The Journal of Open Source Software

TTLocVis: A Twitter Topic Location Visualization Package

Gillian Kant[1], **Christoph Weisser**[1, 2], **and Benjamin Säfken**[1, 2]

1 Centre for Statistics, Georg-August-Universität Göttingen, Germany **2** Campus-Institut Data Science, Göttingen, Germany

DOI: 10.21105/joss.02507

Software

- Review
- Repository
- Archive

Editor: Tania Allard

Reviewers:

- @sara-02
- @linuxscout
- @aozorahime

Submitted: 22 May 2020
Published: 26 October 2020

Summary

The package TTLocVis provides a broad range of methods to generate, clean, analyze and visualize the contents of Twitter data. TTLocVis enables the user to work with geo-spatial Twitter data and to generate topic distributions from Latent Dirichlet Allocation (LDA) Topic Models (Blei, Ng, & Jordan, 2003) for geo-coded Tweets. As such, TTLocVis is an innovative tool to work with geo-coded text on a high geo-spatial resolution to analyze the public discourse on various topics in space and time. The package can be used for a broad range of applications for scientific research to gain insights into topics discussed on Twitter. For instance, the package could be used to analyze the public discourse on the COVID-19 pandemic on Twitter in different countries and regions in the world over time. In particular, data from the recently provided COVID-19 stream by Twitter can be analyzed to research the discussion about COVID-19 on Twitter.[1] In the following, an overview of TTLocVis will be provided. Finally, it will also be discussed how TTLocVis extends existing related software solutions. The installation of TTLocVis can easily be done via pip. Further details on the installation and the package can be found in the packages repository or on the documentation website of TTLocVis.[2]

Statement of Need

In general, Topic Models are generative probabilistic models, that provide an insight into hidden information in large text corpora by estimating the underlying topics of the texts in an unsupervised manner. In Topic Models, each topic is a distribution over words that can be labeled by humans. For the purpose of labelling histograms and word clouds (for example, see graph) provide helpful visualizations for the decision-making process of the user (Blei et al., 2003).

Firstly, the package allows the user to collect Tweets using a Twitter developer account for any area in the world. Subsequently, the inherently noisy Twitter data can be cleaned, transformed and exported. In particular, TTLocVis enables the user to apply LDA Topic Models on extremely sparse Twitter data by preparing the Tweets for LDA analysis by the pooling Tweets by hashtags. The hashtags pooling algorithm (Mehrotra, Sanner, Buntine, & Xie, 2013) is implemented in a parallelized form in order to speed up the heavy computational task. The goal of hashtag pooling is to supply the Topic Models with longer documents than just single Tweets to reduce the problems of Topic Models to process short and sparse texts. The pooling idea can be summarized into the following steps: Pool all Tweets by existing hashtags and check the similarity of an unlabeled Tweet with all labeled Tweets (hashtag-pools). Subsequently, the unlabeled Tweets join the hashtag-pool with the highest cosine

[1] https://developer.twitter.com/en/docs/labs/covid19-stream/overview
[2] https://ttlocvis.readthedocs.io/en/latest/#installation

similarity value, if the value exceeds a certain threshold. This process is repeated for all unlabeled Tweets. The resulting topic distributions that are computed with an LDA model that is trained on the pooled Tweets are substantially improved. When trained with sufficient data, clear topics can be generated, and the shortcoming of LDAs with short and sparse text are minimized.

TTLocVis provides options for automatized Topic Model parameter optimization. Furthermore, a distribution over topics is generated for each document. The distribution of topics over documents can be visualized with various plotting methods (for example, see figure Word Cloud). The average prevalence of topics in the documents at each day can be plotted as a time series (for example see figure Time Series), in order to visualize, how topics develop over time.

Above this, the spatial distribution of Tweets can be plotted on a world map, which automatically chooses an appropriate part of the world, in order to visualize the chosen sample of Tweets. As part of the mapping process, each Tweet is classified by its most prevalent topic and color coded (for example see figure Word Map 1 and figure World Map 2 for the spatial distribution of the same selected topics at different points in time).

