A MODULAR APPROACH TO TOPIC MODELING FOR HETEROGENEOUS DOCUMENTS

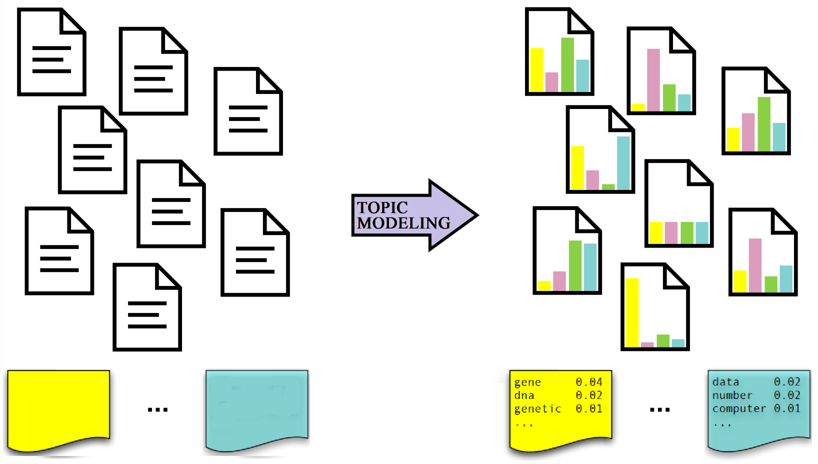
(Extended Abstract)

Giovanni Toto[1], Emanuele Di Buccio[1,2]

[1]Department of Statistical Sciences, University of Padova
[2]Department of Information Engineering, University of Padova

IIR 2022 - 12th Italian Information Retrieval Workshop, June 29th-30th, 2022, Milano, Italy

Topic modeling



Heterogeneous documents

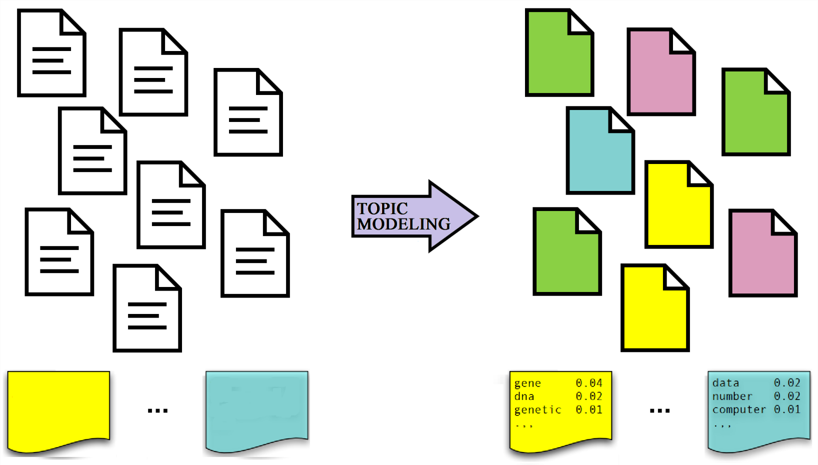
We are interested in dealing with two types of heterogeneity:

- heterogeneity of document length,
- heterogeneity of document descriptors.

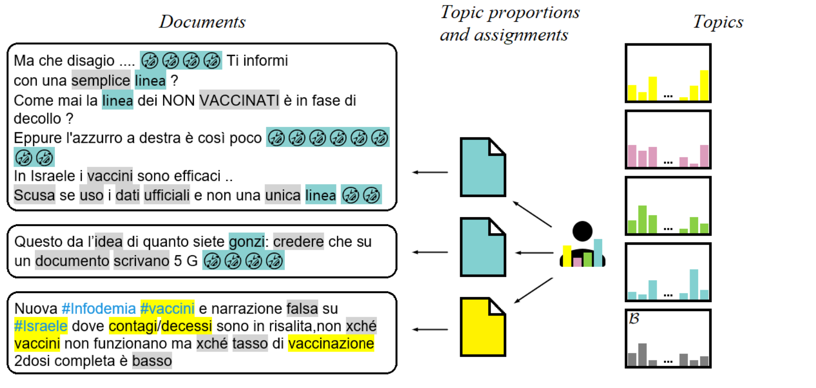
In this stage of the work we focused on *microblogs* – *Twitter* – since:

- posts can be both long and short,
- posts can contain words, *hashtags*, *emoji*, *mentions*, ...

