It's Just A Report
With a subtitle, I guess

BY

Adam Shen

A REPORT SUBMITTED TO

The Department of Data Science, Analytics, and Artificial Intelligence

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

IN

Data Science, Analytics, and Artificial Intelligence

Carleton University
Ottawa, Ontario, Canada

SUPERVISOR: Dr. Joe Mama

NUMBER OF PAGES: vi, 10

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 First section

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor.

Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

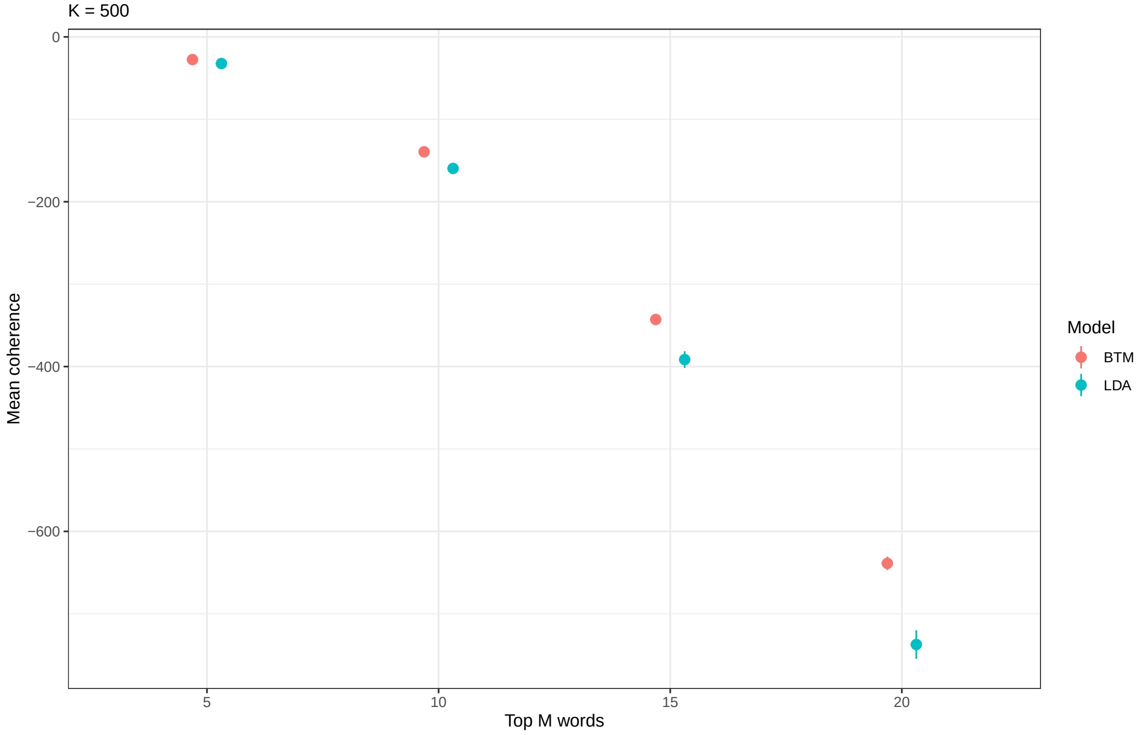

Figure 1.1: This caption explains the above plot.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer. Looking at the figure above (Figure 1.1), I don't remember what I was going to say.

1.2 Second section

1.2.1 First subsection

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

1.2.2 Second subsection

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

1.3 Third section

Everything I do is in R 4.0[1]. `purrr`[2] is a great package. I also used a bunch of other packages[3–5].

From reading list

[1] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: https://www.R-project.org/.

Supplementary readings

[2] Lionel Henry and Hadley Wickham. *purrr: Functional Programming Tools*. R package version 0.3.4. 2020. URL: https://CRAN.R-project.org/package=purrr.

Justification for inclusion: Required for x reason.

[3] Travers Ching. *qs: Quick Serialization of R Objects*. R package version 0.23.3. 2020. URL: https://CRAN.R-project.org/package=qs.

