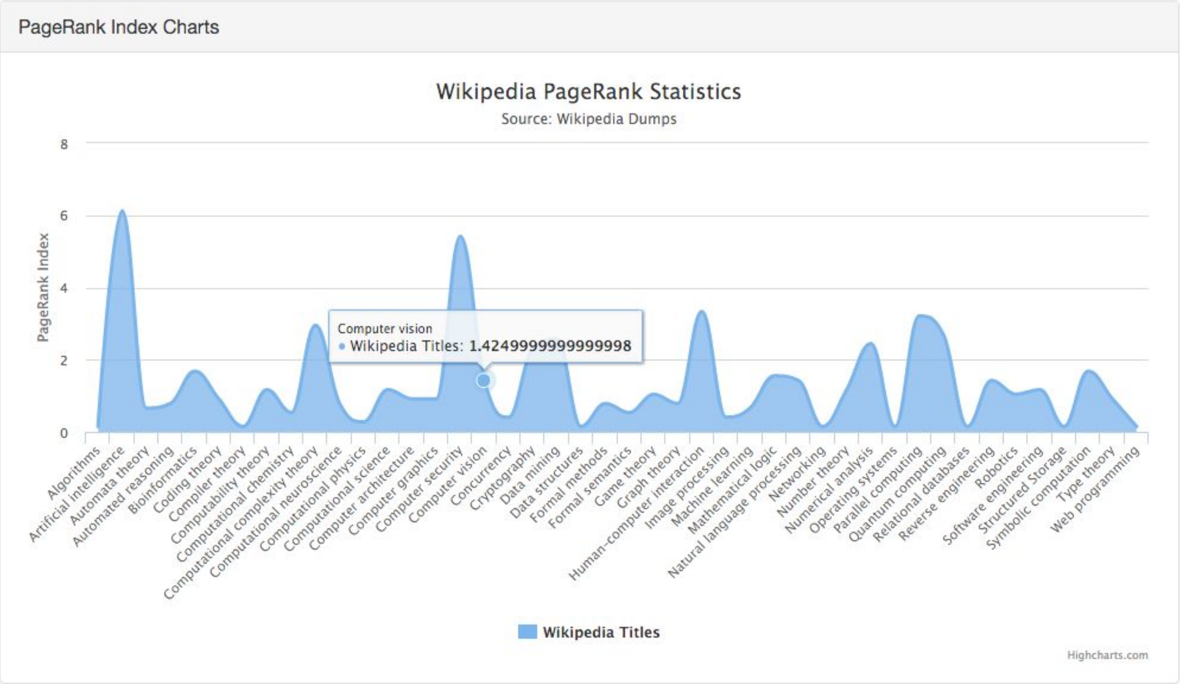
PageRank Index Charts

Wikipedia PageRank Statistics

Source: Wikipedia Dumps

Page Rank

Various CS Areas



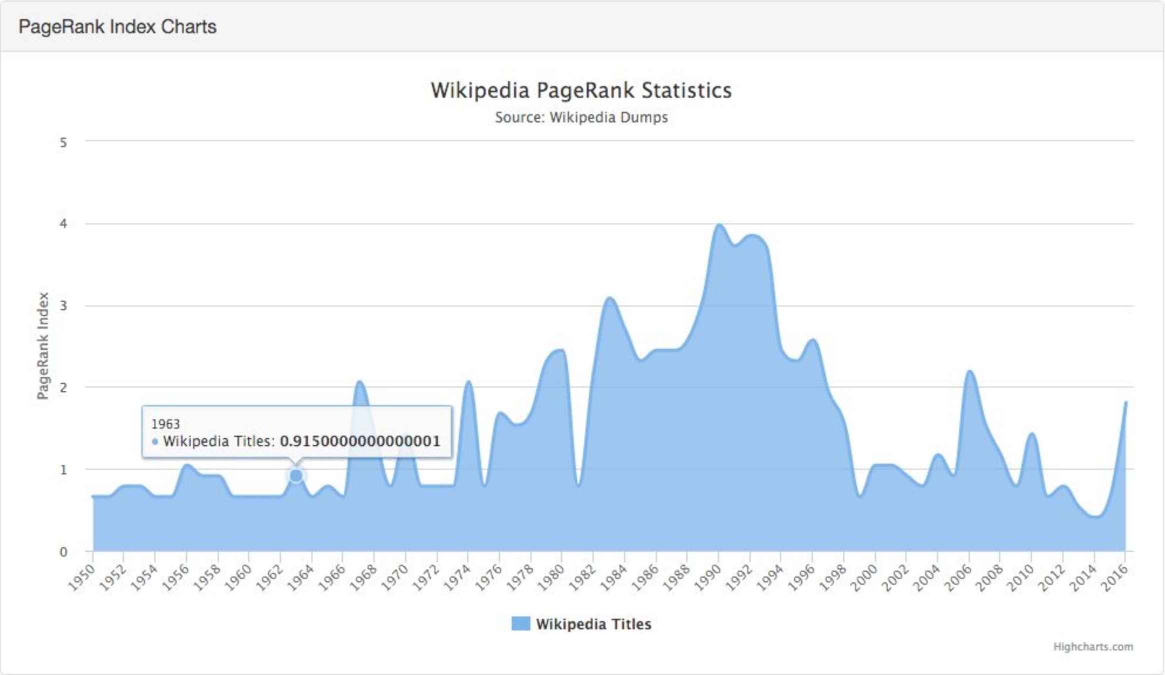
PageRank Index Charts

Wikipedia PageRank Statistics

Source: Wikipedia Dumps

Page Rank

Years



Stumbling Blocks

- Setting up hadoop development environment and infrastructure for processing data.
- Integration of Elastic Map Reduce with MongoDB instance on EC2.
- Visual Chart Asynchronous Refresh.
- Implementing fast autocompletion in Wiki page search box in the Web-app.

