

**Improving the space efficiency of the MS2LDA.org web application**

Radu Petrescu

School of Computing Science

Sir Alwyn Williams Building

University of Glasgow

G12 8RZ

A dissertation presented in part fulfillment of the requirements of the Degree of Master of Science at the University of Glasgow

September, 2019

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic form. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: Radu Petrescu Signature: *Radu Petrescu*

Acknowledgements

S, J, J, – will be added at the end

Abstract

Interesting blablabla

Contents

Chapter 1 Introduction 1

1.1 Context 1

1.2 Aim 1

1.3 Structure 1

Chapter 2 Analysis & Requirements 2

2.1 Statement of Problem & Motivation 2

2.2 Background 3

2.2.1 Latent Dirichlet Allocation 3

2.2.2 LDA in Mass Spectrometry & Metabolomics 6

2.3 Gathering Requirements & Feature List 7

2.4 Surveyed literature and tools 7

Chapter 3 Design & Implementation 8

Chapter 4 Testing & Evaluation 9

Chapter 5 Conclusion 10

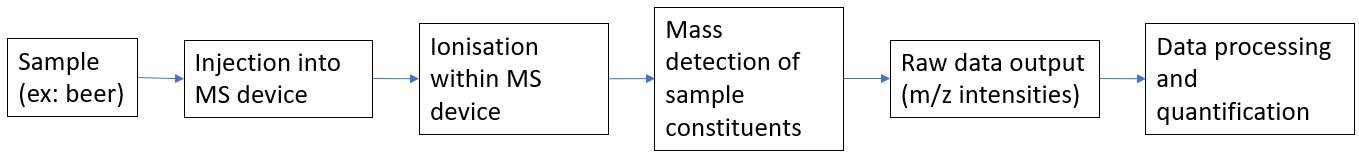
Chapter 6 Bibliography 11

Appendix A <Name of appendix> 1

# Introduction

## Context

Mass spectrometry (hereby “MS”) is a technique for measuring the mass-to-charge (“m/z”) ratio of the ionised particles of a chosen sample.[[1]](#endnote-1) The generated spectra signals are used to identify the presence and characteristics of elements and molecules in the sample. [[2]](#endnote-2) A simple diagram of this workflow is presented below:



As an analytical technique, MS is essential in numerous areas such as pharmaceutical drug development [[[3]](#endnote-3)], metabolomics [[[4]](#endnote-4)] and environmental testing [[[5]](#endnote-5)] to name a few. Taking into consideration the diagram above, this project is only concerned with the final stage, namely the data processing and quantification of the raw output from the MS device (hereby referred to as “*in silico* processing”). *In silico* processing is currently done through the web application MS2LDA.org that maps the masses of individually identified sample constituents into motifs.[[6]](#endnote-6) Motifs, concisely, are visual representations showing a pattern of related constituents, as shown in the following diagram:[[7]](#endnote-7)



The application utilises Latent Dirichlet Allocation (hereby “LDA”), a text mining algorithm that will be further expanded upon in the following sections.[[8]](#endnote-8) MS2LDA.org has a corresponding connected database named MotifDB, which stores motifs and corresponding data.[[9]](#endnote-9)

## Aim

Although the functionality of MS2LDA.org is not contested, the web application is heavily resource intensive in terms of the non-volatile memory required.

**The aim of this project is to modify and enhance the MS2LDA.org web application in order to increase its space efficiency and reduce the amount of non-volatile memory consumed on its host server.**

## Structure

The report is structured into the following main sections:

1. Analysis & Requirements;

2. Design & Implementation;

3. Testing & Evaluation;

# Analysis & Requirements

## Statement of Problem & Motivation

This report will hereby refer to the owners of the MS2LDA.org application as “the Client”.

The Client currently has a functionally working version of the web application available online at ms2lda.org. The application runs on a server with limited space resources that are increasingly running scarce due to the significant amount of data produced by the application. This is the Client’s problem.

The solution to this problem consists in refactoring the system so that it consumes less space resources without impacting the functionality of the web application. The functionality of the application implies both that its behaviour is unchanged and that its computational speed is not significantly reduced.

