EXPLORING THE STACK OVERFLOW API IN 3 WORKBOOKS

adam hogan
@ahogy

Presentation for the Chicago R User Group

Chicago, IL
August 20, 2019

OVERVIEW

- Stack Overflow has made a lot of data available to us:
  - Queries:
    https://data.stackexchange.com/stackoverflow/queries
  - Full data:

    https://archive.org/details/stackexchange
- We're going to do use R to do three explorations:
  - Queries and counts over time
  - Social Network Analysis
  - Natural Language Processing

FORMAT

- We will
  - Introduce a topic and the R tools to interrogate it.
  - Obtain the data.
  - Work through a notebook.
  - Ask questions motivating the group workshop session for going deeper.
  - There won't be "answers" on this, but should get you started.
- Hashtag for this event:
  - #RstackoverflowOURstackoverflow

MOTIVATION

- These are the questions and answers of our community.
- SO has made the data available, and we have the responsibility and the tools to make sense of it, to improve our community.
- Also, it's interesting data!

QUERIES OVER TIME

Let's get that data.

OVERVIEW

- We start simple, counting. Here, we examine tags associated with questions, and see how R is faring over time.
- Let's query the data, and plot tags for top statistical languages/packages/frameworks.
- We'll introduce xts for time series handling, plotting, and the difference operator for rendering a series stationary, which is important in forecasting.

QUERY

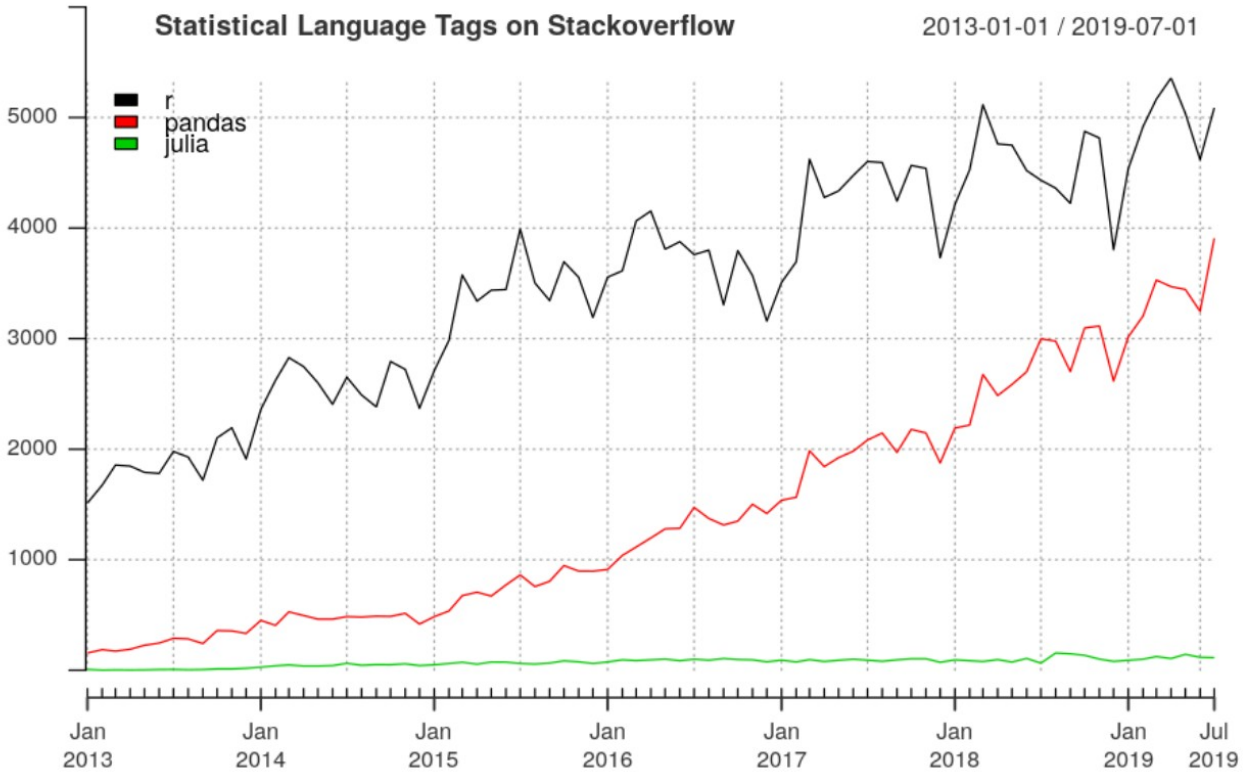
- Query is here:

  https://data.stackexchange.com/stackoverflow/query/1092906/r-v-pandas-v-julia
- Explanation:
  - Give me the posts where post-tags have names in R, Pandas, or Julia, and let's count those up by month.
- Let's open the notebook:
  - https://d-and-a-public.s3.amazonaws.com/counts.ipynb

