Exploratory Data Analysis for English Wikipedia

Ashwin Nimhan
Indiana University, USA
animhan@indiana.edu

Manashree Rao
Indiana University, USA
manarao@indiana.edu

1 Project Description

In this paper, we focus on English Wikipedia data that has been open-sourced by Mediwiki. The primary objective of the project is deploying the Wikipedia data on Apache Spark-HDFS-Yarn stack and analyzing two main sub-types of Wiki data - Clickstream and Pageviews data. Ansible scripts are used to deploy the system on an Openstack cluster. We highlight different aspects of data left on servers each time a user accesses a web page in the form of clicks and page view statistics. We have analyzed pageviews data to analyze different aspects of user like the device he is using, which day of the week user is most active on the website etc. On analyzing wikipedia clickstream data, we can calculate the path that user follows on Wiki, which links are requested but missing, and have also demonstrated how the information on presidential candidates is accessed on Wikipedia.

2 Problem statement

Big Data open source software like as Spark, Hadoop and MongoDB are incredibly important for analyzing data that has large volume and variety. We are using English Wikipedia clickstream and pageviews data to display how to data can be deployed on such big data stacks and analyzed effectively in order to derive insights.

3 Purpose and Objectives

Typical questions that are related to visitor behaviour are the frequency and length of visits during a certain time period, the entrance and exit locations of visitors, the percentage of visitors who reach key pages (such as a signup page, cash register, etc), the paths they take, the traffic trend, the prediction of traffic spikes, the accommodation of server space for increased traffic, the adjustment for browser technology, the evaluation of behaviour variations among subsets of customers and the change during sales, etc.

Pageviews data :
A pageview is each time a visitor views a page the website, regardless of how many hits are generated. Pages are comprised of files. Every image in a page is a separate file. When a visitor looks at a page (a pageview), they may see numerous images, graphics, pictures etc. and generate multiple hits. Hence page views and not hits are analyzed. Pageviews data can be analyzed to check distribution of views over the day, day of the week or even month, the medium used to access the website etc.

Clickstream data:
Is an information trail a user leaves behind while visiting a website. It is typically captured in semi-structured website log files. These website log files contain data elements such as a date and time stamp, the visitors IP address and the destination URLs of the pages visited. Clickstream analysis is the process of collecting, analyzing, and reporting aggregate data about which pages visitors visit in what order - which are the result of the succession of mouse clicks each visitor makes (that is, the clickstream). There are many applications of Clickstream data like finding the most visited path on website, most visited sections of the website, top referrers linking to the website, missing/broken links etc which users go

to, etc.

4 Dataset

4.1 Wikipedia Clickstream Data

The data contains counts of (referer, resource) pairs extracted from the request logs of English Wikipedia. When a client requests a resource by following a link or performing a search, the URI of the webpage that linked to the resource is included with the request in an HTTP header called the ”referer”. This data captures 22 million (referer, resource) pairs from a total of 3.2 billion requests collected during the month of February 2015. Referers were mapped to a fixed set of values corresponding to internal traffic or external traffic from one of the top 5 global traffic sources to English Wikipedia, based on this scheme:

```
Wiki article -> the article title
an empty referer -> other-empty
other Wikimedia project -> other-internal
Google -> other-google
Yahoo -> other-yahoo
Bing -> other-bing
Facebook -> other-facebook
Twitter -> other-twitter
anything else -> other-other
```

Format
The data includes the following 6 fields:

1. **prev_id:** if the referer does not correspond to an article in the main namespace of English Wikipedia, this value will be empty. Otherwise, it contains the unique MediaWiki page ID of the article corresponding to the referer i.e. the previous article the client was on
2. **curr_id:** the unique MediaWiki page ID of the article the client requested
3. **n** the number of occurrences of the (referer, resource) pair
4. **prev_title** the result of mapping the referer URL to the fixed set of values described above
5. **curr_title** the title of the article the client requested
6. **type**
   (a) ”link” if the referer and request are both articles and the referer links to the request
   (b) ”redlink” missing pages
   (c) ”other” if the referer and request are both articles

4.2 Pageviews Data

This file contains a count of pageviews to the English-language Wikipedia from 2015-03-16T00:00:00 to 2015-04-25T15:59:59 grouped by timestamp (down to a one-second resolution level) and site (mobile or desktop).