Justification for inclusion: Yass kween WERK!

[4] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.2. 2020. URL: https://CRAN.R-project.org/package=dplyr.

Justification for inclusion: Required for x reason.

[5] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Justification for inclusion: Required for x reason.

Chapter 2

Data preparation

2.1 First section

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

2.2 Second section

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

user_id	status_id	created_at	text	is_retweet	hashtags	lang
833499844517961728	1255290686250811392	2020-04-29 00:19:56	The NY Dem Presidential Primary on 6/23 is canceled due to COVID-19. Other elections on 6/23 will still take place: https://t.co/0P6PMPq20h - Rock the Vote	FALSE	NA	en
581180392	1254511313331597314	2020-04-26 20:42:59	In the midst of the current #pandemic, it is more important than ever to get creative with your #homeworkout. One great #trainfromhome option includes making a sandbag from a laundry bag. Here's how to implement sandbag circuits: https://t.co/Ottxy2YZAq	FALSE	"pandemic", "homeworkout", "trainfromhome"	en
819296296062201857	1304087649011953665	2020-09-10 16:01:38	#Masks serve 2 purposes: 1) To remind everyone that there is supposed to be a deadly pandemic, despite nobody knowing anybody who is sick with covid, and 2) To increase incidence of pulmonary illnesses, which will later be blamed on covid. Proof: See my pinned tweet.	FALSE	"Masks"	en
2350488460	1322387795718119425	2020-10-31 03:59:53	#GOPBetrayedAmerica decisions are being made to take your voice/vote away. The same #GOPCorruptionOverCountry that won't respond with help for every American during this once a century pandemic https://t.co/Gg5gecx7tb	FALSE	"GOPBetrayedAmerica", "GOPCorruptionOverCountry"	en

Table 2.1: This caption describes the above table.

2.3 Third section

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Did you see Table 2.1?

Algorithm 2.1 Collapsed Gibbs sampling algorithm for BTM.

```
Input: Number of topics $K$, Dirichlet hyperparameters $\alpha$ and $\beta$, collection of biterms $\mathbf{B}$
Output: $\widehat{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\Phi}}$
  Randomly assign a topic to all biterms, $b_i = (w_{i,1}, w_{i,2}) \in \mathbf{B}$
  for $iter = 1$ to $N$ do
    for each biterm, $b_i$ do
      Update topic assignment of $b_i$ according to Equation 2.1
      Update $n_k$, $n_{w_{i,1}|k}$, $n_{w_{i,2}|k}$
    end for
  end for
  Compute estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ using Equation 2.2 and Equation 2.3, respectively
```

$$P(z_i = k \,|\, \mathbf{z}_{(i)}, \mathbf{B}) \;\propto\; (n_{(i),\,k} + \alpha)\frac{(n_{(i),\,w_{i,1}\,|\,k} + \beta)(n_{(i),\,w_{i,2}\,|\,k} + \beta)}{(\sum_{w=1}^{W} n_{(i),\,w\,|\,k} + W\beta + 1)(\sum_{w=1}^{W} n_{(i),\,w\,|\,k} + W\beta)}, \tag{2.1}$$

$$\widehat{\theta}_k = \frac{n_k + \alpha}{N_B + K\alpha} \tag{2.2}$$

$$\widehat{\phi}_{k,w} = \frac{n_{w\,|\,k} + \beta}{\sum_{w=1}^{W} n_{w\,|\,k} + W\beta} \tag{2.3}$$

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor. Go and re-read Algorithm 2.1! I just realized I added LDA to the bib but didn't cite it[1].

Cited

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3.4 (2003), pp. 993–1022.

Not cited

[7] Julia Silge and David Robinson. "tidytext: Text Mining and Analysis Using Tidy Data Principles in R". In: *JOSS* 1.3 (2016). DOI: 10.21105/joss.00037. URL: http://dx.doi.org/10.21105/joss.00037.