



UNIVERSITY OF BIRMINGHAM

SCHOOL OF COMPUTER SCIENCE
COLLEGE OF ENGINEERING AND PHYSICAL SCIENCES

MSc. PROJECT

Analysing news articles about Russia's war on Ukraine using Latent Dirichlet Allocation based topic modelling

Submitted in conformity with the requirements
for the degree of MSc. Artificial Intelligence and Machine Learning
School of Computer Science
University of Birmingham

Nelson Quintanilla Castro
Student ID: 2291960
Supervisor: Dr Mohammed Bahja

September 2022

MSc. Project

Analysing news articles about Russia's war on Ukraine using Latent Dirichlet Allocation based topic modelling

Nelson Quintanilla Castro

Contents

Table of Abbreviations

List of Figures

List of Tables

1	Overview	1
1.1	Abstract	1
1.2	Acknowledgements	1
2	Introduction	2
3	Literature Review	3
3.1	Subsection	3
3.2	Subsection	3
4	Methodology	4
4.1	Data Extraction	4
4.2	Data Pre-Processing	4
4.2.1	Dataset Relevant Features	4
4.2.2	Normalisation and Tokenisation	4
4.2.3	Stop Word Removal	5
4.2.4	Lemmatisation	5
4.2.5	Transforming the documents in a vectorised form (Bag of Words Representation)	5
4.3	Model Training and Parameter Tunning	5
4.4	Model Evaluation and Selection	7
4.4.1	Visualising topic models	7
4.4.2	Model Perplexity and Coherence	7
4.4.3	Coherence	7
4.4.4	Perplexity	7
4.4.5	Using trained models on sample of documents	7
4.4.6	Human Judgement	7
4.4.7	Word intrusion	7
4.4.8	Topic intrusion	7
4.5	Topics Interpretation and Visualisation	8
5	Results	9
5.1	Dataset	9
5.2	LDA model parameters	9
5.3	Experimental comparison between coherence and perplexity	9
5.3.1	Coherence	9
5.3.2	Dataset	9
5.4	Experimental comparison using human judgement	9
6	Discussion	11
6.1	Subsection	11
6.1.1	Subsubsection	11
7	Conclusion	12
7.1	Subsection	12
7.2	Subsection	12
7.2.1	Subsubsection	12
	References	13
8	Appendix One: Accompanying Archive and Instructions	15
8.1	Directory Structure	15
8.2	Running the Provided Code	15

Table of Abbreviations

- LDA – Latent Dirichlet Allocation
- API – Application Programming Interface
- NLTK – Natural Language Toolkit
- BOW – Bag of Words

List of Figures

1	<i>Diagram that summarises steps taken during the development of our work.</i>	4
2	<i>LdaMulticore function from Gensim to train or LDA model.</i>	6
3	<i>Results</i>	9
4	<i>Charts</i>	10

List of Tables

1	<i>Results: word intrusion and topic intrusion experiments</i>	9
---	--	---

1 Overview

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1.1 Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1.2 Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

2 Introduction

The ongoing conflict between Russia and Ukraine is a topic of pivotal importance not only for the protagonist's regions but also for the whole world. This war is already causing many negative economic, political, and social consequences on a different scale for many countries (Caldara et al. 2022). As a result, this has become of paramount concern for governments and, subsequently for all types of media.

This is why there is currently a tremendous amount of information that has been generated about the Russo-Ukrainian conflict and there is no systematic way to analyse this corpora to examine what is currently happening regarding this conflict, to discover the main themes contained within it, and possibly draw conclusions about it.

At the same time, it is not humanly possible to read and study large quantities of articles with this thematic in a short period of time. To do this, experts in machine learning have created probabilistic topic modelling, a collection of several algorithms which are used to identify and interpret sizeable sets of documents that contain topical information. We can use topic modelling to identify the primary ideas that run through a sizable and otherwise unstructured collection of papers. We can arrange the collection in accordance with the themes found by applying topic modelling (Blei 2012).

For the purpose of this work, latent dirichlet allocation (LDA) is going to be the topic modelling algorithm of choice. A collection of texts can be broken down into their key topics using LDA, where a topic is defined as a probability distribution over a vocabulary (Blei n.d.).