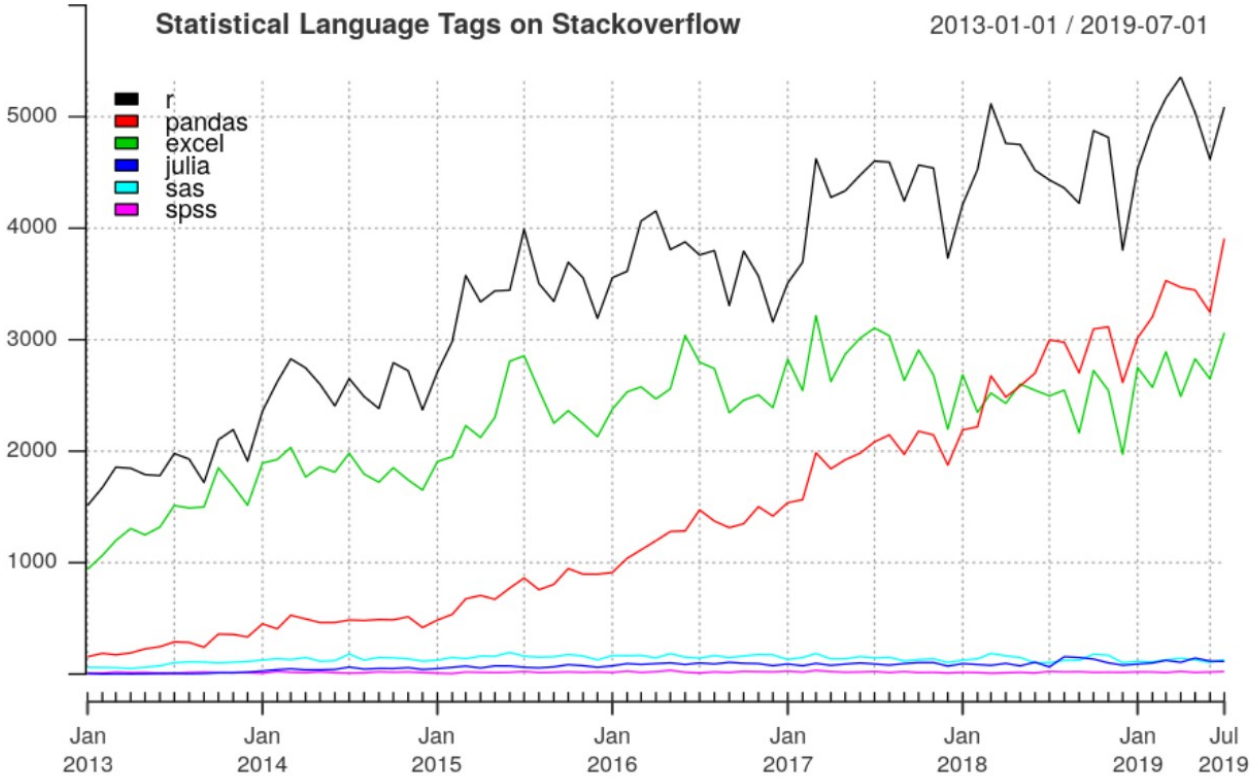
RESULTS: 1

	month	r	pandas	julia
	<date>	<int>	<int>	<int>
4	2019-07-01	5088	3908	114
7	2019-06-01	4619	3247	117
10	2019-05-01	5038	3445	144
13	2019-04-01	5358	3470	106
16	2019-03-01	5166	3530	124
19	2019-02-01	4919	3205	100

R is still #1 for now.



RESULTS: 2



Pandas passed excel!

	month	r	pandas	excel	julia	sas	spss
	<date>	<int>	<int>	<int>	<int>	<int>	<int>
7	2019-07-01	5088	3908	3063	114	129	24
13	2019-06-01	4619	3247	2649	117	106	19
19	2019-05-01	5038	3445	2828	144	128	15
25	2019-04-01	5358	3470	2491	106	142	25
31	2019-03-01	5166	3530	2892	124	127	14
37	2019-02-01	4919	3205	2573	100	110	19

WORKSHOP QUESTIONS

- How would you change the queries we introduced to study what share of R questions are about ggplot? xts? data.table? Tidyverse? Are these changing over time or stable?
- Modeling question: is having more questions actually a good proxy for popularity? A sign of more beginners? What is valuable to model?
- How would you forecast the future of R?
- On what date will pandas overtake R with 95% confidence? Julia? (This is tricky)

SOCIAL NETWORK ANALYSIS

Let's see who we are.