As it will be further explained in the following sections of the report, much of the data stored by the web application on the server can be computed at run time and therefore needs no long-term storage on the server. This idea will be leveraged in producing the desired solution for the Client.

MS2LDA.org deals with large mass spectrometry data sets from experiments in the field of metabolomics.[[10]](#endnote-10) This field is concerned with the study of metabolomes, small molecular cellular end-products.[[11]](#endnote-11) There is a growing interest in combining metabolomics data sets from multiple sources in order to yield comprehensive scientific insights and possibly new discoveries.[[12]](#endnote-12) The main bottlenecks to such developments are the general lack of integrative analytical tools and software[[13]](#endnote-13) as well as the lack of labelled data.[[14]](#endnote-14) In particular, the difficulty of obtaining labelled data is significantly influenced by the fact that data sets can take a great amount of hard disk space due to the multitude of data points in the samples (samples can easily reach millions of data points).[[15]](#endnote-15)

The main motivation of this project is therefore not only concerned with reducing the space burden for a specific server and satisfying the Client, but with providing insights into ways that the space efficiency of data sets and metabolomics tools/software can benefit the scientific community at large and improve the likelihood of advancements in metabolomics, related fields and life sciences in general.

One of the motivations of MS2LDA.org upon its creation was to tackle the abovementioned bottlenecks by providing a method for molecular fragment data set decomposition in the field of metabolomics.[[16]](#endnote-16) The web application is not merely a prototype, but a real functioning system where interested parties from the scientific community can upload their own mass spectrometry data, run analyses and view relevant visualisations of the results. Enhancing the web application therefore is intended to provide a focused benefit not only to the direct Client, but also to its vast user group in the scientific and medical community.

The following section conveys the relevant theoretical background to the project.

## Background

### Latent Dirichlet Allocation

MS2LDA.org uses the text mining algorithm known as LDA adapted for the fields of mass spectrometry and metabolomics. Understanding LDA is crucial for the scope of this project since comprehending the underlying algorithm allows for a more space efficient computation of the relevant data.

LDA is a generative probabilistic model that is commonly used, among other areas, for topic modelling in text corpora.[[17]](#endnote-17) A text corpus is defined as a large and unstructured collection of text documents containing discrete text data. Each document, the sub-constituent of the corpus, is a collection of words. The word can be defined as the atomic discrete unit of text data. The set of all the words in the corpus is referred to as the vocabulary. LDA assumes that documents are distributions over topics, and that topics are distributions over words. A topic is an abstract concept representing the distribution over words. Put differently, words have a certain probability of being part of a specific topic.[[18]](#endnote-18) This is best illustrated through an example: in a corpus, words such as “football” and “championship” are more likely to be belong to the topic “sports” as opposed to the words “cat” and “dog”, the latter having a high probability of being part of the “animals” topic. It must be noted that topics are considered latent/hidden in the LDA model and therefore specific labels such as “sports” are for illustration purposes only.

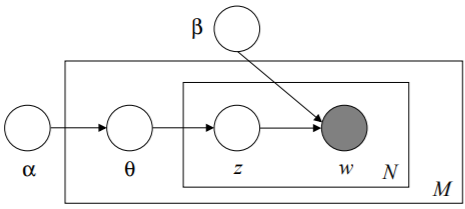
The LDA model assumes that the number of topics is known, fixed and finite. This helps us determine the hyperparameter known as β (Beta). A hyperparameter is simply a parameter applicable to the entire corpus, while Beta is the topic distribution over words and is represented as a matrix where each row is a topic, each column is a unique word in the vocabulary and their intersection cell is the respective probability. Each row must be normalized, summing up to the value of 1. The following figure represents an example Beta matrix for a hypothetical corpus made up of three topics and three words (the grey shaded area is the matrix):

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Word 1** | **Word 2** | **Word 3** |
| **Topic 1** | 0.1 | 0.2 | 0.7 |
| **Topic 2** | 0.2 | 0.2 | 0.6 |
| **Topic 3** | 0.4 | 0.4 | 0.2 |