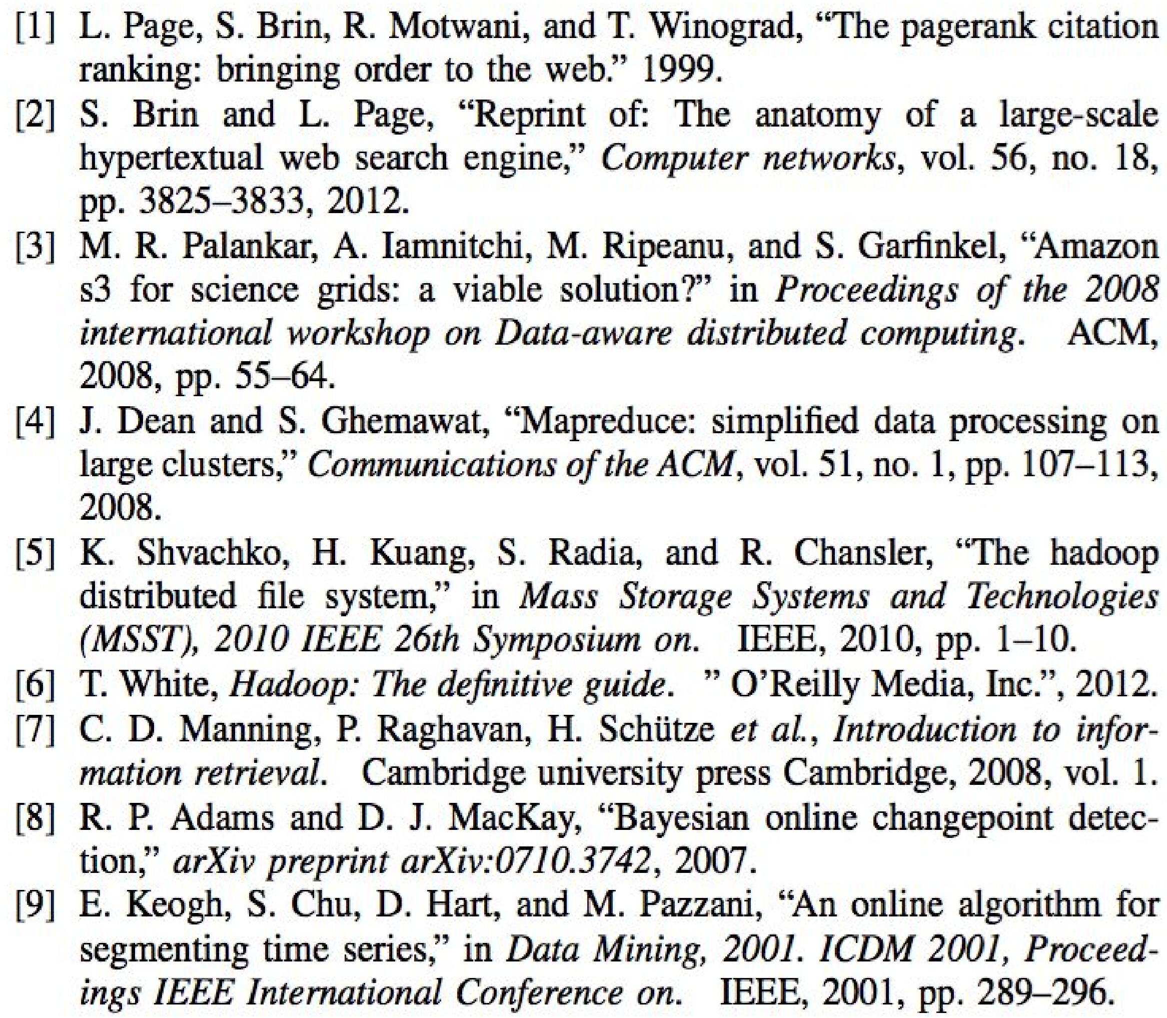
Ideas for extension

- Find weekly popular/trending topics based on calculated trend factor.
- Use page link data to find topic relations like events in Germany in year 2000 based on outbound links on Wikipedia page for year 2000.
- Correlate page view count on Wikipedia pages for movies with the movie reviews.

Acknowledgment

We would like to thank our supervisor, Dr. James Abello, for assisting us with technology, algorithm and constructive suggestions which made our project promising and competitive.

References



[1] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: bringing order to the web.” 1999.

[2] S. Brin and L. Page, “Reprint of: The anatomy of a large-scale hypertextual web search engine,” *Computer networks*, vol. 56, no. 18, pp. 3825–3833, 2012.

[3] M. R. Palankar, A. Iamnitchi, M. Ripeanu, and S. Garfinkel, “Amazon s3 for science grids: a viable solution?” in *Proceedings of the 2008 international workshop on Data-aware distributed computing*. ACM, 2008, pp. 55–64.

[4] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[5] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The hadoop distributed file system,” in *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*. IEEE, 2010, pp. 1–10.

[6] T. White, *Hadoop: The definitive guide*. ” O’Reilly Media, Inc.”, 2012.

[7] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1.

[8] R. P. Adams and D. J. MacKay, “Bayesian online changepoint detection,” *arXiv preprint arXiv:0710.3742*, 2007.

[9] E. Keogh, S. Chu, D. Hart, and M. Pazzani, “An online algorithm for segmenting time series,” in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE, 2001, pp. 289–296.

Any Questions?