OVERVIEW

- Social Network Analysis uses the connections between agents to model their social system using the tools of graph theory.
- Let's ask two questions:
  - Knowing only that users engaged with the same questions, but not the scores they got for engagement, can we back out actual Stack Overflow reputations?
  - Are there quantitatively discernible communities among the R Stack Overflow network?

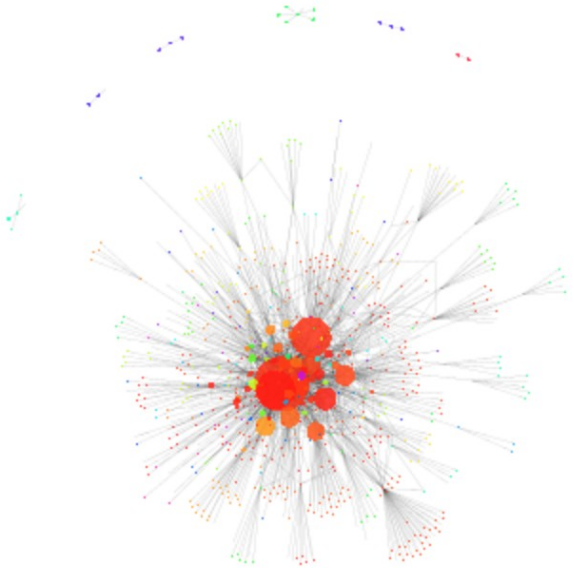
QUERY

- Query is here:

  https://data.stackexchange.com/stackoverflow/query/1092909/questions-and-answers-usernames-and-dates-for-r-tagged-questions

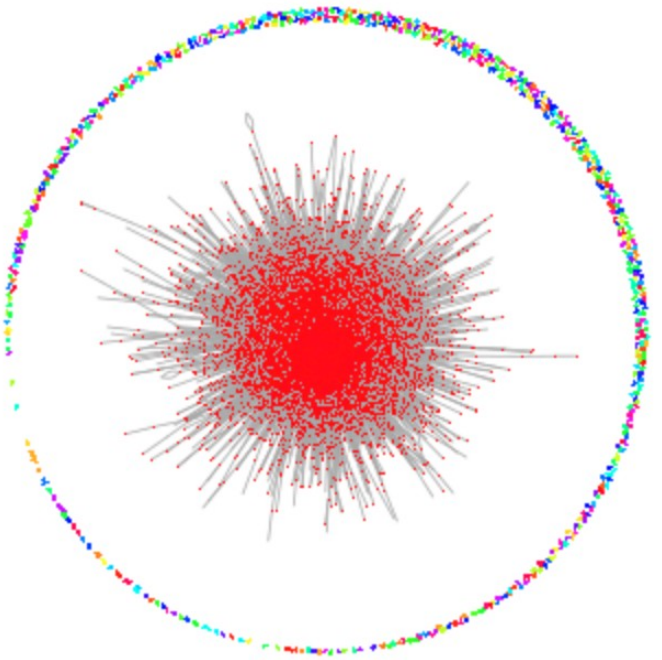
- Explanation:
  - For answered questions, give me (a) the question, (b) the selected answer, (c) the non-selected answers, (d) the basic ids, (e) scores of each, (f) and their authors---for things tagged with R, ordered by creation date
- Let's open the notebook:
  - https://d-and-a-public.s3.amazonaws.com/graph.ipynb

RESULTS



2009

2019



WORKSHOP QUESTIONS

- Lookup your username. What R community do you belong to?
- What additional data could improve a measure of authority over eigenvector centrality or PageRank?
- Are the communities meaningful? Do they represent users who frequently answer the same tags? How would you find out?
- How have the communities changed over time?

NATURAL LANGUAGE PROCESSING

Let’s see what we say.

OVERVIEW

- There are a lot of ways to analyze text. We're going to perform sentiment analysis using the Stanford coreNLP library, and compare it to Facebook's fasttext pre-trained word-vectors.
- n.b., "sentiment" isn't as popular a research model as it once was. The new hipness is not a single abstract representation of hot or cold, but whether you can predict a particular variable of interest: eg, propensity-to-buy, is-this-a-hot-dog-conversation, is-this-fake-news?