The second hyperparameter relevant in the Dirichlet context is referred to as the α (Alpha) parameter. The Alpha parameter is sometimes referred to as the concentration parameter as it represents the probability of the topics appearing in the documents of the corpus. Note that Alpha is not a probability *per se* since it can have values larger than 1.[[19]](#endnote-19)[[20]](#endnote-20) Alpha is therefore a vector (a 1-dimensional matrix) with each concentration value corresponding to a specific topic. The following figure illustrates an Alpha vector for a 3-topic corpus (the grey shaded area is the vector):

|  |  |
| --- | --- |
|  | **Alpha (concentration) value** |
| **Topic 1** | 0.5 |
| **Topic 2** | 0.4 |
| **Topic 3** | 0.1 |

Visualising the interconnection between the elements described in this sub-section can be best described using a plate diagram:[[21]](#endnote-21)



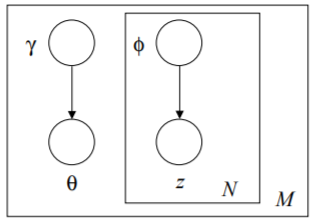
In the above diagram α and β are the abovementioned hyperparameters standing for topic concentration values and topic over word probability distribution respectively. The M plate represents a single specific document. The N plate represents a single specific word in a specific document. The latter plate contains two elements. The first element, w, represents the occurrence of a word. As it can be observed from the arrow directions, each such word occurrence corresponds to a specific topic z (the second element) and to a specific word/topic probability from the Beta matrix. It is known that Alpha is a concentration parameter for all the topics in the corpus, but in order to get the topic probabilities for a specific document, a θ (Theta) parameter is created. The latter is simply a vector of the probabilities of all the topics occurring in a specific document. Assuming a corpus of three topics, specific document M’s Theta vector is illustrated through the gray-shaded area in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Topic 1** | **Topic 2** | **Topic 3** |
| **Document M** | 0.3 | 0.3 | 0.4 |

Note that the Theta vector is normalized as well, meaning that the probabilities for each row must sum to 1. To derive the entire document to topic probability distribution, the respective Theta vectors of all documents in the corpus can be concatenated in order to produce a two-dimensional matrix with documents as rows and topics as columns.

The plate diagram above conveys another important aspect that needs to be taken into consideration. The w parameter is the only one that is grayed since it is the only element that is not hidden.[[22]](#endnote-22) LDA is called “latent” because most of its parameters are hidden and therefore must be inferred. Inference, in this scenario, can be defined as the means to derive conclusions about the unknown parameters of a sample/corpus.[[23]](#endnote-23) There are various methods for inference regarding LDA such as Markov-chain Monte Carlo methods (example: Collapsed Gibb’s sampling), Online (Stochastic) methods or Variational Inference methods.[[24]](#endnote-24) This report will only proceed to go into detail concerning variational inference since this is also used by the MS2LDA.org web application.

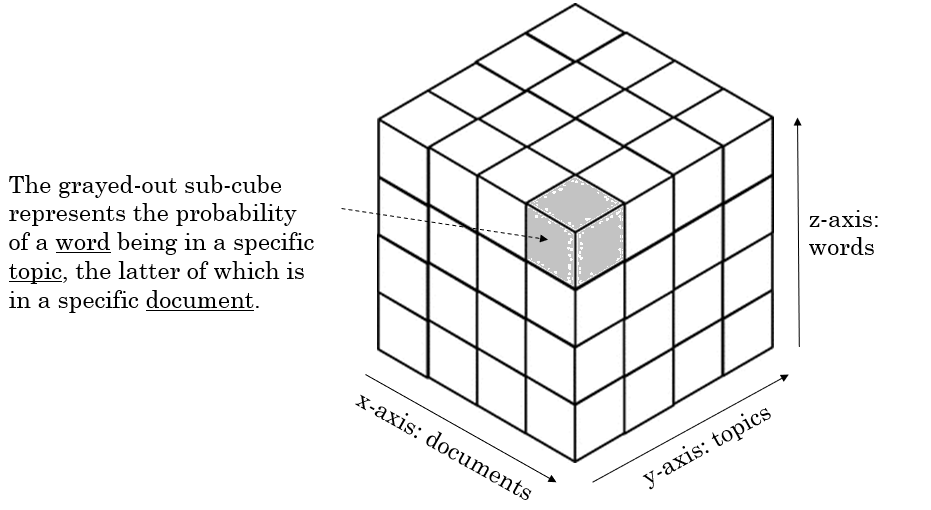
In order to apply variational inference, two variational parameters are needed: γ (Gamma) and φ (Phi). Their relationship to the previous LDA diagram is conveyed in the following secondary plate diagram:



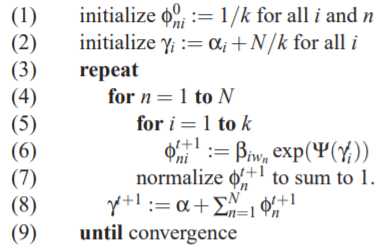
Just like in the first plate diagram above, M is a single document, N is a single word in M, z is the topic corresponding to N and Theta is the distribution of topics per document.

Gamma is very similar to Theta in that it is a vector illustrating the topic incidence for a particular document. The difference lies in the fact that Theta is a normalised version of Gamma. If all the Gamma vectors for all documents are concatenated, a so-called two-dimensional Gamma matrix is created with each row representing a document and each column representing a topic for the given corpus.

Phi is the collection of probabilities of words appearing in topics for specific documents. In other words, phi can be mapped into a three-dimensional matrix resembling a cube composed of substituent smaller cubes, each of the latter representing a single probabilistic value. The x-axis represents documents, the y-axis represents topics and the z-axis represents specific words. This abstraction can be visualized as in the following “cube-of-cubes” diagram:[[25]](#endnote-25)



Variational inference follows a two-step procedure. The first step is called the e-step and involves estimating the two variational parameters Gamma and Phi. The second step is referred to as the m-step and is concerned with re-estimating the hyper-parameters Alpha and Beta given the data in the e-step.[[26]](#endnote-26) This process can be summarised through the following iterative process:[[27]](#endnote-27)



The first two steps initialise the variational parameters Phi and Gamma respectively. Steps 3 to 9 illustrate a nested for loop that repeats until convergence occurs. Convergence occurs when the difference between the values of Beta for two iterations is small enough to be considered insignificant in hindering appropriate inferential approximation. Beta is reset through each of its respective iterations and so is Alpha for its respective iteration. The above algorithm is implemented in the web application through the methods: *init\_vb* (for initialisation), *e-step* (for the above loop, though with slight modifications to perform calculations in log space) and *vb\_step* (the actual resetting/updating of the hyperparameters Alpha and Beta). Ψ above stands for the Psi / Polygamma function.[[28]](#endnote-28)

In order to achieve the objectives of the project, the variational inference steps provide a powerful insight: only the corpus original data and the initial Alpha and Beta parameters are needed to be stored on the hard-disk / in a database. The rest of the parameters can be initialised, computed or updated without having a need for their permanent storage. In other words, the variational parameters Phi and Gamma, as well as the latter’s normalised version (Theta) would only occupy space redundantly if stored in a database. This crucial insight will be further expanded upon in the Design & Implementation section of the report below in order to progress towards achieving the objective.

The next subsection focuses on how mass spectrometry and metabolomics scientific jargon can be mapped to the LDA terminology discussed in this subsection.