Format

The data includes the following fields:

1. **requests:** Count of pageviews
2. **site:** mobile or desktop
3. **timestamp:** timestamp of pageviews

5 Implementation

5.1 Deployment

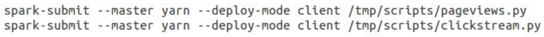
Ansible scripts are used to deploy the system on an Openstack cluster.

Steps to Deploy:

1. Download script.sh from git src folder:

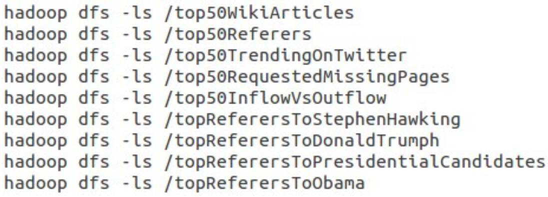
   https://github.iu.edu/animhan/sw-project-template/blob/master/src/script.sh

2. make sure CH-817724-openrc.sh is present under ˜/
3. run: ”bash script.sh”
4. ssh into master0 node
5. switch user: sudo su - hadoop
6. Run following commands



```
spark-submit --master yarn --deploy-mode client /tmp/scripts/pageviews.py
spark-submit --master yarn --deploy-mode client /tmp/scripts/clickstream.py
```

7. To check the output saved to HDFS; run the following commands:
   hadoop dfs -ls /top50WikiArticles



```
hadoop dfs -ls /top50WikiArticles
hadoop dfs -ls /top50Referers
hadoop dfs -ls /top50TrendingOnTwitter
hadoop dfs -ls /top50RequestedMissingPages
hadoop dfs -ls /top50InflowVsOutflow
hadoop dfs -ls /topReferersToStephenHawking
hadoop dfs -ls /topReferersToDonaldTrumph
hadoop dfs -ls /topReferersToPresidentialCandidates
hadoop dfs -ls /topReferersToObama
```

5.2 Analytics

Both datasets have been extensively analyzed to generate insights.

For Pageviews data we try to answer questions like:

1. No of incoming requests in mobile vs desktop
2. No of rows in table for mobile vs desktop
3. Which day of the week does wiki get the most traffic?
4. Compare traffic between both mobile and desktop sites by day of the week?

For clickstream data:

1. Top 10 articles requested from Wiki
2. Who sent the most traffic to wiki in Feb'15
3. Top 5 trending articles on twitter in Feb'15
4. Most requested missing pages?
5. What does traffic inflow vs outflow look like for most requested pages?
6. Analyze traffic pattern for a particular article and visualize it.
7. What percent of page visits are from wikipedia itself?
8. How do people arrive at pages for current Presidential Candidates? Is this different from access pattern to current President?

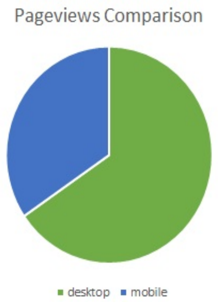
6 Results and Visualizations

6.1 Pageviews

Sum of requests for Mobile & Desktop viewers

```
+--------------------+
|sum(requests)mobile|
+--------------------+
|   4605797962      |
+--------------------+
+---------------------+
|sum(requests)desktop|
+---------------------+
|   8737180972       |
+---------------------+
+----------------------+
|sum(requests) for all|
+----------------------+
|     13342978934     |
+----------------------+
```

Compare Pageviews - Mobile vs Desktop



Analyzing statistics of requests for mobile and desktop

```
For desktop
+------------------+-------------+--------+
|avg(requests)|min(requests)|max(requests)|
+------------------+-------------+--------+
|1279.38      |645          |3292         |
+------------------+-------------+--------+

For mobile
+------------------+-------------+--------+
|avg(requests)|min(requests)|max(requests)|
+------------------+-------------+--------+
|2426.99      |1312         |5695         |
+------------------+-------------+--------+
```