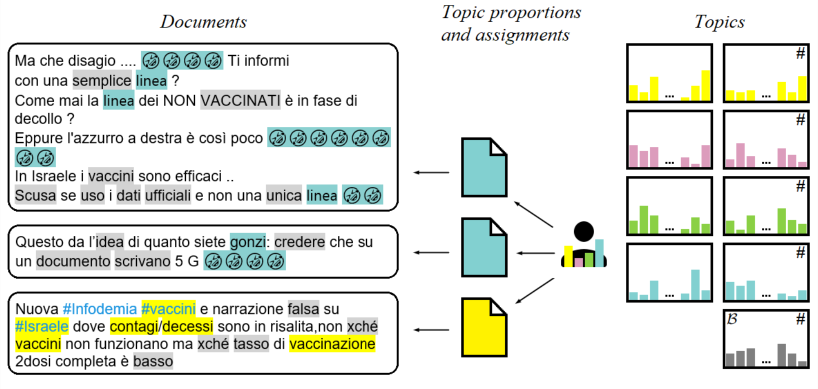
Twitter-LDA e Hashtag-LDA



Twitter-LDA

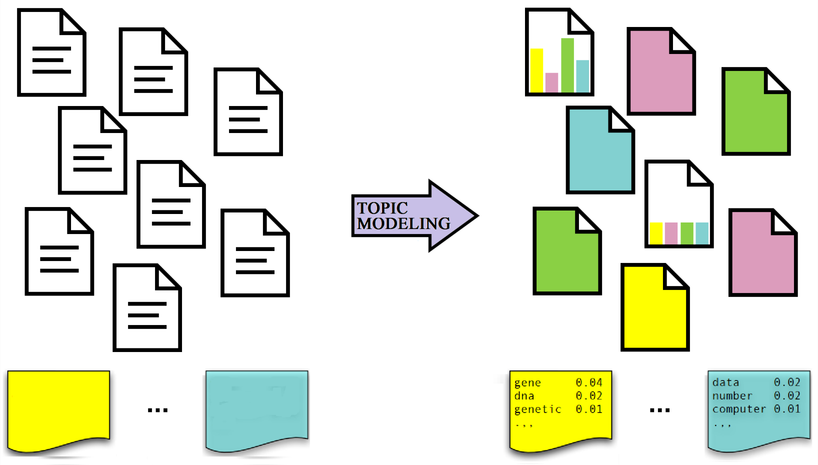

Documents
Topic proportions
and assignments
Topics
Ma che disagio .... Ti informi
con una semplice linea ?
Come mai la linea dei NON VACCINATI è in fase di
decollo ?
Eppure l'azzurro a destra è così poco
In Israele i vaccini sono efficaci ..
Scusa se uso i dati ufficiali e non una unica linea
Questo da l'idea di quanto siete gonzi: credere che su
un documento scrivano 5 G
Nuova #Infodemia #vaccini e narrazione falsa su
#Israele dove contagi/decessi sono in risalita,non xché
vaccini non funzionano ma xché tasso di vaccinazione
2dosi completa è basso
B