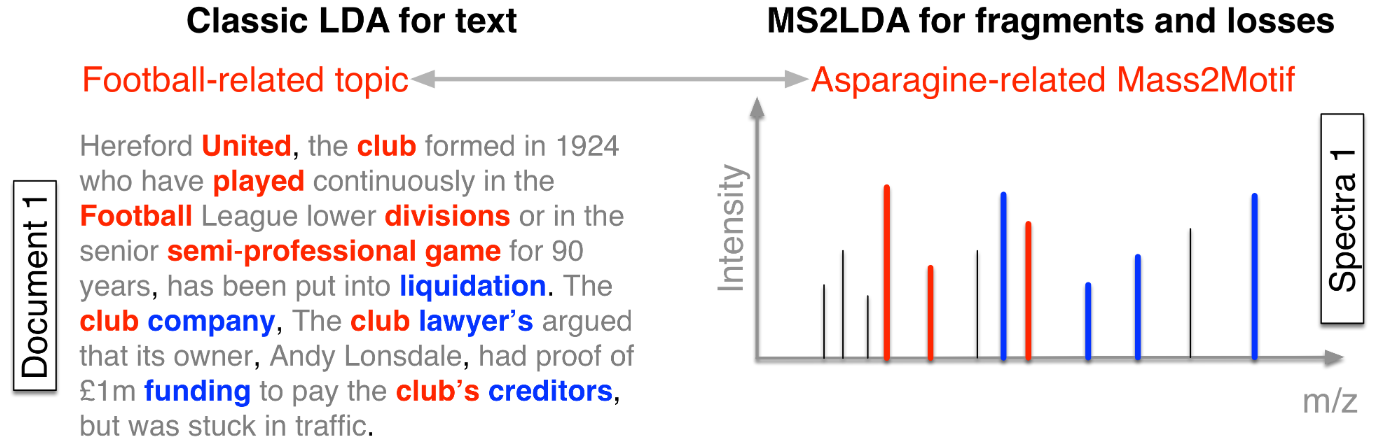
### LDA in Mass Spectrometry & Metabolomics

A brief and simplified introduction to mass spectrometry and metabolomics has been provided in the Introductory sections of this report. As the objective of this report is purely focused on the non-functional requirement of increasing space efficiency, a deep understanding of the actual functional aspects of mass spectrometry and metabolomics is not necessary. However, it is important to be able to map scientific jargon to the LDA terminology presented in the previous subsection.

It can be noticed above that Alpha, Beta, Gamma, Theta and Phi can be expressed through the concepts of corpus, words, topics and documents. The following table provides a mapping to these four key concepts:

|  |  |
| --- | --- |
| **Mass Spectrometry Concept** | **LDA Concept / Term** |
| Feature (or Fragments/Losses) | Word |
| Motif (or Mass2Motif) | Topic |
| Spectrum (or just Document) | Document |
| Experiment / Original data | Corpus |

Visually, the mapping above can also be observed in the following diagram[[29]](#endnote-29):



## Gathering Requirements & Feature List

There are many traditional techniques for gathering requirements such as user interviews, questionnaires and workshops.[[30]](#endnote-30) This project uses a pragmatic tri-partite approach that is described below. There are two important aspects of this report: the software (MS2LDA.org) already exists in a functional state and the objective of the report does not give rise to any functional requirements. Taking these into consideration, the requirements gathering process is composed of:

1. Initial communication with the Client in order to establish the objective and scope of the project (minutes of the meetings are kept in paper form);
2. Periodical (weekly) meetings with the Client in order to clarify any arising issues or in order to perform acceptance testing (minutes of the meeting are kept);
3. Constant contact and communication with the client on an as-required basis using collaboration tools such as Slack for any immediate questions or bottleneck resolutions;

The above requirement gathering process is the most pragmatic and suitable in this context since the Client is in the best position to judge the non-functional successes of the project as well as the success of the preservation of existing functional aspects.

In terms of the actual requirements, there is only a single mandatory must have non-functional feature:

**The web application must use less hard disk space.**

This requirement, however, can be quite vast and through iterations in providing suitable solutions to it one may find that further “should-have” and “could have” requirements result from such iterations. For example, a space-time trade-off is quite common in software and algorithms.[[31]](#endnote-31) As such, given the context, the project takes a pre-emptive approach and is expecting other potential requirements to arise such as:

|  |  |
| --- | --- |
| **Priority** | **Potential Secondary Non-Functional Requirements** |
| Should Have | The web application must not have a significant or noticeable drop in time efficiency. |
| Could Have | The web application uses a caching system in order to support an added value computational time efficiency. |

As it can be observed, potential requirements are adaptive and change with the context. They are also heavily dependent and conditional on the main must-have feature and its implementation. Features have therefore developed with the project, with the main feature in mind. The final version of the ordered list of features, their justification and aims/objectives are presented in appendix 1. **[TBCOMPLETED -include Moscow statements and user stories/constr]**

## Surveyed literature and tools

As MS2LDA.org is a custom-made system and since the objective is to enhance it, competing or similar products have not been heavily scrutinized. A complete list of surveyed literature and software tools is presented in the bibliography.