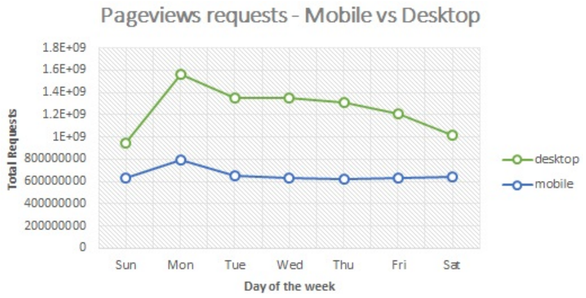
Day of the week with most traffic

```
Overall (Mobile+Desktop)
+---------------+-------------+
|Day of the week|sum(requests)|
+---------------+-------------+
|            Tue|   1995034884|
|            Thu|   1931508977|
|            Sat|   1662762048|
|            Sun|   1576726066|
|            Fri|   1842512718|
|            Mon|   2356818845|
|            Wed|   1977615396|
+---------------+-------------+

For mobile
+---------------+--------------+
|Day of the week|total_requests|
+---------------+--------------+
|Fri            |635169886     |
|Mon            |790026669     |
|Sat            |646334635     |
|Sun            |629556455     |
|Thu            |625338164     |
|Tue            |648087459     |
|Wed            |631284694     |
+---------------+--------------+

For desktop
+---------------+--------------+
|Day of the week|total_requests|
+---------------+--------------+
|Fri            |1207342832    |
|Mon            |1566792176    |
|Sat            |1016427413    |
|Sun            |947169611     |
|Thu            |1306170813    |
|Tue            |1346947425    |
|Wed            |1346330702    |
+---------------+--------------+
```

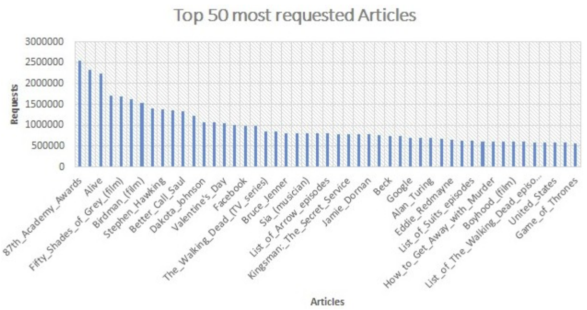
Day of the week with most traffic - Mobile vs Desktop



6.2 Clickstream

Top requested articles

```
+--------------------+---------+
|               Title|  sum(no)|
+--------------------+---------+
|           Main_Page|127500620|
| 87th_Academy_Awards|  2559794|
|Fifty_Shades_of_Grey|  2326175|
|               Alive|  2244781|
|          Chris_Kyle|  1709341|
|Fifty_Shades_of_Grey|  1683892|
|      Deaths_in_2015|  1614577|
|      Birdman_(film)|  1545842|
|Islamic_State_of_...|  1406530|
|     Stephen_Hawking|  1384193|
+--------------------+---------+
```



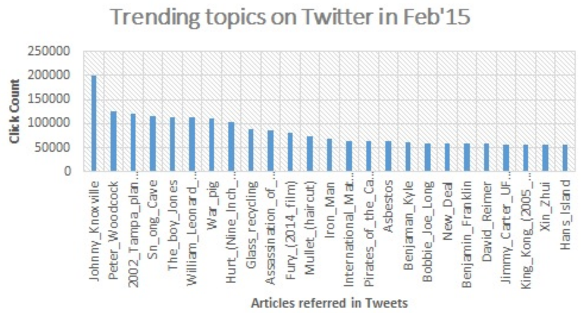
Top Referers

```
+-------------------+----------+
|           Referrer|   sum(no)|
+-------------------+----------+
|       other-google|1496209976|
|        other-empty| 347693595|
|    other-wikipedia| 129772279|
|        other-other|  77569671|
|         other-bing|  65962792|
|        other-yahoo|  48501171|
|          Main_Page|  29923502|
|      other-twitter|  19241298|
|     other-facebook|   2314026|
|87th_Academy_Awards|   1680675|
+-------------------+----------+
```

Trending on Social Media like twitter

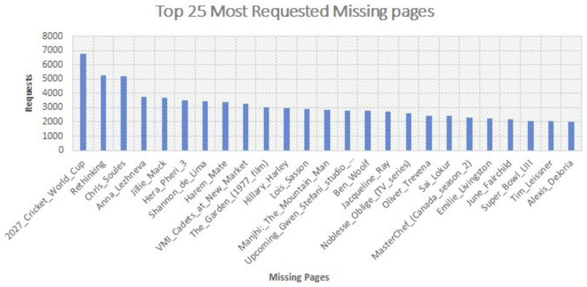
```
+--------------------+-------+
|             Article|sum(no)|
+--------------------+-------+
|     Johnny_Knoxville| 198908|
|       Peter_Woodcock| 126259|
|2002_Tampa_plane_...| 119906|
|         Sn_ong_Cave| 116012|
|        The_boy_Jones| 114401|
+--------------------+-------+
```