Hashtag-LDA

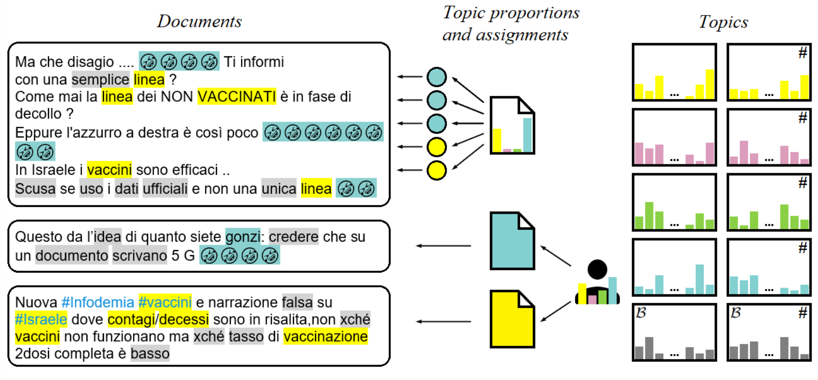


Documents
Topic proportions and assignments
Topics
Ma che disagio .... Ti informi
con una semplice linea ?
Come mai la linea dei NON VACCINATI è in fase di decollo ?
Eppure l'azzurro a destra è così poco
In Israele i vaccini sono efficaci ..
Scusa se uso i dati ufficiali e non una unica linea
Questo da l'idea di quanto siete gonzi: credere che su un documento scrivano 5 G
Nuova #Infodemia #vaccini e narrazione falsa su #Israele dove contagi/decessi sono in risalita,non xché vaccini non funzionano ma xché tasso di vaccinazione 2dosi completa è basso

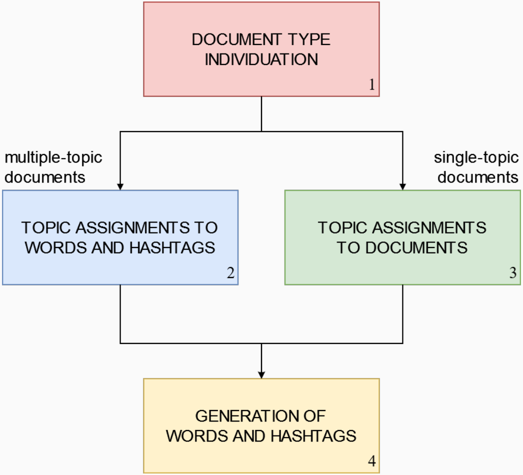
A topic model for heterogeneous documents



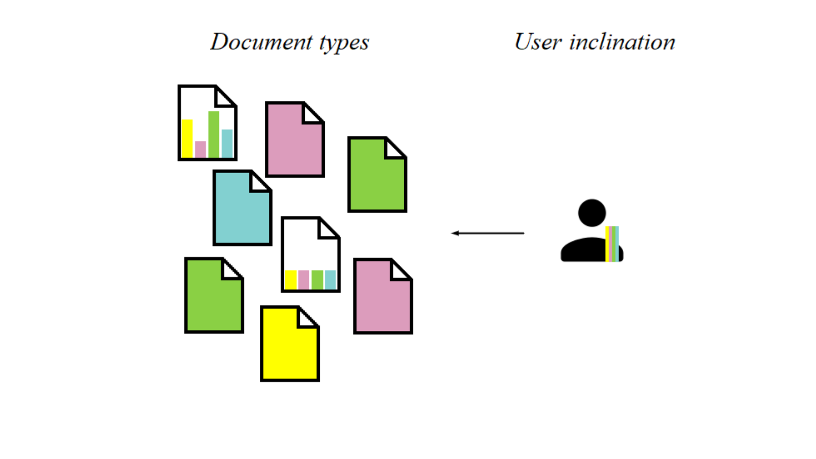
Proposed model



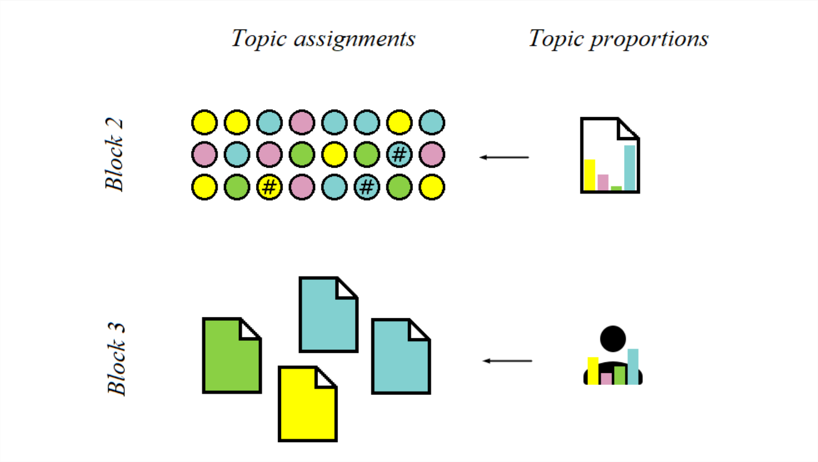
Block division of the generative process