Comparison with existing tools

To the knowledge of the authors, no Python Package with a comparable functionality of TTLocVis is currently available. A web tool that is most related to TTLocVis is TweetViz (Stojanovski, Dimitrovski, & Madjarov, 2014). TweetViz provides word clouds and topic distributions for Twitter data. However, TTLocVis improves on TweetViz by optimizing the LDA input with tweet pooling and options for geo-spatial and temporal analysis. Further, a major limitation of TweetViz is that the number of topics for the LDA estimation is always fixed at 20. TTLocVis gives the option to define a range of potential topic numbers and also includes an algorithm to select the optimal topic number according to coherence scoring.

Alternatively, Twitter data may be analyzed with a web application (Malik et al., 2013) with an LDA Topic Model. The authors use so-called bins resembling time intervals for the Topic Model estimation. For each of these bins, an LDA Topic Model is estimated in order to account for the topical change over time. They then use cosine similarity to align the topics from several bins to a resulting topic. However, in this framework, Topic Models are estimated on very small samples on which LDA Models usually do not perform well. In contrast to this approach, the LDA Model is trained on pooled tweets in TTLocVis in order to improve the estimation results.

A further web application (Onorati, Díaz, & Carrion, 2019) offers functionality to generate word clouds, treemaps and map visualization. In contrast to TTLocVis, topics are not estimated by LDA, but rather by semantic relations. The focus of this application is on the contents of individual tweets with regard to disaster-related classification.

A recent analysis of COVID-19 related Tweets can be found in the integration of the packages Birdspotter and Evently (Kong, Ram, & Rizoiu, 2020), in order to analyze retweet cascades. Birdspotter is a package to analyze the social influence and botness of Twitter users, while Evently can be used to model the temporal spread of information.

Figures

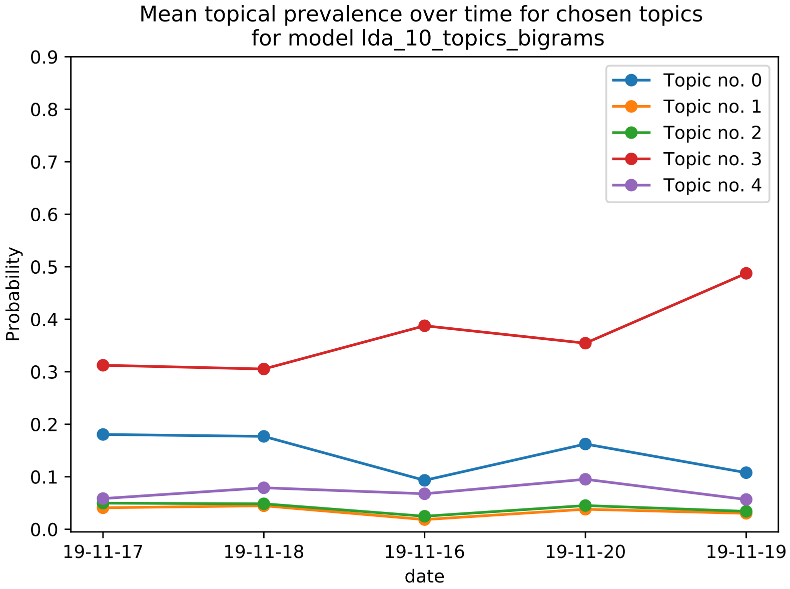

Figure 1: Time Series.

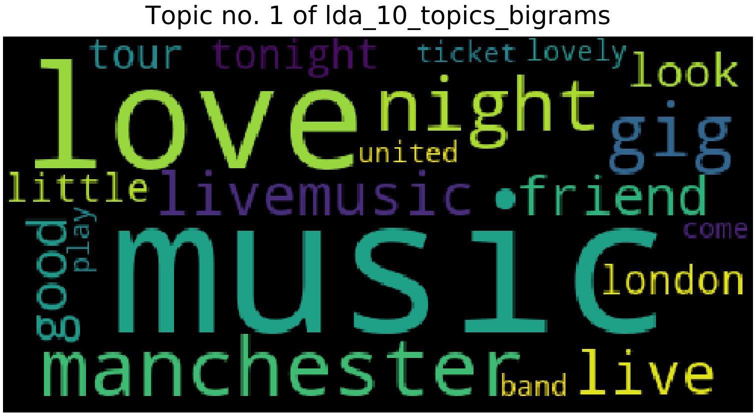

Figure 2: Word Cloud.