A set number of topics are proposed by LDA, and it is assumed that each document in a collection of documents reflects a combination of those themes. Under these presumptions, probabilistic inference techniques identify an embedded theme structure in a document collection. LDA offers a technique to easily summarise, peruse, and search huge document collections with this structure (Blei et al. 2010).

The main motivation of this work is to capture the main themes or topics that can be found on news articles available on the internet that cover events related to Russia's war on Ukraine. This represents an important topic to study because there is a need to derive insightful information from this vast amount of news about the war that are being generated in mass all over the world.

Furthermore, it is of utter importance to observe and detect significant patterns that are being communicated to the general public through news articles so we can predict consequences of the current state of the situation on the short term such as the economical, political, and social impact on human lives.

In this paper we provide an analysis of news articles retrieved from The Guardian newspaper that cover events related to the Russian invasion of Ukraine by utilising LDA based topic modelling with the purpose of obtaining abstract topics within a corpora of news articles. We also offer a way to assess the effectiveness and outcomes of our model. Additionally, we present the validation of our model by comparing it against human judgement. In order to allow others to duplicate our findings, we have also made our data sets publicly available. Further to this, we use pyLDAvis to interpret the topics obtained by our model in a more comprehensible and visual way. This software offers a global view of the topics and how they differ from one another while also enabling a close examination of the terms most closely associated with each specific topic (Sievert & Shirley 2014).

The structure of this paper is as follows. In section 3 we examine the most recent studies in this field. In Section 4 we present the specifics of our approaches, including a description of data extraction, preprocessing, parameter tuning of our model, and evaluation metrics implemented. We provide our findings and engage in a discussion in Section 5. The discussion is presented in part 6, and our conclusion is presented in section 7.

3 Literature Review

3.1 Subsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

3.2 Subsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

4 Methodology

In this section we explain the steps we took in order to build and prepare our dataset, the model generation and parameter tuning process, methods of evaluation for our model, and lastly visualisation and interpretation of our resulting topics. The Figure 1 summarises the method we just mentioned.

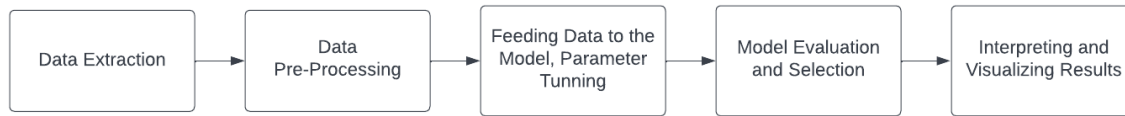


Figure 1: *Diagram that summarises steps taken during the development of our work.*

4.1 Data Extraction

To obtain a significantly large dataset of reliable news articles related to the Russian invasion of Ukraine we decided to use The Guardian digital newspaper. The reasoning behind this decision is because The Guardian is one of the most trusted newspapers in the UK and Europe in terms of objective journalism (Matsa et al. 2018). Additionally, they provide The Guardian – Open Platform, which is their public web service or application programming interface (API) that provides free access to all of the content that the Guardian produces, organised by tags and sections (*theguardian / open platform - documentation / overview* n.d.). This is specially useful in our case because we need to filter the news to reduce noise in our data and have articles relevant to the ongoing conflict between Russia and Ukraine.

To narrow-down the retrieval of the articles, we requested published data between the 2nd of February of 2022 (day in which Russia invaded Ukraine) and the 31st of August of 2022. Additionally, we created a function to interact with the API and retrieve news articles filtered by search term and section. This is to avoid obtaining non relevant news; for instance, we used the section "world" to point to data classified in this category and not other categories such as "sports" or "lifestyle". Furthermore, we used 2 search terms: Ukraine and Russia. We obtained 2 different datasets corresponding to the mentioned search terms. After that, we merged both datasets and removed duplicates that could have existed in both to avoid redundancy. We also removed non pertinent articles that contained information about topics not related the conflict.

The resulting dataset is a json containing a list of 2275 news articles with their id, title of the article, body of the article, and url to read the article on the web. Such dataset is available in the repository included in Section 8 so it can be used for future research or work.

4.2 Data Pre-Processing

Before feeding the data to our model it is required to prepare the data (Jacobi et al. 2016).

