Department of Computer Science

Machine Learning
Valentin Zambelli

Data Modeling, Winter Term 2016/17

Final Report: Analyzing fake news articles using LDA topic modeling

1 Introduction

Topic modeling is a widely used technique to analyze corpora of text data. Its aim is to find latent factors that carry semantic meaning, similar to factor models for non-textual data. In this report a topic model is applied to a corpus of fake news articles from October-November 2016. The most popular topic modeling technique and the one used in this report is Latent Dirichlet Allocation (LDA).

2 Latent Dirichlet Allocation

2.1 Basic algorithm

LDA was developed by Blei et al. (2003). It can be seen as a fully probabilistic extension of Latent Semantic Indexing (LSI) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), building on the work of probabilistic LSI (Hofmann, 1999) . Compared to previous approaches LDA is based on a well-defined generative process. This gives the user a rich set of tools to explore the corpus from different viewpoints and generalizes well to new documents. Figure 1 shows the graphical model representation of the underlying generative process for each word in a corpus D:

1. For each topic,
   (a) Draw a distributions of words $\beta \sim \ Dir_v(\eta)$
2. For each document
   (a) Draw a vector of topic proportions $\theta \sim \ Dir(\alpha)$
   (b) For each word,
      i. Draw a topic assignment $Z_{d,n} \sim \ Mult(\theta)$
      ii. Draw a word $W_{d,n} \sim \ Mult(\beta_{z_{d_n}})$

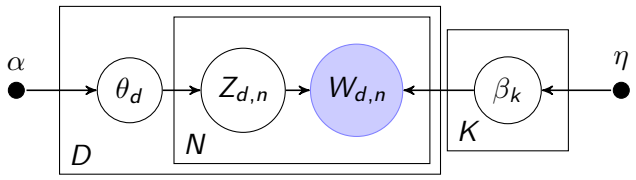

Figure 1: Plate Diagram of LDA.

By applying a generative process to all aspects of the corpus (words, documents, topics) we can not only solve the problem of assigning topics to documents but also gain a convenient tool to analyze the composition of the topics themselves as well as the context of specific words. This flexibility makes LDA models a great solution for information retrieval problems (like document

search engines). LDA models encode some specific assumptions that simplify the learning process. First, the order in which documents appear in the corpus does not matter, they are seen as independent from each other, given the latent variables. An extension of LDA called Dynamic Topic Modeling works without this assumption and can be used to track developments through time (Blei, 2006). Second, also the order in which the words appear in a document is not taken into account. This is called the "bag-of-words" model, where a document is fully described by the counts of occurring words. This means LDA cannot take any grammatical structure into account.

2.2 Hierarchical Dirichlet Process (HDP) LDA

One potential issue with LDA is that the number of topics has to be specified a priori. This becomes an issue when the corpus is completely unknown and LDA is used as an exploratory tool. Whye Teh, Jordan, Beal, and Blei (2006) have developed an approach based on hierarchical Dirichlet processes that can infer the number of topics from the data. HDP LDA seeks to capture the uncertainty in the number of topics by using a Dirichlet process (DP), a distribution of probability measures (such that marginals on finite partitions are again Dirichlet distributed) and estimate the correct number of topics from data. It is straightforward to apply this to simple finite mixture models but for LDA each document has specific mixture proportions. To be able to share topics between documents we need a hierarchical process as introduced by Teh et al. where the the document-specific DPs are themselves drawn from a shared DP. HDPs cannot only be used for topic modeling but are a general purpose technique to share clusters between groups of observations.

3 Applying LDA

The popular Gensim (Rehurek & Sojka, 2010) framework was used for all modeling tasks.

3.1 Description of data

- Muslims BUSTED: They Stole Millions In Gov't Benefits
- SHARIA IN AMERICA? How Minneapolis Muslims Are Still Being Recruited By Terrorists
- Hillary's 'Russian Hack' Hoax: The Biggest Lie of This Election Season

3.2 Preprocessing steps

As with all natural language tasks the data has to undergo substantial processing before it can be used in modeling. This necessary to reduce the size of the corpus and noise in the documents. Also a way has to be found to convert the documents into numerical feature vectors. First, the articles were split into individual, lowercased tokens and all special characters (such as .,/§ etc.) as well as stopwords (from the stoplist implemented in Gensim) removed. All words shorter than 2 characters and longer than 20 were removed. Then all words were stemmed using the Snowball stemmer method, an extension to the popular Porter stemmer. From the remaining tokens a dictionary was created that maps tokens to (numerical) IDs. This created a vocabulary of 93,963 individual tokens. From this large dictionary all tokens were deleted that appear in less than 5 documents or in more than 90% of all documents. The remaining vocabulary had a size of 19,998 tokens.