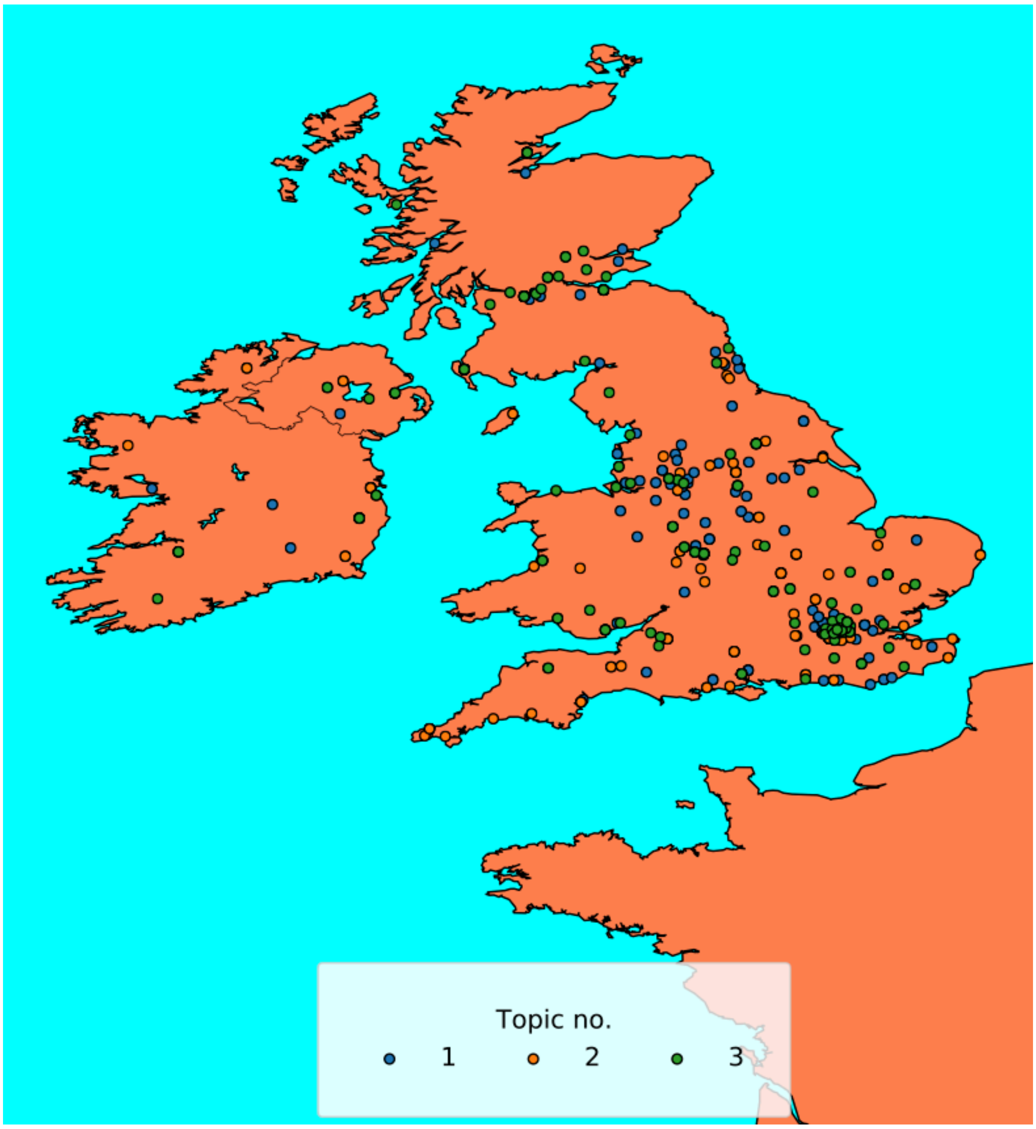

Figure 3: World Map 1.