Trending on Social Media like twitter



Most requested missing pages

```
+--------------------+-------+
|             Article|sum(no)|
+--------------------+-------+
|2027_Cricket_Worl...|   6782|
|          Rethinking|   5279|
|        Chris_Soules|   5229|
|      Anna_Lezhneva|   3764|
|         Jillie_Mack|   3685|
+--------------------+-------+
```



Inflow vs Outflow for Top 50 Most requested pages.

```
+--------------------+---------+--------+-----+
|Articles            |in_count |out_cnt |ratio|
+--------------------+---------+--------+-----+
|Main_Page           |127500620|29923502|0.234|
|87th_Academy_Awards |2559794  |1680675 |0.656|
|Fifty_Shades_of_Grey|2326175  |1146401 |0.492|
|Alive               |2244781  |3480    |0.001|
|Chris_Kyle          |1709341  |869974  |0.508|
+--------------------+---------+--------+-----+
```

Percentage of traffic flow within Wiki itself Percentage of page visits in Wikipedia from other pages in Wikipedia itself: 6.615%

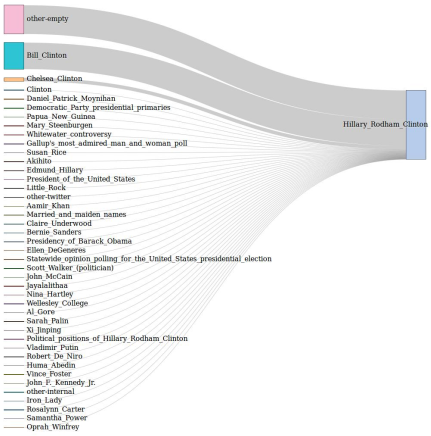
Top referrers to Donald Trump

```
--------------------------------------------+------+
Referrer                                    |typeOf|
--------------------------------------------+------+
The_Apprentice_(U.S._season_14)             |link  |
United_States_presidential_election_2016|link  |
Bill_Rancic                                 |link  |
Roast_(comedy)                              |link  |
Geraldo_Rivera                              |other |
Steve_Rubell                                |link  |
Jamaica,_Queens                             |link  |
Eric_Trump                                  |link  |
List_of_people_in_Playboy_200009            |link  |
other-empty                                 |other |
--------------------------------------------+------+
```

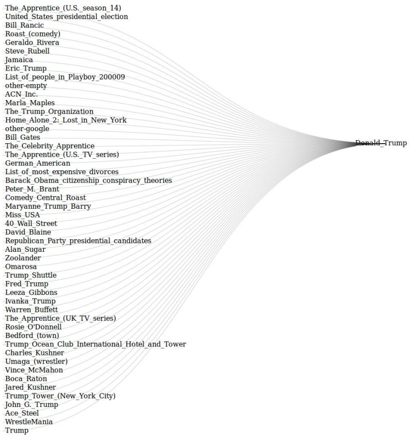
Top referrers to all presidential candidate pages

```
----------------------+-------------------------+
Article               |Referrer                 |
----------------------+-------------------------+
Hillary_Rodham_Clinton|Clinton                  |
Hillary_Rodham_Clinton|Daniel_Patrick_Moynihan  |
Ted_Cruz              |Kay_Bailey_Hutchison     |
Donald_Trump          |The_Apprentice           |
                      |  _(U.S._season_14)      |
Hillary_Rodham_Clinton|Chelsea_Clinton          |
Bernie_Sanders        |Independent_politician   |
Donald_Trump          |United_States_presidential|
                      | _election_2016          |
Bernie_Sanders        |Democratic_socialism     |
Hillary_Rodham_Clinton|Papua_New_Guinea         |
----------------------+-------------------------+
```