Following is a detailed description of the processes we took when completing the analysis for our case study.

4.2.1 Dataset Relevant Features

Before starting pre-processing the data we display important features from our dataset, such as the number of articles to be used and the average word count per article. This is just to have a rough idea of the length of each document we are using for the purposes of training our model. In our case, we have a total of 2275 news articles and an average of 864 words per article.

4.2.2 Normalisation and Tokenisation

A topic model does not analyse documents directly; rather, it employs a so-called document-term matrix that is derived from the documents. This matrix displays the frequency for each term (word) in each document (Jacobi et al. 2016). The initial step in the creation of this matrix is normalisation and tokenisation, for which each page is converted into a list of lowercase words, ignoring short and long words.

4.2.3 Stop Word Removal

Stop words are the words in a stop list that are filtered out prior to processing natural language data (text) due to their insignificance (Rajaraman & Ullman 2011).

We compared three standard lists of stop words from extensively used libraries for machine learning applications: Natural Language Toolkit (NLTK) (Bird n.d.), scikit-learn (Pedregosa et al. n.d.), and spaCy (Honnibal & Montani 2017). There is a file dedicated to do this comparison on our repository included in Section 8. However, it is worth mentioning that our final decision was to merge the lists of stop words since this leads to better results in terms of removing words that are not adding any meaning, for instance "but", "end", and "the".

4.2.4 Lemmatisation

Lemmatisation is an important part of the preprocessing of the data. It reduces all words to their "lemma" by the use of a lexicon and regular conjugation rules. Thus, lemmatization reduces "is" and "were" to their respective lemma (to be) "be" (Jacobi et al. 2016). There is also another technique called stemming which removes the endings of the words leaving just the "stems". Using the commonly employed Porter stemming algorithm, for instance, "weaknesses" becomes "weak" while "failures" and "failure" both become "failur" (Jacobi et al. 2016).

In our case lemmatisation was preferred over stemming because it produced better results as compared to stemming which sometimes returned words that were missing letters at the end and therefore did not exist in the English dictionary.

For the analysis presented here we used the lemmatiser from spaCy (Honnibal & Montani 2017) which assigns tokens' base forms using criteria derived from part-of-speech tags (POS).

4.2.5 Transforming the documents in a vectorised form (Bag of Words Representation)

As we mentioned earlier, before we feed the data to the model we need to transform the documents into a document-term matrix. This way of expressing our dataset is known as Bag of Words (BOW) representation. BOW represents the content of the documents by showing the frequency of the word on every document. Once we have this, we are ready to start training our model.

After using the aforementioned strategies, we have removed a number of terms from our initial tokenised dataset and adjusted the remaining words' structure to make them more effective for training our model. In the following section, we are going to delve into the model training and tuning.

4.3 Model Training and Parameter Tuning

There are numerous tools available for Topic modelling and analysis, including professional and amateur software, commercial and open-source software. Some examples of these well known tools that have been used by a large number of scholars in the past are Mallet (McCallum 2002), TMT (Ramage et al. 2009), and Gensim (Jelodar et al. 2019).

For the training of our model we used Gensim, which is a Python package that provides algorithms for Latent Dirichlet Allocation that are memory-efficient, scalable, and quick (Rehurek & Sojka 2010, Řehůřek & Sojka n.d.).

In order to train our LDA model we need to provide it with different parameters as it can be observed in Figure 2. From all the parameters, one of the most important ones is the number of topics we want to extract from our dataset as this one affects greatly on the quality of the results. Furthermore, choosing the number of topics in LDA is typically determined by assessing the perplexity of numerous models built with varying values of number of topics as a parameter and choosing the one with the lowest perplexity value (Blei et al. 2010).

Perplexity is a metric for evaluating language models, where a low score denotes a model with superior generalisation, as determined by (Asmussen & Møller 2019, Blei & Lafferty 2007, Xu & Raschid 2016, Zhao et al. 2015).

Reducing the perplexity score is equivalent to increasing the chance of all articles belonging to a topic. The criteria for picking the optimal number of topics is to strike a balance between a manageable number of topics and the lowest possible level of perplexity. The optimal amount of topics can vary substantially depending on the purpose of the analysis. As a general guideline, fewer topics are utilised for a general overview, while more topics are used for a more detailed look

(Asmussen & Møller 2019). In our case, we geared our experimentation towards obtaining a general overview.

