INFO 3501/5501: Lab Assignment #4

Due on Wednesday, November 2, 2016 at 9:00am

Professor Brian Keegan
Department of Information Science
College of Media, Communication, and Information
University of Colorado, Boulder

Background & Objectives

We build on the results from all three previous labs to look at attention dynamics in Wikipedia. Lab Assignment #4 will explore different approaches to identify and quantitatively analyze changes in pageviews to Wikipedia articles over time.

1. **Combining data retrieved via API and MySQL**. This lab will continue to demonstrate the capabilities of using the combination of MySQL and API queries to analyze relational and temporal data in Wikipedia. Students will extend functions from previous labs to retrieve data from the Wikimedia pageview API and combine it with data sources and types from previous labs.
2. **Interpreting pageview data**. Wikipedia pageviews provide an alternative measure of user behavior over time to triangulate with editor revision data (from Lab Assignment #1). Students will analyze time series pageview data for a single article as well as compare the pageview behavior of multiple articles.
3. **Integrating pageview data with hyperlink and revision data**. Using concepts from previous labs, students will analyze how pageview behavior is correlated with the pageview behavior of neighboring pages as well as how revision activity and pageviews are correlated with each other.

Instructions

We will be be using Jupyter Notebooks running on the Wikimedia Foundation's PAWS infrastructure to retrieve public data from replica MySQL databases containing the complete history of revisions to articles, data exposed by the the Wikipedia API, as well as data from the Wikipedia Pageview API. We will use the account you created in Lab 1 as well as code from Lab Assignments #1 through 3.

Additional Resources

Here are some links to other resources that provide more details if you are curious and want to learn more.

Pageview API Official documentation about how to access different API endpoints for the Wikimedia Pageview API: `https://wikimedia.org/api/rest_v1/`

Pageview API examples Additional examples of how to make different API calls. `https://wikitech.wikimedia.org/wiki/Analytics/PageviewAPI`

Problem 1

The Wikimedia Pageview API only returns data from approximately July 2015 onward (there are other data sources that are more difficult to use to get different pageview data as far back as January 2008). What is an example of a Wikipedia article that might have unusual information consumption patterns like a brief burst of attention, sustained increase in popular interest, *etc*. since July 2015? What specific events might drive attention to this article or related topics? What kinds of information would people be seeking about the topic from the Wikipedia article? Where else might they find this information instead and how might this competition for attention affect the kinds of conclusions we can draw from this Wikipedia data?

Problem 2

This problem will use the "Interpret page view results" section of the notebook. Does the linear (`logy=False`) or logarithmic (`logy=True`) plot of pageviews for a single article do a better job of capturing the attention dynamics? What are strengths and weaknesses in each plot? Include a plot from the notebook and describe potential explanations for what causes different events in the graph. What days have the largest outliers in pageviews and are these surprising or to be expected (change the `std_threshold` if there are too many or too few results)? How much of the total pageview activity occurred during these peak days? Why might this be important or troubling? Are there particular days of the week where the median attention to this topic is higher? Speculate what might be causing users' information seeking to vary over the course of the week to explain the observed patterns.

Problem 3

This problem will use the "Compare pageviews to another page" section of the notebook. What is another page to which you can compare the pageview behavior of your previous article? What similarities or difference would you expect in the information seeking/consumption behavior for each article? Include a plot comparing the pageview data of both articles and narrate some of the salient dynamics around major events, spikes in attention that are in one but not both, the relative amount of attention, *etc*. Discuss the correlation coefficient between these two time series: how strongly correlated are these two articles and is this higher or lower than you expected? Choose a third and ideally totally unrelated article to compare against your initial page. Is the correlation between this third random article and your initial page higher or lower than you expected? Should we be concerned about the other results in this analysis if even two random articles show high levels of correlated pageview activity?

Problem 4

This problem will use the "Get the pageviews for the hyperlink network", "Most and least correlated articles", and "Pageview correlation and link location" sections of the notebook. What articles in your initial page's hyperlink network have the most total attention? Are any of these surprising and what did you expect to see but is missing in the top-10 list? What does their ranking tell us about attention to different topics on Wikipedia? Which articles within the hyperlink network show the strongest and weakest correlations in attention? What is interesting or unsurprising about the topics that show strongly correlated or anti-correlated attention? Include a plot showing the most correlated and most anti-correlated topics and narrate some of the interesting features in each. Is there a strong relationship between where a link appears in the the article and how well-correlated the pageview behavior is? Discuss what might be contributing to the observed presence or absence of a relationship between link position and pageview correlation.

Problem 5

This problem will use the "Get page revisions" and "Are pageviews and edits correlated with each other?" sections of the notebook. Why might we expect pageviews and editing activity to be correlated with each other? How strong is the correlation between pageviews and editing? How does it compare to other correlations you've come across in the analysis so far? Include a plot visualizing the number of daily edits and pageviews for the article and narrate any interesting spikes in attention or editing that are or are not correlated with each other. Include and discuss the chart that visualizes the cumulative pageviews per edit: what does this ratio substantively capture about the relationship between pageviews and edits over time? Do the spikes in attention on specific days you observed in prior analyses result in positive or negative changes to this ratio? Include the chart showing the scatterplot trajectory of normalized revision and pageview activity. Discuss whether the evidence in the trajectory graph provides additional support for your previous findings about whether Wikipedia can meet the demand for information.

Problem 6

Required for graduate students, extra credit for undergraduates. This problem will use the "Did a burst of pageviews diffuse to adjacent pages?" and the "Did this burst of pageviews translate into edits on these pages as well?" sections of the notebook. Speculate about potential behavioral mechanisms that drive bursts of attention on one article to spread to some — but not all — adjacent articles in the hyperlink network. How many articles in the hyperlink network have their maximum pageviews on the same date as the focal article? Is this number higher or lower than you would expect? What does the list of "co-bursting" pages reveal about the similarity between these topics and are there any outwardly spurious co-bursters? Is the list of co-bursting pages by edits larger or smaller than the co-bursting pages by pageviews? What topics are present on one but not the other and what do these overlaps or exclusions reveal? How many edits were created "as a result" (establishing causality is hard!) of the burst in pageviews? What are some additional analyses you might do to understand the consequences of bursts of activity on Wikipedia content?

Acknowledgements

I want to thank the Wikimedia Foundation for the PAWS system and related Wikitech infrastructure that this workbook runs within. Yuvi Panda, Aaron Halfaker, Jonathan Morgan, and Dario Taraborelli have all provided crucial support and feedback.

Licensing