


# Tweets topic modeling.

REPORT FOR THE NATURAL LANGUAGE PROCESSING PRACTICAL APPLICATION OF THE MSc IN DATA SCIENCE SUBJECT OF INTELLIGENT SYSTEMS.

LORENZO ALFARO, DAVID (49217407Y)

ETSIINF, UPM

January, 2022

# Table of contents

# 1. Introduction

The increasingly overwhelming amount of available natural language motivates the pressing need to find efficient and reliable computational techniques capable of processing and analysing this type of data to achieve human-like natural language understanding for a wide range of downstream tasks. Over the last decade, Natural Language Processing (NLP) has seen impressively fast growth, primarily favoured by the increase in computational power and the progress on unsupervised learning in linguistics.

Dealing with unannotated data poses as an unsupervised learning problem, a paradigm within Machine Learning in which no outputs or target variables are provided, i.e., the model is only trained with inputs $x$, $\mathcal{D} = \{x_i\}_{i=1}^{N}$, with the goal being to discover or extract patterns present in data. Often referred to as *knowledge discover* in the literature, this class of problem is much less defined and since it lacks from ground truth, it is not trivial to find efficient metrics to use for the model to learn about relationships in data and to assess goodness of the identified groups.

In this assignment, we use Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a generative probabilistic model for collections of discrete data such as text corpora in which each item is modelled as a finite mixture over an underlying set of topics. In this context, we attempt to obtain meaningful representations of documents according to the topic probabilities yielded by a LDA model.

# 2. Problem description

For this work we will use a dataset composed of 1.6 million tweets, first introduced (Go et al., 2009). Originally proposed for sentiment classification, in this assignment we are mainly concerned with the text of the messages and not in the label, from which the most prominent topics will be captured via LDA.

LDA constitutes one of the main approaches for topic modeling. It is roughly based on these two principles:

- Every document is a mixture of topics. A document may contain words belonging for several topics. LDA assigns documents to topics in a probabilistic fashion, i.e., membership to topics is inferred from the likelihood estimates.

- Every topic is a mixture of words. A topic can be characterized by a set of words. For example, given the topic of "entertainment", one could think of words such as "music",

"singer", "actress", "movie", etc. This implies that a word may belong to more than one topic.

# 3. Methodology

The programming language used throughout the assignment to obtain the hereby described models is *R,* relying on *quanteda* and *spacyr* packages for preprocessing purposes, *textmineR* to induce a topic model from data using LDA and to perform queries at inference time; packages *readr* and *rlist* to serialize information allowing to load (store) data from disk to provide with faster execution times by avoiding repeating calculations; and *plotly*, used for generating interactive graphs to visualize topic assignations for words and documents. Furthermore, we grant reproducibility of all experiments hereby described to ensure deterministic execution of the program.

# 4. Results

Prior to the modeling phase, we divide the original dataset into the training and test datasets, composed of $2 \cdot 10^5$ and $10^5$ instances, respectively, yielded by sampling. Then, we apply some pre-processing steps to the raw tweets to deal with the heterogeneous nature of the data, normally written in an informal register. First, we remove (1) punctuation characters, (2) URLs beginning with *http(s)* and *www*, (3) special symbols, (4) numbers; (5) English stop-words, (6) *ampersand* and *quot* symbols, (7) non-ASCII characters; and (8) twitter usernames. Subsequently, words are converted to lower case and later replaced to their lemma, using English spaCy pretrained models.

Upon pre-processing the tweets, we induce a term co-ocurrence matrix (TCM). The TCM is a square matrix whereby rows and columns are characterized by each of the tokens in the vocabulary of the corpus. The TCM is based on the rationale that terms that appear conjointly across the corpus are prone to be semantically related.

To that end, and accounting for the length of the tweets (short in nature), we follow a term-context approach, where the whole document serves as a context. Another prominent method consists in using window contexts, known as *skip-grams*, which has been found to be very effective in neural language modelling (Mikolov et al., 2013).

Note that the cardinality of the vocabulary increases, correlations between words tend to be more marginal, in the sense that out of all the words, only a small proportion of them are meaningfully related. Therefore, sparse representations are more suitable (the equivalent dense representations have a quadratic order of space complexity). In this

problem, we have a vocabulary size of 68984 tokens, which leads to a TCM with 4,758,792,256 entries.

Once learned the TCM for the training dataset, we fit an embedding model via LDA to find the top twenty topics that capture the most salient information present in data. For such aim, we run 200 iterations of the *Gibbs* sampler (a powerful method for sampling observations, assuming they approximate to an underlying multivariate probability distribution); and by adjusting the value of alpha ($\alpha$) to ensure adaptiveness in the learning process.

The topics identified, their prevalence (density of tokens along each embedding dimension), coherence (degree of semantic similarity between high scoring words in the topic) and the top terms associated to each topic, are reported in Table 1.