COHERENCE...

Having mentioned that, once we generated our corpus or Bag of Words representation of our dataset and our dictionary (a dictionary object that contains every single word on the entire dataset without duplicates), we created a function to generate many models at once with a different value of number of topics.

The first run we did was for models with 2, 8, 14, 20, 26, 32, and 38 number of topics. As we will explain further, we did several runs with different number of topics. At this point we computed the coherence and perplexity for each of the models generated and as it can be observed in figure ??, We assessed the calculated perplexity and coherence and decided to do a second run for models with 2, 4, 6, 8, 10, 12, and 14 number of topics.

We assessed the coherence and perplexity values, as well as the associated words with the highest probability for each of the topics of the generated models and based on that we started adding custom words to our list of stop words. The purpose of this is to avoid getting some of the words repeated in many topics (this is the case for words that occur in the majority of the documents). Therefore, we removed the following words: Ukraine, Ukrainian, Russia, Russian; and ran our model a third time.

We assessed again the metrics in addition to the associated words for each of the topics of the model generated and this time we removed the following words: war, people, city; and ran our model a fourth time.

Lastly, we assessed in a similar manner, removed the words: country, year, day; and ran our model a fifth time.

For each of the runs in the process previously described we also did several other runs to tune the remaining parameters of the model by trial and error. When the results were better, we continued with the next iteration of runs.

```
lda_model_ = gensim.models.ldamulticore.LdaMulticore(
    corpus=corpus_,
    num_topics=num_topics_,
    id2word=id2word_,
    workers=None,
    chunksize=chunksize,
    passes=passes,
    alpha=alpha,
    eta=eta,
    decay=decay,
    offset=offset,
    eval_every=eval_every,
    iterations=iterations,
    gamma_threshold=gamma_threshold,
    minimum_probability=minimum_probability,
    random_state=random_state,
    minimum_phi_value=minimum_phi_value,
    per_word_topics=per_word_topics,
    dtype=dtype
)
```

Figure 2: *LdaMulticore* function from *Gensim* to train or *LDA* model.

4.4 Model Evaluation and Selection

4.4.1 Visualising topic models

4.4.2 Model Perplexity and Coherence

4.4.3 Coherence

4.4.4 Perplexity

4.4.5 Using trained models on sample of documents

4.4.6 Human Judgement

4.4.7 Word intrusion

4.4.8 Topic intrusion

4.5 Topics Interpretation and Visualisation

5 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Experiment	Model	Correct Answers	Total Answers	Ratio
Word intrusion	Model 1	13	50	0.260
	Model 2	10	50	0.200
	Model 3	14	50	0.280
Topic Intrusion	Model 1	16	30	0.533
	Model 2	17	30	0.567
	Model 3	13	30	0.433

Table 1: *Results: word intrusion and topic intrusion experiments*

Word Intrusion

pen
price
le
party
government
macron

☐ A. pen
☐ B. price
☐ C. le
☐ D. party
☐ E. government
☐ F. macron

Topic Intrusion

Please read the following article and then answer the following 3 questions about it. The URL to the article if you prefer to read it on the web: <https://www.theguardian.com/world/2022/apr/21/russia-using-banned-weapons-to-kill-ukrainian-civilians-pictures-suggest>

Russia using cluster bombs to kill Ukrainian civilians, analysis suggests

Russian troops have used a number of weapons widely banned across the world, which have killed hundreds of civilians in the Ukrainian region of Kyiv, research by the Guardian suggests.

Evidence collected during a visit to Bucha, Hostomel and Borodianka, where Russian occupiers have been accused of atrocities against residents, showed that Russian troops had used cluster munitions, cluster bombs and extremely powerful unguided bombs in populated areas, which have destroyed at least eight civilian buildings.

Choose the group of words which **does not** belong with the article (Model * 1).