Block 1



Block 2 and Block 3



Block 4

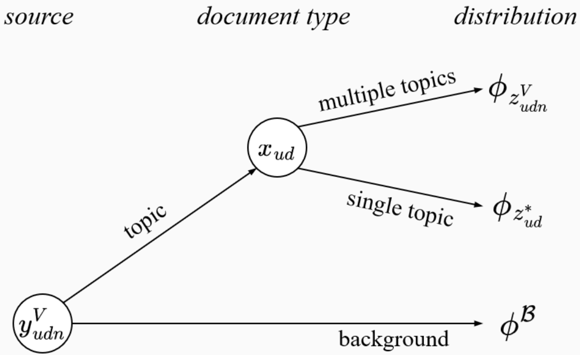


Figure: Generation of a word *udn*.

Experimental evaluation and Posterior inference

A first evaluation was carried out on a collection of tweets:

- 8895 tweets in Italian about COVID-19,
- 101 distinct users.

A *Collapsed Gibbs Sampler* is used to perform the approximate posterior inference:

- 1000 iterations with a burn-in period of 700 iterations
- *Monte Carlo* on one iteration every ten

Quantitative analysis

Topic Coherence metrics

- a topic is perceived as useful and coherent if its top words tend to occur together

Distance from the corpus distribution

- a topic is perceived as useless or overly general if it is similar to the corpus distribution

Comparison of topic models

	TC-PMI	TC-LCP	JS div.
LDA	**1.3023**	-3.1437	0.2020
TLDA	1.2026	-2.9671	0.2703
HLDA	0.9863	**-2.9563**	0.2159
MLDA	1.2909	-2.9866	**0.2745**

Double representation of a topic

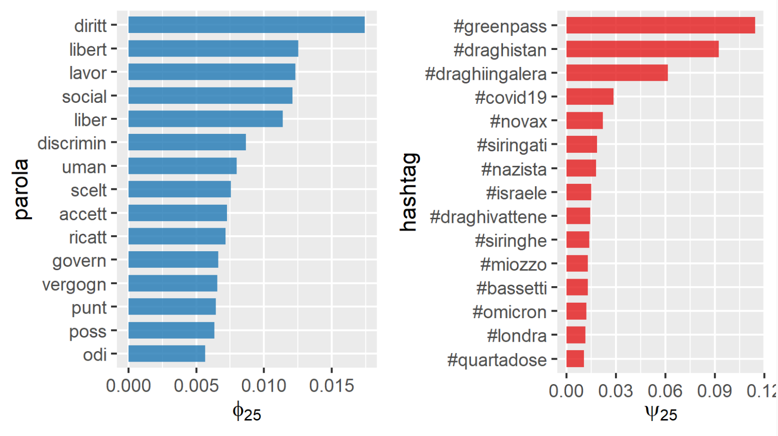


Figure: List of the 15 *top words* and 15 *top hashtags* of topic 25.

Future work

The subsequent steps will be:

- investigate in detail the effect of the number of topics on the proposed approach
- investigate how to tailor the model to heterogeneous text collections
- extend the set of adopted baselines
- evaluate the effectiveness in diverse tasks such as (hash)tag recommendation, text classification, and clustering
- perform a qualitative analysis through a case study

References

- Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.
- Lin, T., Tian, W., Mei, Q. & Cheng, H. (2014). The Dual-Sparse Topic Model: Mining Focused Topics and Focused Terms in Short Text. *Proceedings of the 23rd International Conference on World Wide Web*, 539–550.
- Zhao, F., Zhu, Y., Jin, H. & Yang, L. T. (2016). A Personalized Hashtag Recommendation Approach Using LDA-Based Topic Model in Microblog Environment. *Future Gener. Comput. Syst., 65* (100), 196–206.
- Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. & Li, X. (2011). Comparing Twitter and Traditional Media Using Topic Models. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee & V. Mudoch (Cur.), *Advances in Information Retrieval* (pp. 338–349). Springer Berlin Heidelberg.

Probabilistic graphical model