**Table 1.** Topics identified in LDA, characterized by their prevalence, coherence and their top terms.

| Topic Id | Prevalence | Coherence | Top terms |
|---|---|---|---|
| t_1 | 3523.064 | 0.263 | day, today, nice, work, rain |
| t_2 | 2717.960 | 0.297 | nt, ca, wait, happy, birthday |
| t_3 | 3996.424 | 0.428 | hurt, feel, hair, bad, back |
| t_4 | 3728.096 | 0.301 | twitter, follow, tweet, read, send |
| t_5 | 3468.122 | 0.352 | miss, friend, fun, girl, love |
| t_6 | 4092.678 | 0.251 | work, find, iphone, phone, update |
| t_7 | 3178.196 | 0.299 | feel, bad, school, hate, sick |
| t_8 | 3382.311 | 0.358 | show, tonight, live, meet, night |
| t_9 | 3224.035 | 0.294 | work, night, time, sleep, bed |
| t_10 | 3538.090 | 0.292 | love, song, lt, hey, awesome |
| t_11 | 2956.934 | 0.359 | good, hope, great, morning, day |
| t_12 | 3070.464 | 0.259 | buy, day, year, pay, money |
| t_13 | 3700.561 | 0.172 | hehe, la, sa, ko, de |
| t_14 | 4149.243 | 0.305 | lol, haha, yeah, nt, ur |
| t_15 | 2720.268 | 0.357 | na, gon, play, game, lose |
| t_16 | 3248.841 | 0.367 | car, house, break, run, sit |
| t_17 | 3303.145 | 0.320 | watch, movie, sad, love, tv |
| t_18 | 3707.799 | 0.322 | eat, make, food, good, dinner |
| t_19 | 3814.527 | 0.296 | make, thing, people, life, lot |
| t_20 | 3462.242 | 0.345 | back, home, week, work, leave |

As it can be seen, some of the topics are semantically coherent. For instance, topic 9 is characterized by words *work, night, time, sleep* and *bed*, where one could expect people

tweeting about how they need to have some rest after a long day of work. Similarly, topic 10 is characterized by words *love, song, lt, hey* and *awesome*. We argue words *love*, and *awesome* are usually employed to describe songs. On the other hand, the top 5 most meaningful words for topic 4 are *twitter, follow, tweet, read* and *send,* which, yet again, one could find natural to find them in such context.

To further prove our point, we conducted several experiments to test whether similar words to those characterizing the topics were attributed to the proper topic, finding that word *college* is related to topic 7 (feel, bad, school, hate, sick), and topic topic 20 (back, home, week, work, leave); word *tired* is associated to topic 9 (work, night, time, sleep, bed); word *lunch* is associated to topic 18 (eat, make, food, good, dinner); word *android* is related to topic 6 (work, find, iphone, phone, update); and that word disease is associated to topic 3 (hurt, feel, hair, bad, back).

Eventually, we embed documents from the test set under the LDA model to obtain topics of local contexts. To examine the per-document-per-topic probabilities, we first obtain a document-term matrix, which describes the frequency of terms that occur in a collection of documents, i.e., a standard bag-of-words representation.

When querying for tweets talking about *Harry Potter*, the model identifies the main topic for all of them is topic 17 (watch, movie, sad, love, tv). Moreover, when querying for random documents, we find that, in general, the model is able to capture salient information about the documents whenever they can be somehow directly linked to some of the top $k$ words of a topic, whereas for those that cannot be associated clearly to any topic the per-topic probability estimates are uniformly distributed.

# 5. Discussion

One of the main issues is how to determine the optimal number of topics, denoted $k^*$ to be found. In this work, we simply took 20, albeit it is arguably suboptimal due to the large cardinality of the training set. There are some approaches that provides efficient heuristically-driven methods to tailor down the problem search space, e.g., via density-based model selection (Cao et al., 2009). In (Arun et al., 2010) some of the main approaches to find $k^*$.

Furthermore, because tweets are constraint in length, it is hard to find messages covering different topics. We argue that it would be convenient to, once learned the model from a batch of tweets, to perform inferences on larger tests to further test whether the model is able to capture the topics in a more meaningful fashion.

# 6. Conclusion

Throughout this work we have found LDA to be very effective onto finding prominent topics in vast amounts of raw, unannotated data. However, because words are treated as mere indices, the model fails to resolve ambiguities inherent to natural language (e.g., lexical, syntactic, semantic and pragmatic ambiguities).

Furthermore, the notion of similarity between words is solely based on words appearing in similar contexts, and not in the intrinsic meaning of each word (e.g., the semantic information a word conveys, given the context in which it appears), where Neural Natural Language Processing, which favours from using Deep Learning models to learn language representations, poses as a better alternative, e.g., BERT (Devlin et al., 2018), ELMo (Peters et al., 2018) and some off-the-likes.

# References


Arun, R., Suresh, V., Madhavan, C. E. V., & Murty, M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6118 LNAI*(PART 1), 391–402. https://doi.org/10.1007/978-3-642-13657-3_43

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

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/J.NEUCOM.2008.06.011

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. http://arxiv.org/abs/1810.04805

Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report, Stanford*, *1*(2009), 1–12.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, January 16). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://arxiv.org/pdf/1301.3781.pdf

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *NAACL HLT*

*2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 2227–2237. https://doi.org/10.18653/v1/n18-1202