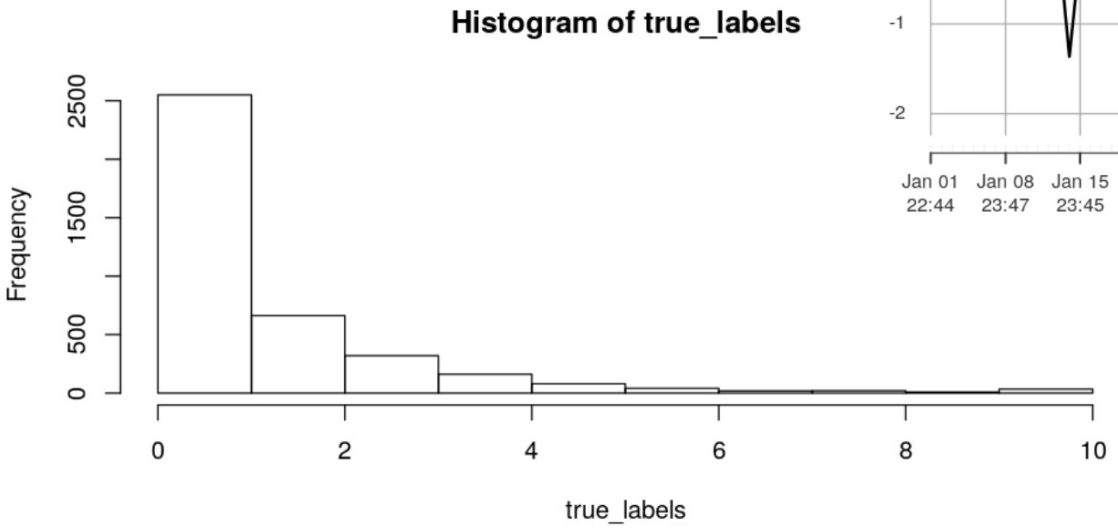
QUERY

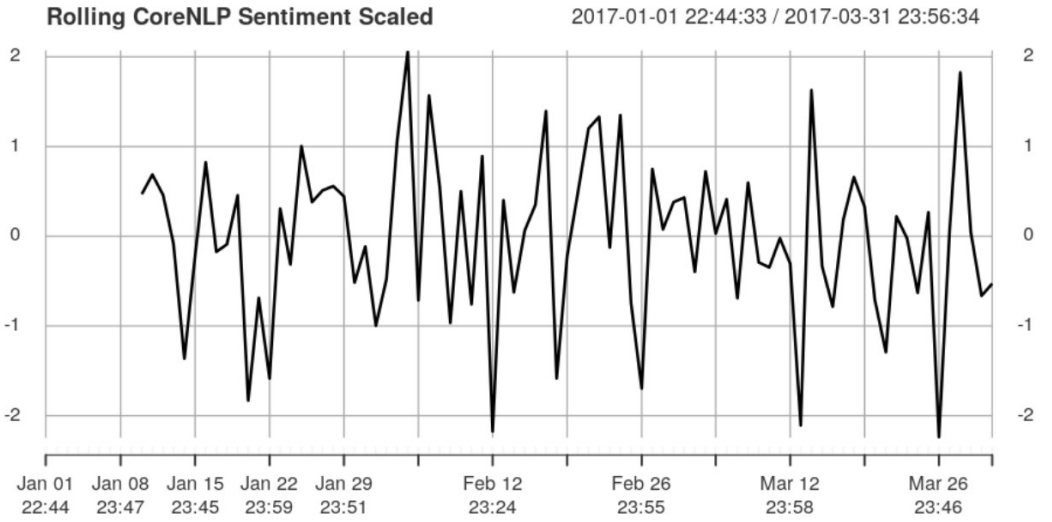
- Query is here:

  https://data.stackexchange.com/stackoverflow/query/1093476/questions-answers-usernames-dates-and-full-text-for-r-tagged-questions
- Explanation:
  - For answered questions, give me data about the question, the question-asker, the answers, the basic ids, scores of each and their authors, and the full text---for things tagged with R, ordered by creation date
- Let's open the notebook:
  - https://d-and-a-public.s3.amazonaws.com/nlp.ipynb

RESULTS

Predict votes from text





Changes in sentiment

WORKSHOP QUESTIONS

- Let's improve the sentiment prediction by finishing the fasttext training.
- What other data could we use to make an even better supervised model for "sentiment?"
- Are people getting "nicer" over time?
- Recent research Stack Overflow released (Aug 14) doing NLP on this data:
  - https://stackoverflow.blog/2019/08/14/crokage-a-new-way-to-search-stack-overflow/
- For all the marbles, can you make a generative model for answers based on Stack Overflow questions?

QUESTIONS?

Thank you!

(I think I know where they answer those)

adam hogan
@ahogy