CE807 — Text Analytics Assignments 1 and 2

Instructor: Shoaib Jameel **Due Dates: Interim Report (Assignment 01)**: 28 February 2022 at 11:59:59. **Assignment 02**: 26 April 2022 at 11:59:59
School: School of CS and EE, University of Essex **Combined Assignments:** 1 & 2
Electronic Submission: https://www1.essex.ac.uk/e-learning/tools/faser2/
Queries? Please use the discussion forum for general questions. If you have something very specific, e.g., an extenuating circumstance, then please write to me at shoaib.jameel@essex.ac.uk

Your "Individual" Experiments with Probabilistic Topic Models

As you have learnt during your lectures and labs that probabilistic topic models [1] are a class of unsupervised machine learning models that help find the thematic patterns present in data. They have been widely applied to a range of domains from text to images. The basic topic model is called Latent Dirichlet Allocation (LDA) is being depicted by a graphical model in plate notation as shown in Figure 1[1].

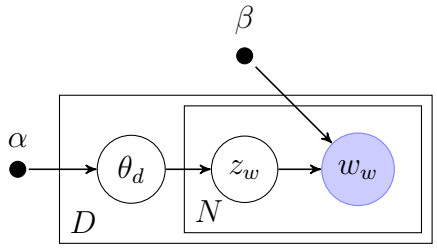

Figure 1: Plate Diagram of LDA [2].

1 Equations

See [2].

1.1 Generative Process

1: **for** document d_d in corpus D **do**
2: Choose $\theta_d \sim$ Dirichlet(α)
3: **for** position w in d_d **do**
4: Choose a topic $z_w \sim$ Multinomial(θ_d)
5: Choose a word w_w from $p(w_w|z_w, \beta)$, a multinomial distribution over words conditioned on the topic and the prior β.

[1] LaTeXsource credit: https://github.com/vvcephei/LaTeX-Topic-Model-Menagerie/blob/master/lda.tex

6: **end for**
7: **end for**

These models are primarily generative models that describe how data, e.g., words in a document are generated. We have also studied that a common "research" shortcoming in probabilistic topic models is finding an appropriate number of topics. Given a dataset, it is always difficult to find how many topics are ideal to describe that dataset. This assignment is all about this bit!

To determine the ideal number of latent topics, what has been usually done in the literature are the following *(crucial hints!)*

- Varying the number of topics arbitrarily [2] and shown with respect to some measure that one result or value obtained through that metric/measure is relatively *better* than other values;
- Using a tuning strategy [3] where the training data is split into training and tune data. The topic model is tuned on the training data and evaluated on the tuned data. The best evaluation result is then chosen as the ideal number of topics and then this is applied to the test data;
- Using a Bayesian non-parametric [4] approach to finding the number of topics automatically.

In this work, we will cover two goals:

1. Which papers or research works have studied the problem of determining the number of topics (Assignment 1)
2. How can you determine the number of topics that are ideal for the dataset (Assignment 2)

2 Assignment 1 (Weightage - 25%)

In the first assignment, we will spend some time finding related work which has studied the problem of determining the number of topics, e.g., these works could be using automated methods to find the number of topics or using data tuning techniques to determine the number of topics using some dataset.

2.1 What do you have to do?

In **one page** document only, please summarise the existing literature which has proposed ways to automatically determine the number of topics in a topic model. Please cite at least six works, and more are surely welcome, but you must describe how they have found the number of topics. Please do not just summarise those works or what they do in general, you must mention clearly how they have determined the number of topics in their work; this is all that matters in this work. You do not have to restrict yourselves to just the Latent Dirichlet Allocation model, you can cite any work as long as it is about

topic models. Your document should be Arial font 12 and must be only one page with references and figures, tables, etc. if you wish to include them (references certainly will be there). I will ignore anything beyond one page and will only mark what is within a page document. You may use Appendix, but that still has to be within a page. You will also convert your document to PDF and then submit it. *I will be very surprised if two assignments have the same set of citations! This can automatically be detected by our plagiarism checking model.*

2.2 Sample Assignment

A sample assignment is shared with you here. This is only for reference and you will **not cite** these papers in your report. You will also not copy its text in your report. Again, this copy is only for your reference and not for copying. I will regard it as plagiarism if I find the same text or cited papers in your report. Note that this sample document is in MS Word, but you will upload a PDF copy.

3 Assignment 2 (Weightage - 75%)

In this assignment, you will continue with your document used in Assignment 1. It means that the first page is your assignment 1, the second page will be your assignment 2. Your goal is to conduct experiments on a computer and apply data analysis techniques to find the number of topics in a text collection.

