TorStar - Data Engineering Assignment

Kenneth Blahut

January 12, 2020

1 Using the Code

The code used to complete the assignment is 'TorStar_assignment.py'. This code is used to answer all questions. To run the code, open a Terminal or PowerShell and direct yourself to the directory containing the model code. Then type the following command :

```
python TorStar_assignment.py #
```

where the '#' can be replaced with '1', '2', '2a', '3' or '4' for each question of the assignment.

2 Solution Overview

2.1 Question 1

Using the code provided, the number of posts in June 2016 is **30727**.

To get this result using the code simply write:

```
python TorStar_assignment.py 1
```

into the command line of PowerShell, from the directory where the code is saved.

2.2 Question 2

The number of posts containing the tag 'combinatorics' is **33394**. From these posts, there are **33270** posts which do not contain the tag 'fibonacci-numbers'.

To get the results for each of these questions write:

```
python TorStar_assignment.py 2
python TorStar_assignment.py 2a
```

respectively into the command line of PowerShell, from the directory where the code is saved.

2.3 Question 3

Figure 1 shows the distribution of posts containing the tag “graph-theory” sorted by month. All posts were submitted between July 20, 2010 and June 3, 2018. This means that both July and June may be slightly under represented since they each included partial months. The most popular month was November with 1483 posts, followed closely by April with 1444 posts. These two months coincide with end of term final exams for a typical school year in Canada, and could explain why these months are so popular.

To run the code for this solution type:

```
python TorStar_assignment.py 3
```

into the command line of PowerShell, from the directory where the code is saved. The code will save the histogram as “graphtheory_hist.pdf” in the same directory as the code.

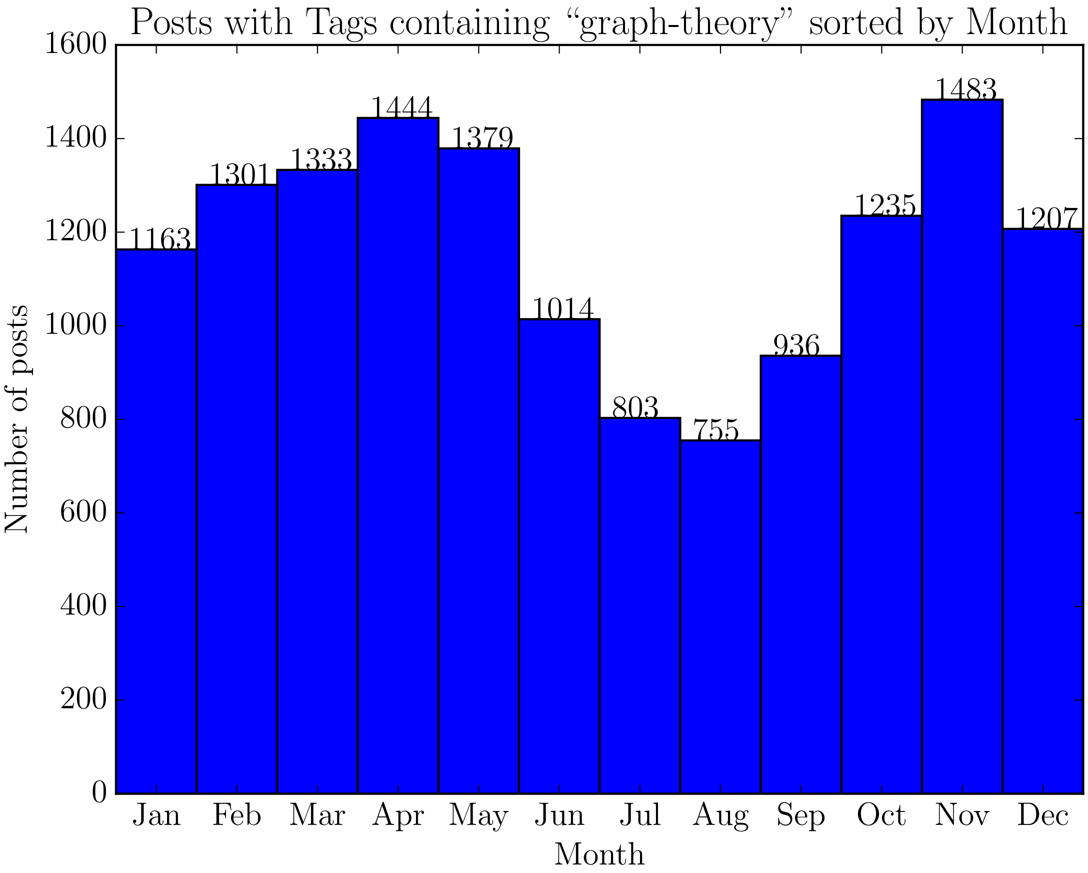

Figure 1: Number of posts containing the tag “graph-theory” sorted by month. The most popular month is November followed closely by April. This could be due to final exam being studied for during these months.