Health scares	**Lock her up!**	**Obamacare**	**Jews and Muslims**	**Syria**
health	clinton	obama	israel	russian
study	fbi	american	world	russia
cancer	email	state	jewish	military
drug	hillary	president	christian	war
use	investigating	campaign	muslim	forces
research	comey	law	palestinian	syria
body	elect	obamacare	state	state
effect	trump	federal	said	syrian
food	campaign	hillary	american	said
disease	new	video	years	united

Table 1: Five select topics and most likely words

3.3 Feature engineering

The next task was to generate a numerical feature vector for each document that can be used to train the model. Since LDA assumes a bag of words model this is also what I choose to encode the feature vectors. Under a bag-of-words model each document is represented by a vector of word ID and count tuples. It is possible to extend the bag-of-words model to bigrams or trigrams to add information about structure of the documents.

3.4 Selecting the number of topics

There are different ways to select the number of topics k for an LDA model. The first and simplest one is to create multiple models with varying k and assess the usefulness of the found topics. This approach is of course subject to bias by the researcher. If the LDA model will be used as the basis for a task other than exploration (for example text classification) it is recommended to treat the number of topics as a hyperparameter of the final task and tune k accordingly. A second approach is to compare the *perplexity* of different models (see for example Wallach (2006)). Perplexity is the negative log-likelihood of unseen documents. The better the model the lower the perplexity, as it captures all properties of the corpus to sufficiently describe unseen data. It should be noted that perplexity is likely to decrease as long as we increase the number of topics. It is also worth noting that perplexity has been shown to not correlate very well with how humans would asses the usefulness of a topic model (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009) . A third approach is to use the a HDP as described in the algorithm section. HDP has been shown to produce solutions with equal perplexity to optimal LDA models (Whye Teh et al., 2006), without requiring any kind of tuning.

4 Results

Fitting a HDP model resulted in **20** found topics. Analyzing the development of perplexity on a holdout set of 200 documents, by fitting 30 models from one to 30 topics, produces an expected knee-shaped curve. The results can be seen in figure 2. According to the perplexity measure, 8-10 topics would already have sufficiently captured the structure in the corpus. This observation is aided by the fact that not all of the 20 topics found by the HDP model were actually useful and some were more or less duplicates of others. However, I have found that this also held true for a model with only 10 topics, resulting in a smaller number of useful topics than found by the HDP model. I therefore choose to stick with the number of topics found by the HDP model for my analysis. An additional benefit of HDP models is their runtime. Creating 30 topic models and analyzing their perplexity took roughly an hour on a modern Intel Quadcore CPU, whereas

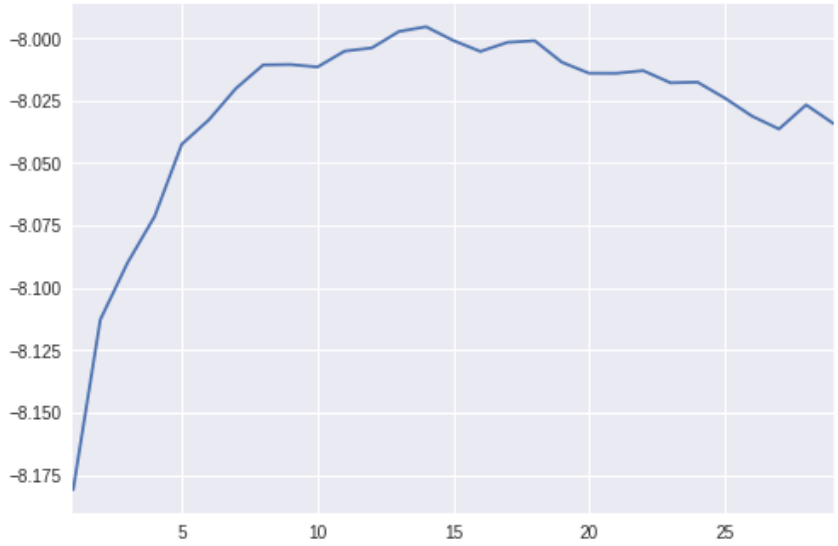

Figure 2: Perplexity for models ranging from one to 30 topics

fitting a HDP model (which itself is more complex than a standard LDA model) only takes about 90 seconds. Given a trained LDA model there are multiple ways we can use it to analyze the data. In the following sections I will look at the topics found, the most prevalent topics in the given corpus, classifying documents with the model as well as finding topics for a specific word. In all the tables the stemming was reversed for improved readability.

Topic perspective

Table 1 shows 5 of the 20 topics found in the corpus. A full list of topics can be found in the appendix. As we can see there is a mix of fake news mainstays like **Health Scares** and **Jews and Muslims** as well as more recent ones like **Lock her up!**, dealing with Hillary Clinton's email scandal, or **Obamacare**.

