smoke test

francesca giannetti

8/16/2017

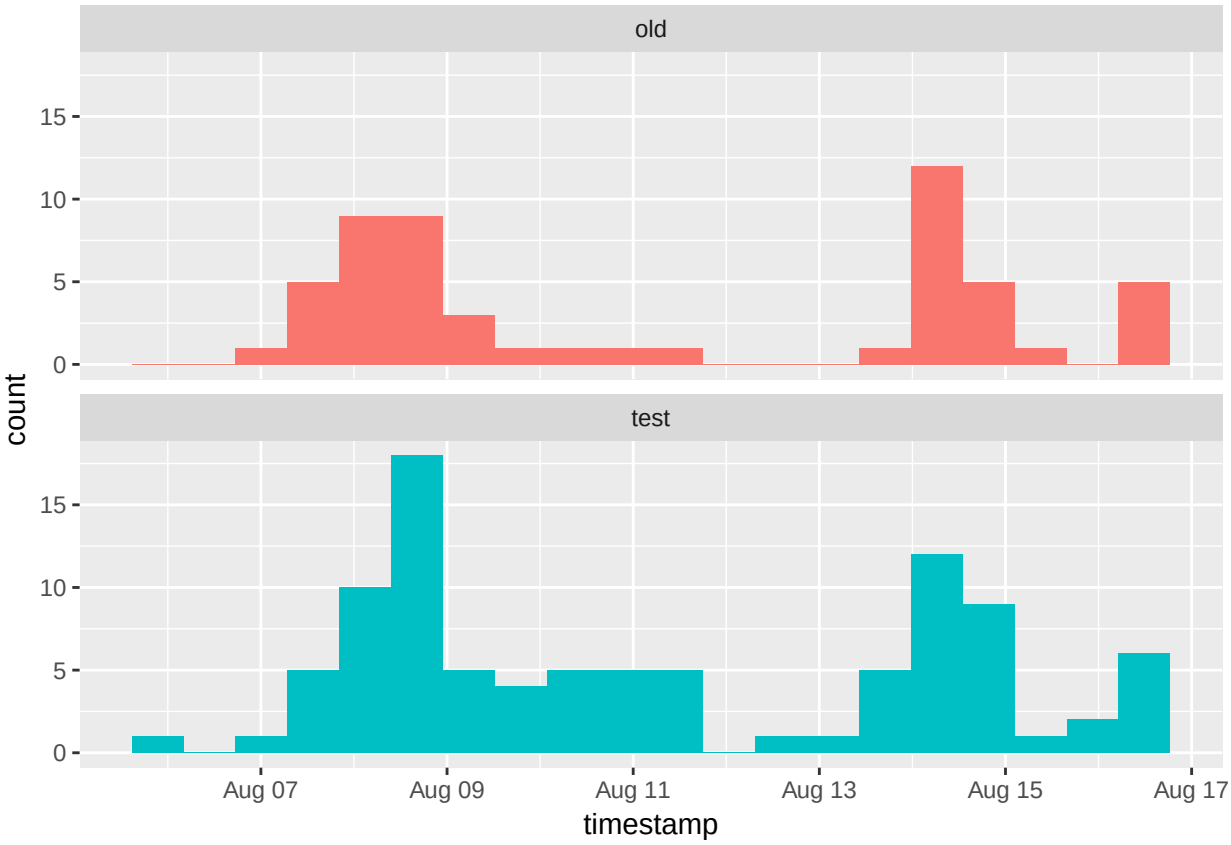
About

The purpose of this file is: 1. evaluate how much data was lost by not using the base URL of British Library Sounds as one of the search strings in the TAGS tool; 2. determine if the lost data is significant enough to skew analysis results.

Code adapted from: Silge, J. and Robinson, D. "Case study: comparing Twitter archives" *Text Mining with R*.

Histogram

The tweet pattern for the old and test datasets is similar but reveals that there is some missing data in the old dataset.