2.4 Question 4

Similar to the previous section, here I show how the number of posts containing the tag "graph-theory" increases over time (Figure 2).

Again, to run the code for this solution type:

```
python TorStar_assignment.py 4
```

into the command line of PowerShell, from the directory where the code is saved. The code will save the time series plot as "graphtheory_timeseries.pdf" in the same directory as the code.

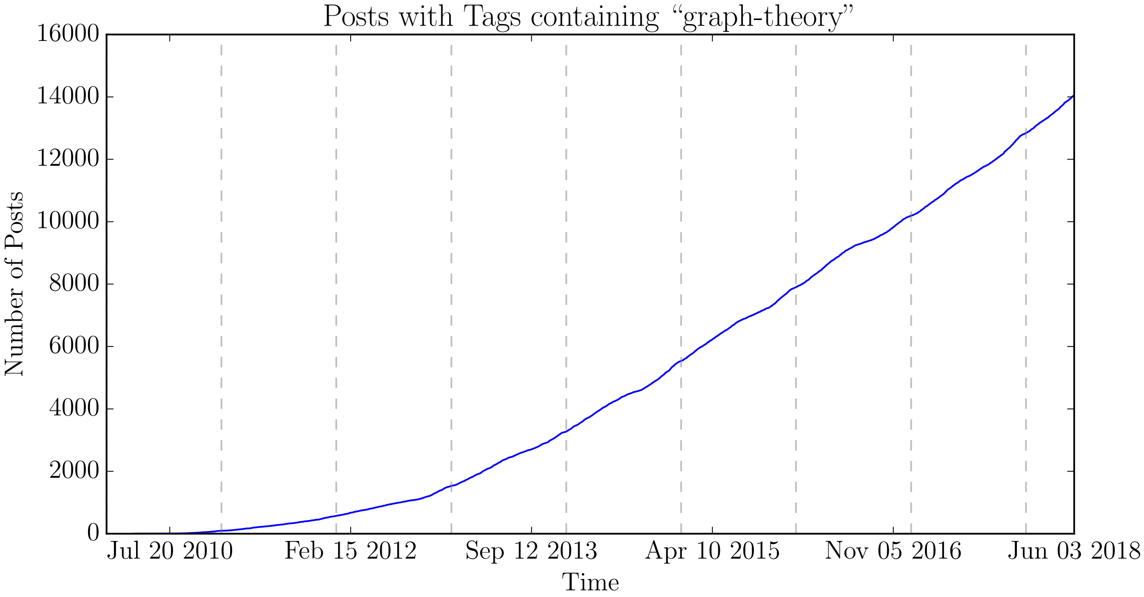

Figure 2: Number of posts containing the tag "graph-theory" over time. The vertical dashed grey lines represent the start of a new calendar year.

3 TorStar – Data Engineering Assignment

3.1 Overview

The goal here is to get an idea of your technical skills and what it would be like to work with you everyday. Once you finish the assignment we'd like you to come in and basically do a code review with our team. The evaluation process will be like a code/solution review: design choices, implementation, reusability of code, etc. As long as you can justify your choices, pretty much anything goes. Please give us a time estimate for this as soon as you can so we can proceed with further scheduling.

3.2 Tech Stack

The only requirement from our side is that you use one of the following languages: Java, Python, Scala or SQL. Bonus points for using MapReduce or Spark. If you don't want to use ANY of the languages, we'd like you to use Talend Open Studio. You can choose to use any/no other tools as you see fit. Please keep scale in mind while developing the solution. While this is a toy problem in an interview setting, we'd like to see how you would approach something we would want as part of our production pipelines.

3.3 Requirements

We've provided a copy of the Posts.xml file (2.5G) from Math Stackexchange[1] dated June 5th (the entire archive is available here) . It contains approximately 2.3 million posts from math.stackexchange.com. We would like you to process this data set and answer the following questions:

1. Count the number of posts made in June 2016.
2. Calculate the number of posts tagged with "combinatorics".
   (a) For all posts tagged with "combinatorics" find the count of those not tagged with "fibonacci-numbers".
3. Determine the month that was most popular for posts tagged with "graph-theory"
4. Create a timeseries plot of the number of "graph-theory" posts across all time captured in the dataset by quarter.

3.4 Deliverables

Please share your solution with us as an archive file via your favorite file sharing cloud or as a public code repo (GitHub, BitBucket etc.). You can continue to make changes to your code after you submit if you like. For the in-person review process, please bring your own laptop and walk us through your design process and solution live. If you would like to use one of our laptops, please provide us with all steps required to build and run your solution on a Windows 10 machine.

[1] Stack Exchange Creative Commons data now hosted by the Internet Archive ”All community-contributed content on Stack Exchange is licensed under the Creative Commons BY-SA 3.0 license. As part of our commitment to that, we release a quarterly dump of all user-contributed data (after carefully sanitizing it to protect user private data, of course).” https://archive.org/details/stackexchange