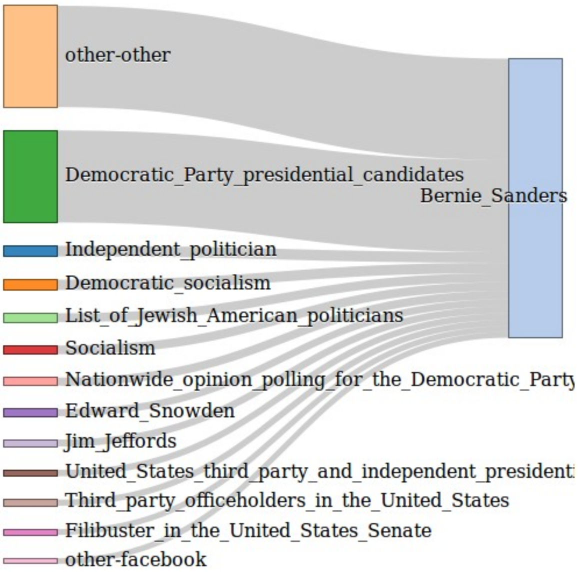
Top Referrers to Hillary Clinton



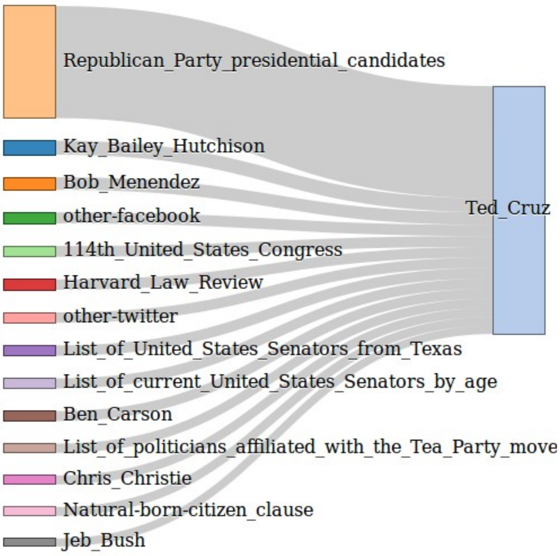
Top Referrers to Donald Trump



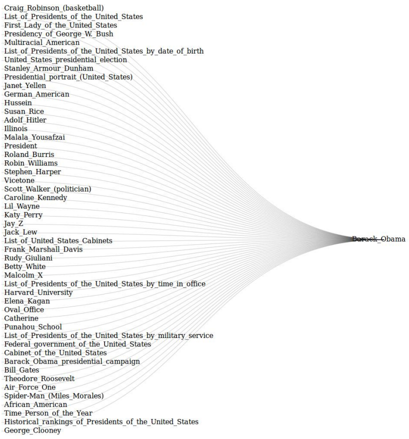
Top Referrers to Bernie Sanders



Top Referrers to Ted Cruz



Top Referrers to Obama



7 Findings

For Pageviews Data

1. Dataset contains 2 rows for every second (one for mobile and one for desktop) which we verify by ordering timestamp after changing its datatype from String to timestamp.
2. **curr_id:** the unique MediaWiki page ID of the article the client requested
3. Caching data and query results leads to faster execution of queries.
4. As we have used both SparkSQL and spark commands to run queries we have found that SparkSQL supports a lot of SQL functionality and is mre intuitive but does not support UPDATE or DELETE as it is typically used for batch analysis of data.
5. Mobile requests were less as compared to desktop.

For Clickstream Data:

1. The most requested pages are about the media(song/video/movie/series) that were popular in February 2015 with very few exceptions.
2. The top Referer is Google by a large margin. Next is the refererless traffic ie., other-empty (usually HTTP clients). The third largest sender of traffic to Wiki are Wikipedia pages that are not in the main namespace.
3. When clients went to the ”Alive” article, almost nobody clicked any links from this article to go to another article, but 49.2% of people who went to ”Fifty Shades of Grey” article and 65.6% of people who went to ”87th Academy Awards”, clicked on another link in article and continued to browse wikipedia.
4. Analyzing presidential candidates:
   (a) The top referrer for Donald Trump is The_Apprentice show where he was the executive producer and host.
   (b) The top referrers for Hillary Clinton page are HTTP clients and Bill Clinton wiki page.
   (c) The top referrer for Bernie Sanders is the page for Democratic Party presidential candidates.
   (d) The top referrer for Ted Cruz is the page for Republican Party presidential candidates.
5. Comparing the Presidential Candidates' Ref to current President -
   Top referrers are : List of other presidential candidates, the first lady, President George Bush, Multiracial American The referrers are quite different for President vs Presidential Candidates.

References

https://spark.apache.org/docs/1.5.1/

http://docs.ansible.com/ansible/index.html

https://d3js.org/.

https://datahub.io/dataset/wikipedia-clickstream

https://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream

https://github.com/futuresystems/big-data-stack/

http://bdossp-spring2016.readthedocs.io/en/latest/projects.html

https://datahub.io/dataset/