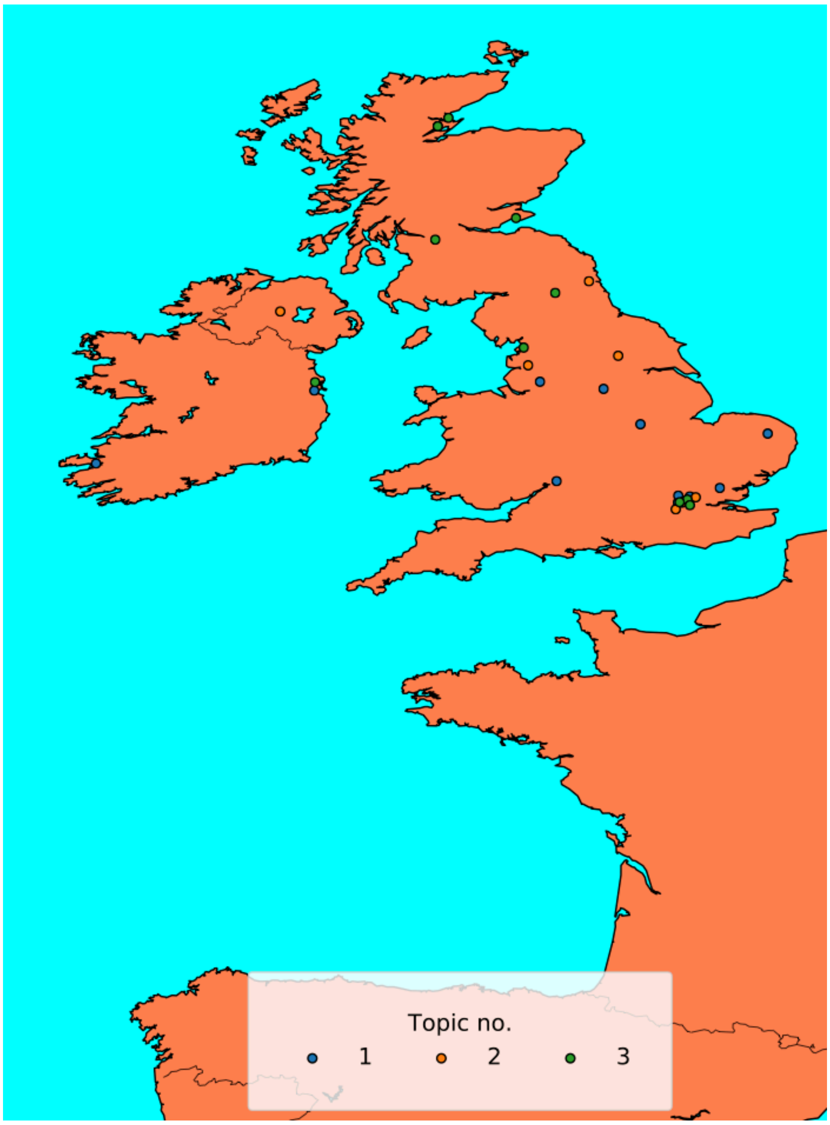

Figure 4: World Map 2.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

Kong, Q., Ram, R., & Rizoiu, M.-A. (2020). Evently: A toolkit for analyzing online users via reshare cascade modeling. Retrieved from http://arxiv.org/abs/2006.06167

Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., & Shneiderman, B. (2013). TopicFlow: Visualizing topic alignment of twitter data over time. In *Proceedings of the 2013 ieee/acm international conference on advances in social networks analysis and mining* (pp. 720–726). doi:10.1145/2492517.2492639

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 889–892). doi:10.1145/2484028.2484166

Onorati, T., D́ıaz, P., & Carrion, B. (2019). From social networks to emergency operation centers: A semantic visualization approach. *Future Generation Computer Systems*, 829–840. doi:10.1016/j.future.2018.01.052

Stojanovski, D., Dimitrovski, I., & Madjarov, G. (2014). Tweetviz: Twitter data visualization. In *Proceedings of the data mining and data warehouses* (pp. 1–4).