Sequential Analysis

140

Process part of a large-scale using sequential programs on a cloud sequentially

Filtering an Hour's worth of Data

In this part, we will analyze a single hour's worth of data from the logs and find out the most popular english wikipedia articles. If you are working with EC2 instances for the first time, please refer to the Amazon EC2 page in the Project Primer and the Amazon EC2 Getting Started Guide for help.

1. Provision a **`t1.micro`** or **`m1.small`** instance using the AMI ID: **`ami-ed30ba84`**
2. **Note**: For this project, assign the tag with Key: **`Project`** and Value: **`1.1`** for all resources

Download a single hour's worth of logs from

`s3://wikipediatraf/201306-gz/pagecounts-20130601-000000.gz`

using **`s3cmd`**

The file contains all of the pages from all WikiMedia projects. Our aim is to identify trending topics from the English Wikipedia articles. In order to do this, develop a script or write a program in any language to:

1. Filter out all pages that are not english wikipedia. (This means that the log lines should start with **`en`**, without any suffix attached).
2. There are many special pages in wikipedia that do not need to be considered when trying to find trending topics. Exclude any pages whose title starts with the following strings:

```
Media:
Special:
Talk:
User:
User_talk:
Project:
Project_talk:
File:
File_talk:
MediaWiki:
MediaWiki_talk:
Template:
Template_talk:
Help:
Help_talk:
Category:
Category_talk:
Portal:
Wikipedia:
Wikipedia_talk:
```

3. Wikipedia policy states that all English articles must start with an uppercase character. Filter out all articles that start with lowercase English characters. You may notice that some articles have non-english

titles, you should choose to retain them in the analysis.

4. You may also get results which refer to image files, exclude any article that ends with the following extensions (Keep all other extensions intact). (**`.jpg, .gif, .png, .JPG, .GIF, .PNG, .txt, .ico`**)
5. Finally, there are some boilerplate articles which are returned by Mediawiki, which should be excluded as well. Articles with titles that exactly (case sensitive) match any of the following strings should be excluded:

```
404_error/
Main_Page
Hypertext_Transfer_Protocol
Favicon.ico
Search
```

6. Once the filtering is done, output the remaining articles in the following format:

```
<article name>\t<page views>
```

(Note that the **`\t`** stands for the tab character) The output should sorted in the order of number of page views.

When you are done, complete the following checkpoint quiz:

checkpoint

Sequential Analysis