Corpus perspective

Table 2 shows the five most prevalent topics in the fake news corpus. As expected the corpus focuses on the 2016 Election. The topics were choosing according to the UMass topic coherence index (Mimno, Wallach, Talley, Leenders, & Mccallum, 2011) .

Document perspective

Given a specific document we can find the topics associated with it. The following examples show abstracts of news articles and their associated topics and topic likelihoods.

2016 Election	**Blacks like Trump**	**America and Russia**	**Dakota pipeline**	**Lock her up!**
trump	trump	world	pipelin	clinton
clinton	poll	people	people	fbi
hillary	percent	war	rock	email
elect	clinton	state	dakota	hillary
president	elect	country	people	investigating
donald	black	russia	north	syria
american	peope	american	oil	comey
people	voter	nation	nativ	elect
obama	vote	new	access	trump
vote	october	power	campaign	united
UMass topic coherence index				
-125.98	-141.36	-152.17	-162.92	-167.13

Table 2: Five most likely topics in the corpus

FBI Reopens Investigation! "Hillary Caught Selling Political Favors to Israel"?

Lock her up! (0.82)	2016 Election (0.054)	Global banking system (0.048)	Jewish conspiracy (0.040)

Homeless TRUMP Supporter Guards Donald Trump's Star on Hollywood Blvd

Trump (0.67)	Blacks like Trump (0.31)		

SHARIA IN AMERICA? How Minneapolis Muslims Are Still Being Recruited

Jews and Muslims (0.65)	Trump (0.26)		

Word perspective

Finally, similar to the document perspective, we can find topic associated with a specific word. This is especially interesting for search, as it allows us to build a system that finds documents based on a shared, underlying concept instead of simple metrics like *"is this word contained in the document"*. The term *sick* for example is associated with topics about Hillary Clinton, Obamacare and health scares.

5 Discussion

I have found two main problems when working with LDA topic models. Due to the non-deterministic nature of the model re-runs with the same number of topics and the same corpus produced quite different results. What made this worse was the fact that the found topics were sometimes a lot harder to interpret. Fixing the random state of the model beforehand also held the results constant but nonetheless some trial and error was necessary to generate easy to interpret results. This behavior could, however, be due to the large number of tokens in the vocabulary compared to the corpus size. Another problem, as mentioned in the results, was that a lower number of topics does not necessarily produce the "useful subset" of a model with more

topics. It seems a certain amount of noise topics is necessary to create clean topics. Despite automated ways of choosing the number of topics it is ultimately still up to the researcher to select which topics carry enough semantic meaning to be useful.

6 Conclusion

In this report I have applied LDA topic modeling to a corpus of fake news articles from late 2016. I have then shown in what ways a LDA model can be used to explore a corpus and have provided examples from the fake news dataset. Despite the drawbacks described in the discussion I was able to gain insight into the composition of the corpus and have found LDA to be a useful tool for this task. The modeling code as well as the dataset can be found on Github `https://github.com/vazamb/data_modeling_lda_project`.

References

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Uncertainty in Artificial Intelligence*.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & Mccallum, A. (2011). Optimizing Semantic Coherence in Topic Models. , 262–272.

Rehurek, R., & Sojka, P. (2010, 5). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the lrec 2010 workshop on new challenges for nlp frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Wallach, H. M. (2006). Topic Modeling: Beyond Bag-of-Words.

Whye Teh, Y., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Source Journal of the American Statistical Association*, *101*(476), 1566–1581. Retrieved from `http://www.jstor.org/stable/27639773http://www.jstor.org/page/info/about/policies/terms.jsphttp://www.jstor.org`

Appendix

Topic	Words
Retired General	forc state retir general peopl trump day said white offic
Health scares	health studi cancer drug use research bodi effect food diseas
Filler words	like time peopl know thing dont children women abort caus
Jewish conspiracy	israel jew isra jewish christian muslim palestinian state said american
Lock her up!	clinton fbi email hillari investig comey elect trump campaign new
Fearmongering	war govern polic kill peopl law forc state mosul said
Russia and America	world peopl war state countri russia polit american nation new
Russia and Syria	russian russia militari war forc syria state syrian said unit
Obamacare	obama american state presid campaign law obamacar feder video hillari
Dakota Pipeline	pipelin water stand rock dakota peopl north oil nativ access
Filler words	peopl year world time live like need state energi water
Election 2016	trump clinton hillari elect presid donald american peopl obama like
Syria	syria war state trump govern russia peopl clinton protest pipelin
The Police	polic time offic law report said children comment like use
Trump	like peopl time year dont thing work american trump way
Voting	vote elect state voter trump ballot report said hillari day
Clinton and Wikileaks	clinton state hillari russia email russian podesta campaign washington wikileak
Global banking system	bank money govern gold financi time new year dollar state
Blacks like Trump	trump poll percent clinton elect black peopl voter vote octob
Clinton emails	clinton email state said news hillari report campaign trump depart