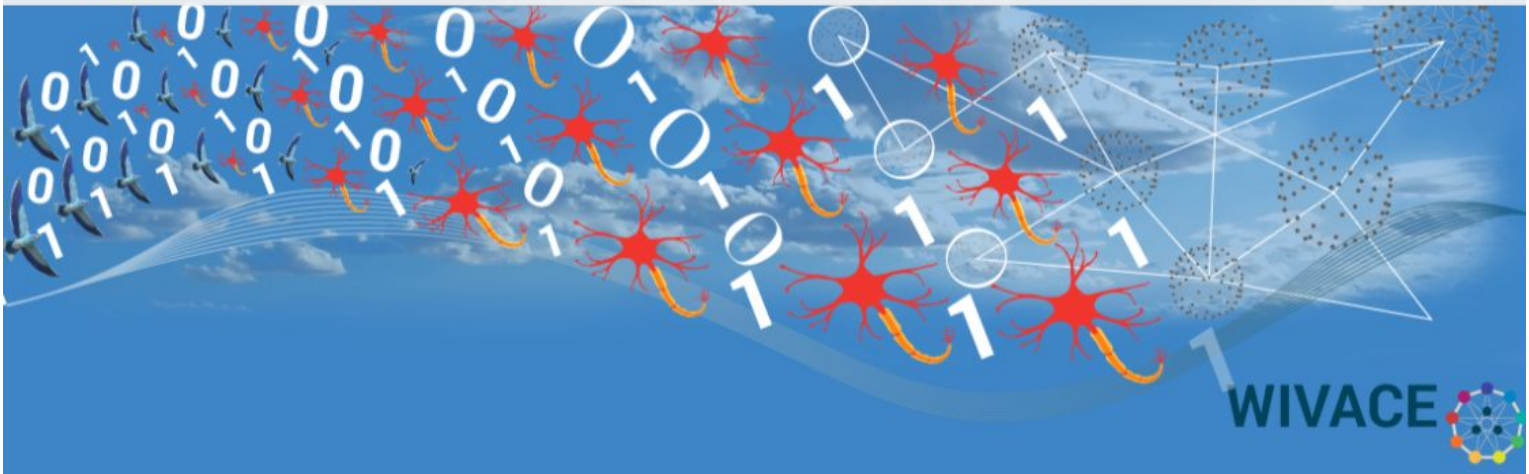
European Centre for Living Technology



WIVACE 2017

Book of Abstracts

Venice, 19-21 September 2017

For info please visit: http://wivace.org/2017

Università Ca'Foscari Venezia
Dipartimento di Scienze Ambientali, Informatica e Statistica

CCIS

Springer

Ca'Foscari Zattere
Cultural Flow Zone

WIVACE 2017 - Book of Abstracts
Andrea Roli, Debora Slanzi, Marco Villani eds.

Università degli Studi di Modena e Reggio Emilia
Dip. di Scienze Sociali, Cognitive e Quantitative
Viale Allegri 9, 42121 Reggio Emilia (RE)

ISBN 978-88-903581-3-5 (online)

Index

Modeling emerging topics on sustainable urban development perception: the case of Hangzhou Future Sci-Tech City.

Debora Slanzi[1,2], Valentina Anzoise[1], and Irene Poli[1,2]

[1] European Centre for Living Technology, S. Marco 2940, 30124 Venice, Italy
[2] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Cannaregio 873, 30121 Venice, Italy
{debora.slanzi, valentina.anzoise, irenpoli}@unive.it

Abstract

Recent social and economical literature has been particularly concerned in the investigation of urban development [1,2,3]. Especially in China, where the process of accelerated urbanization is a result of rapid economic growth and dedicated policies, cities and countrysides are changing at an unprecedented scale and pace. As a consequence, landscape and lifestyle are radically transformed raising social, economic, and environmental sustainability issues and stability problems [4]. Due to the complexity of the study of urban systems, it emerges that there is an increasing need of adopting appropriate methods for analyzing and modeling social data, both from a quantitative and a qualitative perspective [5]. One particular approach for the analysis of social systems is the textual data analysis. Textual documents in fact provide a valuable source of data for the identification and the measurement of latent variables, and statistics and machine learning researchers have developed several approaches to study these structures of data [6,7,8]. Among them, Topic Modeling approaches (TM) aim to automatically inferred from textual data the rich latent topics of a set of documents or texts [6]. TMs have been successfully used across a variety of fields as they can discover complex latent structure in the data [9,10]. Topics are estimated with probabilistic distributions over a vocabulary of words and according to the co-occurence of words within each analyzed text according with a probabilistic generative process. This process considers a collection of D documents (or texts), each containing $N_d \subseteq V$ words, $d = 1, \ldots, D$, and V represents the set of distinct elements of the vocabulary used in the analysis. Moreover a set of K latent topics is defined and assumed to be representative of the documents. The probabilistic generative process consists then of the following steps:

- a V-dimensional Dirichlet probability distribution, $\beta_k \sim Dir(\eta)$, is determined for each topic k, $k = 1, \ldots, K$, assessing the probability according to which words are generated from the k-th topic;
- a K-dimensional Dirichlet probability distribution, $\theta_d \sim Dir(\alpha)$ is determined for each document d, $d = 1, \ldots, D$, assessing the expected proportion of words that can be attributed to each topic;

– for each word in the document
  - a value $z_{d,n}$ for a multinomial distribution $Z_{d,n} \sim Mult_K(\theta_d)$, $n = 1, \ldots, N_d$, is sampled denoting which topic is associated with such word, and
  - a word value $w_{d,n}$ from a multinomial distribution $W_{d,n} \sim Mult_V(\mathbf{B}z_{d,n})$, is sampled where the matrix $\mathbf{B} = [\beta_1 \cdots \beta_K]$ encodes the distributions over words in the vocabulary associated with the K topics.

