Topic Modeling in Twitter: Aggregating Tweets by Conversations

David Alvarez-Melis, MIT CSAIL & Martin Saveski, MIT Media Lab

Massachusetts Institute of Technology

ABSTRACT

We propose a new pooling technique for topic modeling in Twitter, which groups together tweets occurring in the same user-to-user conversation. Under this scheme, tweets and their replies are aggregated into a single document and the users who posted them are considered co-authors. To compare this new scheme we train topic models using Latent Dirichlet Allocation (LDA) and the Author-Topic Model (ATM) on datasets consisting of tweets pooled according to the different methods. We experimentally show that it outperforms other pooling methods in both clustering quality and document retrieval.

TWEET-POOLING SCHEMES

Key issue: Tweets are too short to compute robust per-document term co-occurrence statistics and generating coherent topic models is hard.

Solution: Tweet-pooling, merging related tweets to obtain longer documents.

Appealing since we can use off-the-shelf topic modeling toolkits.

Existing pooling schemes

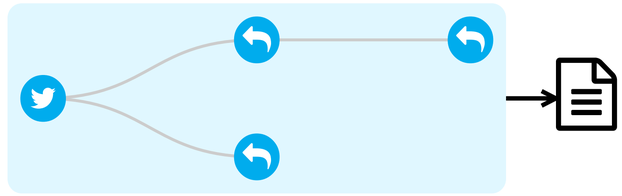
Tweet-pooling	Document = tweet (Baseline)
User-pooling	Document = all tweets posted by a single user
Hashtag-pooling	Document = all tweets that contain a certain hashtag.

DATA COLLECTION

We start with a set of 14 topics and for each topic we select the top 25 most influential users (according to wefollow.com). We retrieve all public tweets posted by these users as well as all tweets that mention them during a period of one week in April 2014.

14	350	101K
Categories	Seed Users	Tweets

POOLING BY CONVERSATIONS



A document consists of a seed tweet, the tweets written in reply to it, tweets written in reply to these and so on.

Motivation: User-to-user interaction in Twitter tends to be around related topics, so pooling tweets by conversations can lead to a more coherent document aggregation and more relevant topic extraction.

CLUSTERING EVALUATION

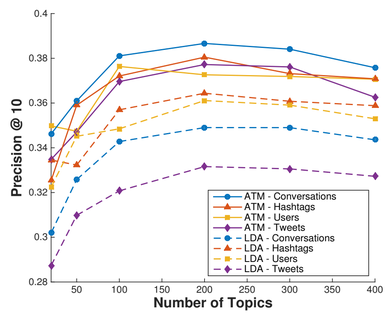
We cluster the test tweets by assigning the predicted most likely topic and compare these clusters to the underlying (noisy) categories.

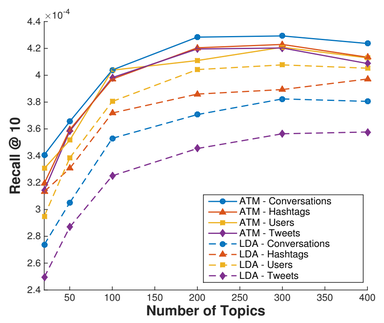
		(Scale: $\times 10^2$)	T = 20	T = 50	T = 100	T = 200	T = 300	T = 400
NMI	ATM	Conversations	**7.819***	7.420	**8.021**	**8.502***	**9.489***	**9.967***
		Hashtags	6.668	**7.810***	7.850	8.230	8.846	9.835
		Users	7.500	6.812	7.968	7.769	9.200	9.842
		Tweets	7.031	6.819	7.456	8.445	9.213	9.595
	LDA	Conversations	4.998	5.937	6.532	7.276	8.066	8.837
		Hashtags	6.559	5.995	7.011	7.532	8.595	9.270
		Users	7.537	6.983	6.933	7.734	8.399	9.313
		Tweets	4.562	5.363	5.946	6.636	7.513	8.201
Adj. Rand Index	ATM	Conversations	**4.514**	1.828	1.170	**0.654***	**0.542***	0.390
		Hashtags	2.875	**1.965***	1.173	0.583	0.406	0.376
		Users	4.407	1.702	**1.195**	0.517	0.478	**0.399***
		Tweets	3.637	1.549	0.973	0.628	0.464	0.351
	LDA	Conversations	2.551	1.469	0.867	0.515	0.332	0.297
		Hashtags	3.361	1.705	1.035	0.632	0.471	0.373
		Users	4.508	1.614	0.879	0.522	0.374	0.332
		Tweets	2.374	1.339	0.761	0.441	0.297	0.254

ATM models outperform their LDA counterparts in both metrics. Pooling by conversations and hashtags frequently achieve the best or second best result. ATM-Conversations has a clear advantage over ATM-Hashtags for T ⩾ 100.

DOCUMENT RETRIEVAL EVALUATION

We treat every test tweet as a query and retrieve training tweets that are topically most similar to the query. Retrieved tweets are considered relevant if they have the same category as the query tweet.

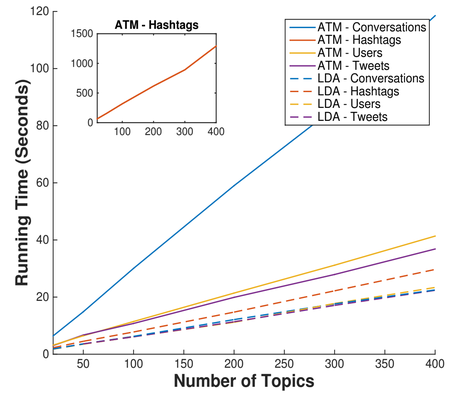




- ATM > LDA
- ATM-Conversations > all other pooling techniques
- LDA-User/Hashtags > LDA-Conversations/Tweets
- In overall, ATM-Conversations performs best

RUN TIMES

ATM models are slower than their LDA counterparts. Pooling by hashtags causes duplication of tweets and thus longer training times. ATM-Hashtags is one order of magnitude slower than for any of the other pooling techniques.



MAIN TAKEAWAY

Pooling tweets by conversations and applying ATM leads to the best results. The best alternative, ATM-Hashtags, achieves similar and sometimes better results, but takes considerably longer time to train.