Improving LDA topic modeling in Twitter with graph community detection

Federico Albanese[1,2] and Esteban Feuerstein[1,3]

[1] Instituto de Ciencias de la Computación, CONICET - UBA, Argentina
[2] Instituto de Cálculo, CONICET- UBA, Argentina
[3] Departamento de Computación, UBA, Argentina
falbanese@dc.uba.ar

Abstract. Texts can be characterized from their content using machine learning and natural language processing techniques. In particular, understanding their topic is useful for different tasks such as personalized message recommendation, fake news detection or public opinion monitoring. Latent Dirichlet Allocation (LDA) is an unsupervised generative model for the decomposition of topics, which seeks to represent texts as random mixtures over topics with a Dirichlet distribution, and each topic is characterized by a distribution over words. However, this method is challenging to apply when the text is short and sometimes incoherent, as is often the case with posts on social networks such as twitter. Therefore, different works have shown that tweet pooling (aggregating tweets into longer documents) improves LDA results, but its performance depends on which method was used to aggregating the texts.
We propose the new method to detect topics on twitter: "Community pooling". In this novel scheme, first we define the retweet graph where users are the nodes and retweets between them are the edges. Then, we use the Louvain method for community detection in order to uncover the communities (a group of users who mainly interact with each other but not with other groups). Finally we aggregate into a single document all the tweets authored by all users in a community. Therefore, this method drastically reduces the number of total documents and makes denser word co-occurrence matrix, which is beneficial to LDA algorithm.
With the intention of evaluating our model, we created two datasets of tweets with different characteristics. A first generic dataset involving various topics such as music, health and movies and a second dataset corresponding to an event: Biden's presidential inauguration day in the United States. We compare the performance of our model with state of the art schemes and previous pooling models in terms of document retrieval performance, cluster quality and supervised machine learning classification score. Results showed that Community pooling had a better performance on all datasets and tasks, with the only exception of the retrieval task on the event dataset. Moreover, Community polling was faster than all other aggregation techniques (less than half the running time), which is particularly useful in big data scenarios.

Keywords: Topic modeling · Community detection · Twitter · Text mining · Text clustering