☐ A. israel, world, israeli, soviet, medvedchuk, medvedev, nazi, palestinian
☐ B. time, work, like, know, tell, putin, family, man
☐ C. force, kyiv, civilian, military, mariupol, attack, soldier, kill

Figure 3: *Results*

5.1 Dataset

5.2 LDA model parameters

5.3 Experimental comparison between coherence and perplexity

5.3.1 Coherence

5.3.2 Dataset

5.4 Experimental comparison using human judgement

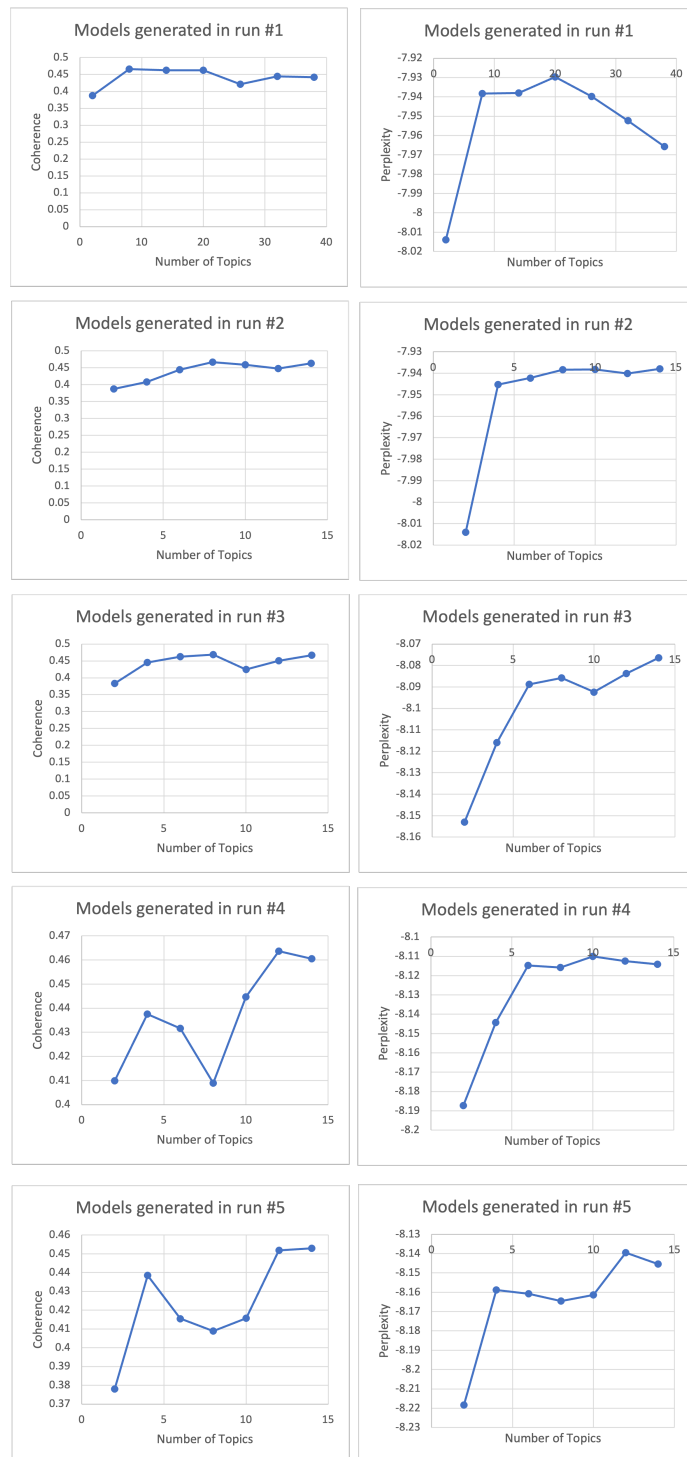


Figure 4: *Charts*

6 Discussion

6.1 Subsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

6.1.1 Subsubsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

7 Conclusion

7.1 Subsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

7.2 Subsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

7.2.1 Subsubsection

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

References

- Asmussen, C. B. & Møller, C. (2019), ‘Smart literature review: a practical topic modelling approach to exploratory literature review’, *Journal of Big Data* **6**(1), 1–18. Number: 1 Publisher: SpringerOpen.
URL: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0255-7>
- Bird, S. (n.d.), ‘Natural Language Processing with Python’, p. 504.
- Blei, D., Carin, L. & Dunson, D. (2010), ‘Probabilistic Topic Models’, *IEEE Signal Processing Magazine* **27**(6), 55–65. Conference Name: IEEE Signal Processing Magazine.
- Blei, D. M. (2012), ‘Probabilistic topic models’, *Communications of the ACM* **55**(4), 77–84.
URL: <https://dl.acm.org/doi/10.1145/2133806.2133826>
- Blei, D. M. (n.d.), ‘Latent Dirichlet Allocation’, p. 30.
- Blei, D. M. & Lafferty, J. D. (2007), ‘A correlated topic model of Science’, *The Annals of Applied Statistics* **1**(1), 17–35. Publisher: Institute of Mathematical Statistics.
URL: <https://projecteuclid.org/journals/annals-of-applied-statistics/volume-1/issue-1/A-correlated-topic-model-of-Science/10.1214/07-AOAS114.full>
- Caldara, D., Conlisk, S., Iacoviello, M. & Penn, M. (2022), ‘The Effect of the War in Ukraine on Global Activity and Inflation’.
URL: <https://www.federalreserve.gov/econres/notes/feds-notes/the-effect-of-the-war-in-ukraine-on-global-activity-and-inflation-20220527.html>
- Honnibal, M. & Montani, I. (2017), spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Jacobi, C., van Atteveldt, W. & Welbers, K. (2016), ‘Quantitative analysis of large amounts of journalistic texts using topic modelling’, *Digital Journalism* **4**(1), 89–106.
URL: <http://www.tandfonline.com/doi/full/10.1080/21670811.2015.1093271>
- Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. & Zhao, L. (2019), ‘Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey’, *Multimedia Tools and Applications* **78**(11), 15169–15211.