When additional information regarding the documents is available, it can be included in the model as a set of covariates $\mathbf{X}$. The Structural Topic Model (STM) proposed by Roberts et al. [11,12,13] represents a particular class of TMs where the inclusion of covariates of interest can affect the topical prevalence (i.e., the frequency with which a topic is discussed) and topical content (i.e., the words used to discuss a topic) of the model. The covariates are introduced in the TM approach by means of different prior distributions for document-topic proportions and topic-word distributions. For the specific procedure on how these prior distributions are defined and how the Topic Model estimation process is modified, we refer to [13].

In this work we analyze the citizens' perception on the urban development of a recently established high-tech zone in China, i.e. Hangzhou Future Sci-Tech City, collected through several face-to-face interviews which have been conducted in spring-summer 2016. This area, 113 km^2 large, was previously covered by farmlands and recently is benefitting of dedicated-national policies to implement talents strategies, improve scientific and technology innovation and foster new entrepreneurship. The planning of this new territorial entity is producing different effects on the economy, on the environment and on the landscape, both positive and negative, and this generates different perception and understanding in the social system. The interviews were conducted using a composition of photos of the area. Images are inherently polysemic, but each of them poses the focus on a different -even controversial- aspects of the urban development of the area. This can be seen in Figure 1, in which the wordcloud shows the relationship between the most frequent words (stemmed with standard pre-processing textual analysis techniques) and the content of the photos used in the interviews.
We then develop a Structural Topic Model using the photos of the interview as covariates in the model to extract the key topics of collected textual data. The introduction of covariates in the model is able to highlight if particular visual stimuli bring out specific perceptions or latent issues. The emerging categories of perception are presented in Figure 2, where each category represents an estimated topic described by its most frequent and exclusive words. From the results of this analysis we notice that the perception of the people interviewed is mostly of great appreciation for the great economic development, with some, but minor, concerns on the negative effects of this development on the society and on the environment. Future developments of this research will concern the estimation of the network of relationships among these categories with Probabilistic Graphical Models (PGMs). Next step of the research will the estimation of the network of relationships among these emerging categories by means of

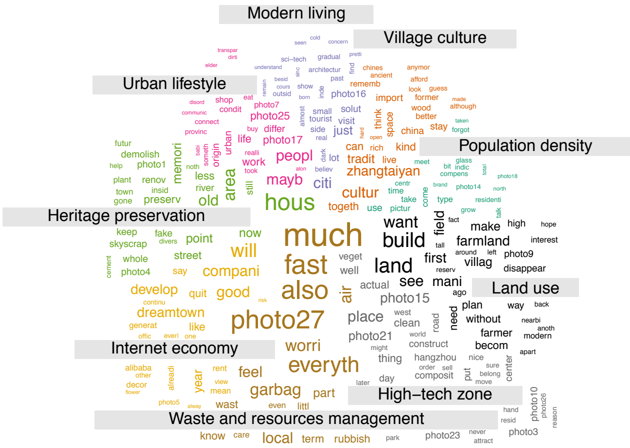


Fig. 1: Wordcloud of the most frequent words (vocabulary) associated with the content of the photos used in the interviews.

Probabilistic Graphical Models (PGMs). PGMs are in fact very efficient and effective statistical models to estimate the complex structures of probabilistic dependences and independencies which characterized complex social systems.

References

1. Zhao, P., Li, P.: Rethinking the relationship between urban development, local health and global sustainability. Current Opinion in Environmental Sustainability. 25, 14–19 (2017)
2. Yang, B., Xu, Y., Shi, L.: Analysis on sustainable urban development levels and trends in China' cities. Journal of Cleaner Production. 141, 868–880 (2017)
3. Riffat, S., Powell, R., Aydin D.: Future cities and environmental sustainability. Future Cities and Environment. 2:1 (2016)
4. Wu, F.: Emerging Chinese Cities: Implications for Global Urban Studies. The Professional Geographer. 68(2), 338–348 (2016)
5. Dolfin, M., Leonida, L., Outada. N.: Modeling human behavior in economics and social science. Physics of Life Reviews. In press (2017)
6. Blei, D.M.: Probabilistic topic models. Communications of the ACM. 55(4), 77–84 (2012)
7. Grimmer, J., Stewart, B.: Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents. Political Analysis. 21(3), 267–297 (2013)
8. Li, G., Feng, S., Jun, T.: Textual analysis and machine leaning: Crack unstructured data in finance and accounting. The Journal of Finance and Data Science. 2(3): 153–170 (2016)

Fig. 2: Estimated topics with the emergent categories of perception.

9. Tvinnereim, E., Liu, X., Jamelske, E.M.: Public perceptions of air pollution and climate change: di erent manifestations, similar causes, and concerns. Climatic Change 2016.1–14 (2016)
10. Reich, J., Tingley, D., Leder-Luis, J., Roberts, M.E., Stewart, B.M.: Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses. Journal of Learning Analytics. 2(1), 156–184 (2015)
11. Roberts, M.E., Stewart, B.M., Tingley, D., Airoldi, E.M.: The Structural Topic Model and Applied Social Science. Neural Information Processing Society (2013)
12. Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G.: Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science. 58(4),1064-1082 (2014)
13. Roberts, M.E., Stewart, B.M., Airoldi, E.M.: A model of text for experimentation in the social sciences. Journal of the American Statistical Association. 111(515), 988-1003 (2016)