3.1 What do you have to do?

You will:

1. Select any available text dataset which you think is sufficient to conduct topic modelling, e.g., you could use the 20Newsgroups dataset that is suitable for document classification, or use OHSUMED dataset that is suitable for clustering (there is OHSUMED classification dataset too), or you may use tweet dataset or Amazon review dataset. You have full freedom to try out things that work and which do not work! Besides, you must be careful that your chosen dataset is not too huge such as the Wikipedia dataset comprising of millions of documents which will make your life much tougher. Your assignment 1 will help you to automatically make this decision for you as by now you know the related literature and the datasets which they have used.

2. You will do all the necessary pre-processing. Again, as I have repeatedly mentioned, there is no set of rules what is an ideal set of pre-processing techniques for a dataset.

3. You will then conduct topic modelling and do an analysis which could clearly show that n topics are sufficient for that dataset. You can choose either a classification or clustering setting. You will justify how you arrived at this value n. This is a core part of the work where,

- You will describe which topic modelling code you used. Note that you only have to use the basic Latent Dirichlet Allocation topic model here, which is a unigram topic model. We have already gone through several existing codes which implement this model during our lectures and labs.
- You will describe how different parameters were chosen by you in the topic model, the number of iterations, values of α, β, and others.
- You will describe how you found the number of topics in a data-driven way.

4. You will write your report on a page, yes again! You are free to do anything on this one page, e.g., tables, figures, references in Arial font 12. You will describe how you conducted the entire set of experiments without leaving out any details.

5. You will submit your document as a PDF.

Note that you do not have to be a computer science major to do assignment 2. All you need to do is to run a few commands and follow your labs carefully. What is challenging in assignment 2 is working towards a solution to find a reasonable number of topics using a data-driven way. Besides, the assignment requires problem-solving skills than in-depth knowledge of computer science. Assignment 2 revolves around "text analytics"!

3.2 Assignment 2 Help

I am not sharing any sample assignments at this time. However, there are already plenty of help materials available to you. Here is one post that will be of huge help.

4 What I am looking for in your assignments?

1. Using full one page. Following the font size mentioned above.

2. Citing relevant references. You may follow the referencing style which I have used in this document.

3. At least two references in each of the three categories mentioned above on determining latent topics, so six works or more in total (hint again!).

4. In assignment 1, mentioning how the research papers have determined the number of topics, not describing what those works are. I do not care about what they propose, I care about how they determined the number of topics.

5. In assignment 2, you mention all your experimental settings followed by your results so that anyone can easily "replicate" your work. Missing information will result in a loss of mark. So, you simply have to summarise what you do to reach the ideal number of topics. Note that changing α and β will have an impact on the number of topics. Copying and pasting commands in your report will not help, please refrain from writing commands, you must explain those commands if you want. Graphs, plots, tables would be an excellent addition to the report.

6. You have only used the Latent Dirichlet Allocation model code in assignment 2.

5 Note on the marking criteria

- If you cite six different papers and clearly explain how the authors have determined the number of topics including other details such as the dataset, there is no way marks can be deducted. There are videos on Moodle that must be watched before attempting assignment 1 to understand the requirements.
- Assignment 1 is worth 25% and you have to cite a minimum of six papers; therefore, each cited paper is worth 4.1% of marks out of 25%.
- Assignment 2 will be marked in the following way:
  - The result tables will be worth 50% of the 75% because they contain the core of your work.
  - The rest of the description takes another 50%.

6 Need Help?

Please have a look at the following resources before attempting your assignments:

- Link 1: https://moodle.essex.ac.uk/course/view.php?id=3700&section=9
- Link 2: https://moodle.essex.ac.uk/course/view.php?id=3700&section=6
- Link 3: https://moodle.essex.ac.uk/course/view.php?id=3700&section=19

References

[1] D. M. Blei and J. D. Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10(71):34, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] S. Jameel, W. Lam, and L. Bing. Supervised topic models with word order structure for document classification and retrieval learning. *Information Retrieval Journal*, 18(4):283–330, 2015.

[4] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

7 Plagiarism

The work you submit must be your own. Any material you use, whether it is from textbooks, classmates, the web or any other source must be acknowledged in your work.

All submissions are fairly and transparently checked for plagiarism. Please make sure that you provide frequent citations. But also make sure that each sentence is written is originally yours, i.e. the material is read, understood and the report is written using your own words and own language only. Do not copy and paste and rephrase the copied text.

There are many different forms of what is considered plagiarism. For example, based on the feedback from the SAO officer, many students were not aware that, e.g. copying entire paragraphs without clearly identifying them as quotes etc. is a form of plagiarism etc. Thus, please check back with your scientific writing module, before you submit it!

Further note that also plainly reusing software code or merely slightly adapting existing software code and submitting as one's own fulfils the matter of plagiarism. Cite any code that you reuse, too.

In 2019, 20% of the submitted reports were plagiarised. There were also multiple cases of software code plagiarism. This number is too high and shall be 0% in 2021!