Learning to effectively model tweets containing “humanities.”

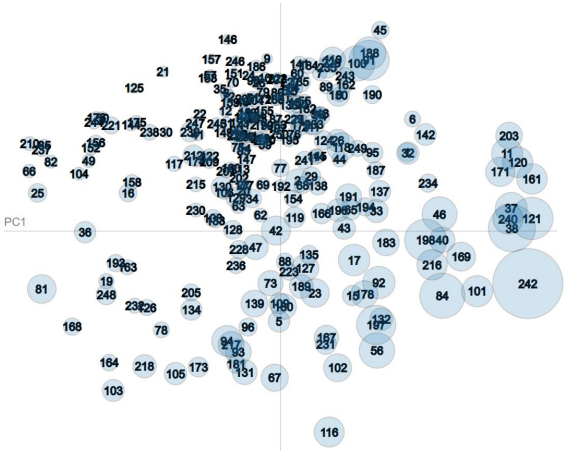
After generating an initial model of our corpus of Twitter posts during 2014-2017 containing the word “humanities,” we encountered issues specific to such a Twitter corpus. These issues pertain to hashtags, mentions, links, and many abbreviations not commonly used in the news and other documents that comprise the main WE1S corpus. Due to Twitter’s 140-240 character limits, we decided that mentions, hashtags, and abbreviations that carry rich semantic meaning (e.g., *yolo*--”you only live once”) should be included in our models.

We thus modeled our corpus in that state as a collection of hundreds of thousands of documents, each containing an individual tweet. However, this very high number of documents created problems for our visualization tools. Most of our own researchers were not even able to load the model on their computers!

We thus refashioned our Twitter corpus in two ways and generated new topic models. First we omitted authors who tweeted only once in our corpus and created our Collection 28 (C-28). Topic models for this collection treated each tweet as a single document. The simple act of excluding single-tweet authors significantly improved usability, as it reduced the number of documents by 26%. Upon inspection, we found our models for C-28 were most coherent in the range of 150 to 250 topics.

Secondly, we aggregated all tweets by individual authors and created our Collection 29 (C-29), which treats everything an author had posted as though it were a single “document.” This resulted in an 88% reduction in documents. However, aggregation also resulted in documents of wildly varying sizes (many authors had only two tweets, while others had hundreds), which may skew the topic models of C-29.

Weighing the trade-offs between C-28 and C-29 based on the coherence of their topic models, we presently think that our best method for modeling our Twitter corpus is to treat the corpus as in C-28 where each tweet is a single document. The number of documents is higher in a way that impacts useability (though not as much as if we had retained single-tweet authors). But the increased coherence of models repays the useability cost.



Tweets By Author in C-28 (250 topics), viewed in pyLDavis

Resources

WE1S Twitter Collections: C-28 (models), C-29 (models)

Related Articles: Asbjørn Steinskog, et al., “Twitter Topic Modeling by Tweet Aggregation” (2017)