SLaTE:
A System for Labeling Topics with Entities

Anne Lauscher
anne@informatik.uni-mannheim.de
University of Mannheim, Germany

Federico Nanni
federico@informatik.uni-mannheim.de
University of Mannheim, Germany

Simone Paolo Ponzetto
simone@informatik.uni-mannheim.de
University of Mannheim, Germany

In recent years, the Latent Dirichlet allocation (LDA) topic model (Blei, Ng, and Jordan, 2003) has become one of the most employed text mining techniques (Meeks and Weingart 2012) in the digital humanities (DH). Scholars have often noted its potential for text exploration and distant reading analyses, even when it is well known that its results are difficult to interpret (Chang et al, 2009) and to evaluate (Wallach et al, 2009).

At last year's edition of the Digital Humanities conference, we introduced a new corpus exploration method able to produce topics that are easier to interpret and evaluate than standard LDA topic models (Nanni and Ruiz, 2016). We did so by combining two existing techniques, namely Entity linking and Labeled LDA (L-LDA). At its heart, our method first identifies a collection of descriptive labels for the topics of arbitrary documents from a corpus, as provided from the vocabulary of entities found within wide-coverage knowledge resources (e.g., Wikipedia, DBpedia). Then it generates a specific topic for each label. Having a direct relation between topics and labels makes interpretation easier, and using a disambiguated knowledge resource as background knowledge limits label ambiguity. As our topics are described with a limited number of unambiguous labels, they promote interpretability, and this may sustain the use of the results as quantitative evidence in humanities research (Lauscher et al, 2016).

The contributions of this poster cover the release of: a) a complete implementation of the processing pipeline for our entity-based LDA approach; b) a three-step evaluation platform that enables its extensive quantitative analysis.

Entity–based Topic Modeling Pipeline

Figure 1 illustrates the computational pipeline of our system; python classes are represented in rectangles. First of all, a set of text files is imported into the system and several preprocessing steps are applied to the textual content. Next, the data is sent to the entity linking system TagMe (Ferragina and Scaiella, 2010), which disambiguates against Wikipedia. As a result of this step, for each document a set of related Wikipedia entities is retrieved. Now, the data is inserted into a MySQL database.
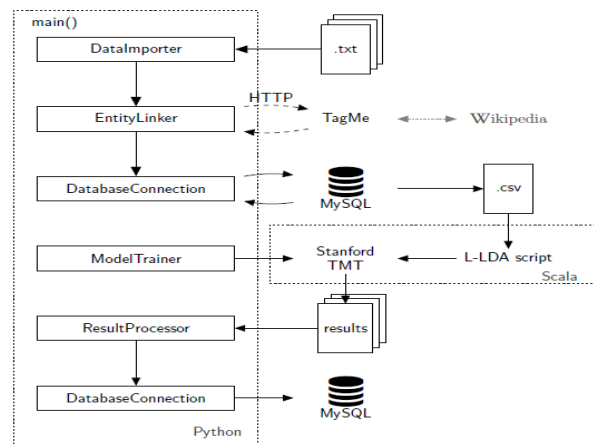

Figure 1. Architecture of the pipeline

Afterwards, the TF-IDF measure is computed over the entities, which we use to rank all the entities for each document in descending order. Then, the top k entities as well as their corresponding documents are exported into a comma-separated values file that is given as input to the L-LDA implementation of the Stanford Topic Modeling Toolbox. Finally, after running L-LDA and applying several post-processing steps, we obtain a document-topic distribution saved in the database in which each topic is described by an unambiguous label linked to Wikipedia.

The whole source code is available for public download on Github. Given a working Python, Java, and Scala runtime as well as a running MySQL installation our pipeline is ready directly out-of-the-box. The specific configuration according to the user's needs can be made via a simple text file.

Three–Step Evaluation Platform

Document Labels

In order to assess the quality of the detected entities as labels we developed a specific browser-based evaluation platform, which permits manual annotations. This platform presents a document on the right of the screen and a set of possible labels on the left (See Figure 2). Annotators are asked to pick labels that precisely describe the content of each document. In case the annotator does not select any label, this is also recorded by our evaluation system.