URL: <http://link.springer.com/10.1007/s11042-018-6894-4>
- Matsa, K. E., Silver, L., Shearer, E. & Walker, M. (2018), ‘Western Europeans Under 30 View News Media Less Positively, Rely More on Digital Platforms Than Older Adults’.
URL: <https://www.pewresearch.org/journalism/2018/10/30/western-europeans-under-30-view-news-media-less-positively-rely-more-on-digital-platforms-than-older-adults/>
- McCallum, A. K. (2002), Mallet: A machine learning for language toolkit.
<http://www.cs.umass.edu/mccallum/mallet>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A. & Cournapeau, D. (n.d.), ‘Scikit-learn: Machine Learning in Python’, *MACHINE LEARNING IN PYTHON* p. 6.
- Rajaraman, A. & Ullman, J. D. (2011), *Mining of Massive Datasets*, Cambridge University Press, USA.
- Ramage, D., Rosen, E., Chuang, J., Manning, C. D. & McFarland, D. A. (2009), Topic modeling for the social sciences, in ‘Workshop on Applications for Topic Models, NIPS’.
URL: <http://vis.stanford.edu/papers/topic-modeling-social-sciences>
- Rehurek, R. & Sojka, P. (2010), Software framework for topic modelling with large corpora, in ‘IN PROCEEDINGS OF THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS’, pp. 45–50.

- Sievert, C. & Shirley, K. (2014), LDAvis: A method for visualizing and interpreting topics, *in* ‘Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces’, Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 63–70.
URL: <http://aclweb.org/anthology/W14-3110>
- theguardian / open platform - documentation / overview* (n.d.).
URL: <https://open-platform.theguardian.com/documentation/>
- Xu, Z. & Raschid, L. (2016), Probabilistic Financial Community Models with Latent Dirichlet Allocation for Financial Supply Chains, *in* ‘Proceedings of the Second International Workshop on Data Science for Macro-Modeling’, DSMM’16, Association for Computing Machinery, New York, NY, USA, pp. 1–6.
URL: <https://doi.org/10.1145/2951894.2951900>
- Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y. & Zou, W. (2015), ‘A heuristic approach to determine an appropriate number of topics in topic modeling’, *BMC Bioinformatics* **16**(13), 1–10. Number: 13 Publisher: BioMed Central.
URL: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S13-S8>
- Řehůřek, R. & Sojka, P. (n.d.), ‘Gensim—Statistical Semantics in Python’, p. 1.

8 Appendix One: Accompanying Archive and Instructions

8.1 Directory Structure

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

8.2 Running the Provided Code

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.