# Design & Implementation

* Functional aspects lda process
* The learning process for genism,pikle, python tools and so on
* Prototyping Jupyter notebook ot understand LDA code
* Refactoring
* Caching

4-5 pages

System architecture major design decisions and rationale (~2-3 pages) Implementation details ~1 page

Screenshots or other similar things ~1 page

− product well-designed, functional, reliable, robust, efficient, usable, maintainable, and well-documented and demonstrated

# Testing & Evaluation

4-5 pages overall

Software Testing – strategy & statistics 1-2 pages

Explanation of Evaluation Strategy (1/2 page).

Evaluation Results – 1-2 pages (4/5 pages covering each aspect of evaluation and discussion of results.)

Show how you plan to organise your work, identifying intermediate deliverables and dates.

software tested and user evaluation, suggestions for further work

# Conclusion

2-3 pages

Conclusions should state the achievements, and a reflection on achievements (what wasn't achieved/ why and what more could have been done or done differently in hindsight - pick one or two issues to discuss in \*depth\* rather than trying to be comprehensive. should include future work as well.

# Bibliography

* https://onlinelibrary.wiley.com/doi/abs/10.1002/0471238961.1301191913151518.a01.pub2
* https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S7-S9
* https://www.future-science.com/doi/full/10.4155/bio.11.232
* https://onlinelibrary.wiley.com/doi/abs/10.1002/mas.20108
* https://www.sciencedirect.com/science/article/pii/S0021967304021648
* https://github.com/sdrogers/ms2ldaviz ; ms2lda.org
* https://www.ncbi.nlm.nih.gov/pubmed/28968802; ms2lda.org/user\_guide
* http://www.jmlr.org/papers/v3/blei03a.html
* <http://eprints.gla.ac.uk/176662/>

<https://github.com/blei-lab/onlineldavb>

<https://www.youtube.com/watch?v=vYIuvSwLsMc>

tango with Django

<https://github.com/ddbourgin/numpy-ml/blob/master/lda/lda_smoothed.py>

<https://www.machinelearningplus.com/python/numpy-tutorial-part1-array-python-examples/>

<https://www.machinelearningplus.com/nlp/gensim-tutorial/>

<https://mimno.infosci.cornell.edu/info6150/readings/HoffmanBleiBach2010b.pdf>

<https://www.youtube.com/watch?v=fCmIceNqVog>

<https://www.youtube.com/watch?v=2pEkWk-LHmU>

<http://times.cs.uiuc.edu/course/598f16/notes/lda-survey.pdf>

<https://github.com/ChangUk/pyGibbsLDA/tree/master/src>

<http://theses.gla.ac.uk/7928/1/2017wandyphd.pdf>

* <https://www.sciencedirect.com/science/article/pii/S1877705811018674>
* <https://www.thoughtvector.io/blog/lda-alpha-and-beta-parameters-the-intuition/>
* <https://datascience.stackexchange.com/questions/199/what-does-the-alpha-and-beta-hyperparameters-contribute-to-in-latent-dirichlet-a>
* <http://www.jmlr.org/papers/volume18/15-595/15-595.pdf>
* <http://times.cs.uiuc.edu/course/598f16/notes/lda-survey.pdf>
* <https://www.statisticshowto.datasciencecentral.com/posterior-distribution-probability/>
* Wandy 2017 -phd thesis joe <http://theses.gla.ac.uk/7928/1/2017wandyphd.pdf>
* Petre Stegaroiu diss level 4
* Joe Frew diss level 4
* <http://metabolomicssociety.org/resources/metabolomics-software>
* <https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d>
* <https://onlinelibrary.wiley.com/doi/abs/10.1002/0471238961.1301191913151518.a01.pub2>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5860206/>
* <https://www.pnas.org/content/113/48/13738>
* <http://eprints.gla.ac.uk/147960/>
* <https://www.tandfonline.com/doi/full/10.1080/21670811.2015.1093271>
* <https://www.thoughtvector.io/blog/lda-alpha-and-beta-parameters-the-intuition/>
* <http://videolectures.net/mlss09uk_blei_tm/>
* <https://datascience.stackexchange.com/questions/199/what-does-the-alpha-and-beta-hyperparameters-contribute-to-in-latent-dirichlet-a>
* <https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158>
* <https://www.google.com/search?q=what+are+the+latent+variables+in+lda&rlz=1C1CHBF_en-GBGB841RO842&oq=what+are+the+latent+variables+in+lda&aqs=chrome..69i57j33.6560j1j7&sourceid=chrome&ie=UTF-8>
* <https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd>