Entities and Topic Words

In order to establish if the selected entities were the right labels for the topics produced, we developed two additional evaluation steps. Inspired by the topic intrusion task (Chang et al, 2009), we designed a platform that permits to evaluate the relations between labels and topics using two evaluation modes: For one evaluation mode (that we called Label Mode - Figure 3), the annotator is asked to choose, when possible, the correct list of topic-words given a label. For the other, he/she was asked to pick the right label given a list of topic words (aee Figure 4). In both cases, the annotator is shown three options: one of them is the correct match, while the other two (be they words or labels) come from other topics related to the same document.
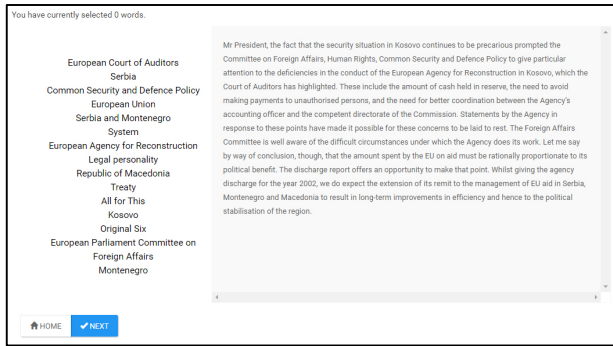

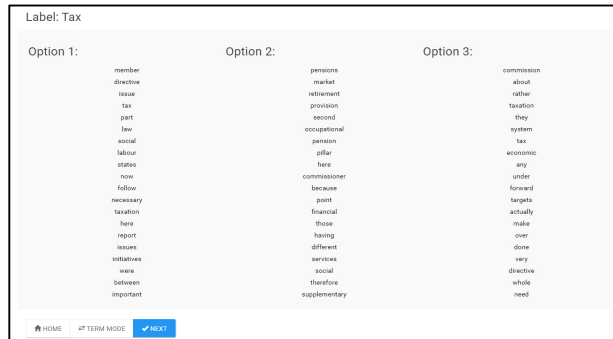Figure 2. Entities as Labels evaluation interface.
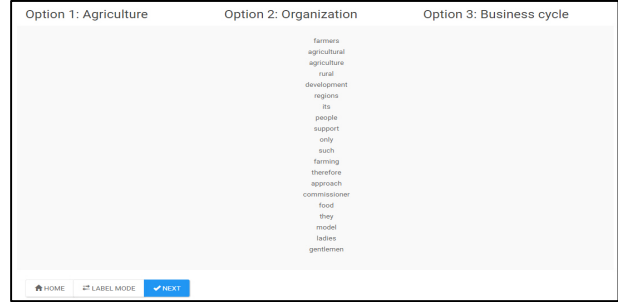

Figure 3. Label-Mode Evaluation

Figure 4. Term-Mode Evaluation

Bibliography

Blei, D. M., Ng, A.Y., and Jordan, M. (2003) "Latent dirichlet allocation." *Journal of machine Learning research.*

Chang, J., Boyd-Graber, J.L., Gerrish, S., Wang, C. and Blei, D.M. (2009) "Reading tea leaves: How humans interpret topic models." *Advances in neural information processing systems.* 21. 288-296.

Ferragina, P., and Scaiella, U. (2010) "TAGME: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities)." In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management.*

Lauscher, A., Nanni, F., Ruiz Fabo, P., and Paolo Ponzetto, S. (2016). "Entities as topic labels: combining entity linking and labeled LDA to improve topic interpretability and evaluability." IJCol-Italian journal of computational linguistics. 2(2): 67-88.

Meeks, E., and Weingart, S.B. (2012). "The digital humanities contribution to topic modeling." *Journal of Digital Humanities* 2.1.

Nanni, F., and Ruiz Fabo, P. (2016) "Entities as topic labels: Improving topic interpretability and evaluability combining Entity Linking and Labeled LDA." *DH2016.*

Wallach, H. M., Murray, I., Salakhutdinov, R. and Mimno, D. (2009). "Evaluation methods for topic models." *Proceedings of the 26th Annual International Conference on Machine Learning.*