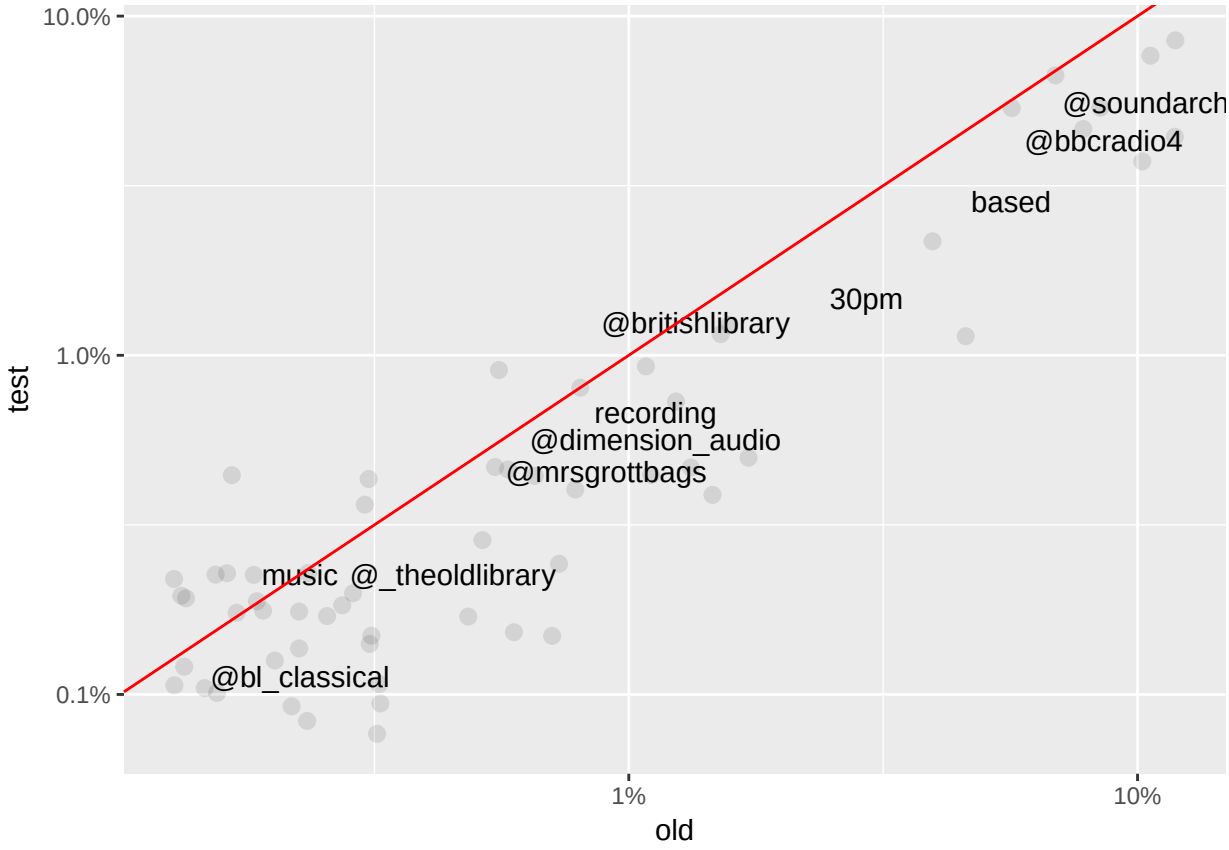


Word Frequencies

Use tidytext to process the text of the tweets in both datasets.

Frequency Plot

All terms hover close to the reference line, indicating that both the old and test datasets use the same words at the higher frequencies.



Comparing word usage

These are the words that are likely to come from both the old and test datasets.

word	old	test	logratio
breakwell	0.0021692	0.0095109	1.478078
ian	0.0021692	0.0095109	1.478078
minutes	0.0021692	0.0095109	1.478078
website	0.0021692	0.0095109	1.478078
cats	0.0021692	0.0108696	1.611609
sounds	0.0021692	0.0108696	1.611609
#internationalcatday	0.0021692	0.0122283	1.729393
history	0.0021692	0.0149457	1.930063
oral	0.0021692	0.0149457	1.930063
interview	0.0021692	0.0163043	2.017075
listen	0.0021692	0.0163043	2.017075

Distinctive Words

The interesting thing one observes with the most distinctive words test is that there are words that are showing up in the old dataset that are **not present** in the test dataset, even though the test dataset was created using a query that contains the old query. Given that data is missing from both datasets, this reinforces the point that Search API provides a *sample* of tweets and not a complete dataset for the period examined.