− list, in alphabetical order by author and date, all articles that you have consulted. For each article, give full bibliographic details

###### <Name of appendix>

<Use Heading 6 for the Appendix heading>

1. https://onlinelibrary.wiley.com/doi/abs/10.1002/0471238961.1301191913151518.a01.pub2 [↑](#endnote-ref-1)
2. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S7-S9> [↑](#endnote-ref-2)
3. <https://www.future-science.com/doi/full/10.4155/bio.11.232> [↑](#endnote-ref-3)
4. <https://onlinelibrary.wiley.com/doi/abs/10.1002/mas.20108> [↑](#endnote-ref-4)
5. <https://www.sciencedirect.com/science/article/pii/S0021967304021648> [↑](#endnote-ref-5)
6. <https://github.com/sdrogers/ms2ldaviz> ; ms2lda.org [↑](#endnote-ref-6)
7. <https://www.ncbi.nlm.nih.gov/pubmed/28968802>; ms2lda.org/user\_guide [↑](#endnote-ref-7)
8. <http://www.jmlr.org/papers/v3/blei03a.html> [↑](#endnote-ref-8)
9. <http://eprints.gla.ac.uk/176662/> [↑](#endnote-ref-9)
10. <https://www.ncbi.nlm.nih.gov/pubmed/28968802> [↑](#endnote-ref-10)
11. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S7-S9> [↑](#endnote-ref-11)
12. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5521158/> [↑](#endnote-ref-12)
13. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5521158/> [↑](#endnote-ref-13)
14. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S7-S9> [↑](#endnote-ref-14)
15. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S7-S9> [↑](#endnote-ref-15)
16. <http://eprints.gla.ac.uk/147960/> [↑](#endnote-ref-16)
17. <http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf> [↑](#endnote-ref-17)
18. <https://guides.library.uq.edu.au/research-techniques/text-mining-analysis/language-corpora> [↑](#endnote-ref-18)
19. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4550296/> [↑](#endnote-ref-19)
20. <https://www.seas.upenn.edu/~cis520/lectures/LDA.pdf> [↑](#endnote-ref-20)
21. Blei’s article [↑](#endnote-ref-21)
22. <https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd> [↑](#endnote-ref-22)
23. <https://www.sciencedirect.com/topics/neuroscience/statistical-inference> [↑](#endnote-ref-23)
24. <http://times.cs.uiuc.edu/course/598f16/notes/lda-survey.pdf> [↑](#endnote-ref-24)
25. <https://etc.usf.edu/clipart/42700/42755/cubes-25_42755.htm> [↑](#endnote-ref-25)
26. <https://www.seas.upenn.edu/~cis520/lectures/LDA.pdf> [↑](#endnote-ref-26)
27. Blei’s article as well as init\_vb from ms2ldaviz repo [↑](#endnote-ref-27)
28. <https://www.mathworks.com/help/matlab/ref/psi.html> [↑](#endnote-ref-28)
29. <https://www.pnas.org/content/113/48/13738> [↑](#endnote-ref-29)
30. Mike Cohn Book, chapter 4, user stories applied [↑](#endnote-ref-30)
31. <https://books.google.co.uk/books?id=NUW_5rs6K-wC&pg=SA8-PA3&dq=space-time+tradeoff&hl=en&sa=X&ved=0ahUKEwiN7N7miNnjAhX9QEEAHWdIAQIQ6AEIKjAA#v=onepage&q=space-time%20tradeoff&f=false> [↑](#endnote-ref-31)