United Nations General Debates: Uncovering International Political Topics through Machine Learning

P. Prado

07/01/2020

Abstract

This paper analyses the United Nations (UN) General Debates dataset provided by Harvard's Dataverse with the objective to uncover the main topics discussed over the years from 1970 to 2018. The UN General Debates dataset includes documents with the yearly speeches delivered by world leaders, from which the main topics in those documents can be revealed by (i) data preprocessing and cleaning, (ii) application of Machine Learning, more specifically and mainly the LDA algorithm, and (iii) data analysis. The algorithm was set to identify 20 topics of which 12 were chosen to further this study; those were (i) Peace in Africa, (ii) War & Terrorism, (iii) Korea, (iv) Israel & Palestine, (v) Peace in Iraq, (vi) Security Council, (vii) Human Rights, (viii) European Conflicts, (ix) Nuclear Weapons, (x) South Africa & Namibia, (xi) Climate Change, and (xii) Economic Development. The time series analysis on those topics revealed trends aligned with known historical events such as the African conflicts (Independence of Namibia in 1991), Yugoslav wars, the Global Financial Crisis and Climate Change, with the latter being the most prevailing topic in the last decade. Although satisfactory results are achieved, not only in determining the topics but also in revealing the trend overtime as well as relationships between topics and continents, further improvements are still possible. This could include tuning the algorithm to find the best number of topics to the LDA algorithm input, the employment of alternative unsupervised Machine Learning algorithms such as PCA and K-means, and even a combination with supervised learning techniques such as Random Forests. The current analysis can also be expanded to combine sentiment analysis to understand the regions' - or countries' - views on the given topics.

1 Introduction

The United Nations (UN) General Debates are held every year as part of the yearly General Assembly meeting. On this occasion, world leaders gather together to discuss and share their views on topics that affect the world, coutries and entities they represent. The opening statement from each leader is made available in the United Nations General Debates dataset.[1] The dataset contains the documented speeches for the period from 1970 to 2018 which is valuable in understanding how the countries concerns - or international political agenda - varied over time.

The goal of this study is to apply Machine Learning techniques to uncover the topics discussed in the UN General Debates documents, a task that otherwise would require extensive human resources to read and categorise them. The process of Unsupervised Machine Learning, also referred to as Topic Modelling, together with robust data analysis allows to answer questions such as: (i) which topics dominated the debates over the 49-year period and (ii) the relationship between topics and continents.

2 Method and Analysis

Natural Language Processing (NLP) involves the challenge of analysing unstratuctured data. As such, the key steps in this study are the data preprocessing and cleaning, the transformation of the texts into word tokens, the application of Machine Learning algorithms and the visualisation of the correlated data. While seemingly simple, there is extensive work required particularly in data cleaning and preprocessing as well as the implementation of the Latent Dirichlet Allocation (LDA) algorithm to identify topics, tasks that may result in a few hours of computation processing time.

[1] Jankin Mikhaylov, Slava; Baturo, Alexander; Dasandi, Niheer, 2017, "United Nations General Debate Corpus", https://doi.org/10.7910/DVN/0TJX8Y, Harvard Dataverse, V5

2.1 *Exploration*

The UN General Debates dataset contains 8,093 opening statements from the world leaders that attend the annual meeting. A look at the dataset shows the following information:

```
## [1] "data.frame"
```

```
## Observations: 8,093
## Variables: 7
## $ doc_id       <chr> "ALB_25_1970", "ARG_25_1970", "AUS_25_1970", "AUT_25_197…
## $ text         <chr> "33: May I first convey to our President the congratulat…
## $ country      <chr> "ALB", "ARG", "AUS", "AUT", "BEL", "BLR", "BOL", "BRA", …
## $ country_name <chr> "Albania", "Argentina", "Australia", "Austria", "Belgium…
## $ continent    <chr> "Europe", "Americas", "Oceania", "Europe", "Europe", "Eu…
## $ session      <int> 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, …
## $ year         <int> 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 19…
```

A full summary is provided with the code below, demonstrating that the dataset comprises statements from 1970 to 2018.

```
un_data.stats <- summary(un_data)
un_data.stats
```

```
##     doc_id              text              country          country_name
##  Length:8093        Length:8093        Length:8093        Length:8093
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   continent            session           year
##  Length:8093        Min.   :25.00   Min.   :1970
##  Class :character   1st Qu.:40.00   1st Qu.:1985
##  Mode  :character   Median :52.00   Median :1997
##                     Mean   :51.23   Mean   :1996
##                     3rd Qu.:63.00   3rd Qu.:2008
##                     Max.   :73.00   Max.   :2018
```

Another summary can be made focusing on the number of unique documents, countries and continents. The 8,093 opening statements were delivered by 200 world leaders in 5 continents over the 49-year period of the dataset.

```
##   documents years countries continents
## 1      8093    49       200          5
```

It is worth noting that the General Debates in 2018 included 196 countries/documents (193 UN country members, the European Union and the observer states of the Holy See and the State of Palestine) as seen in the extract below. This number is lower than 200 countries in the data set, which derives from the political changes that resulted in the merger or separation of countries (e.g. Yugoslavia) over time.

```
## # A tibble: 1 x 4
##    year documents countries continents
##   <int>     <int>     <int>      <int>
## 1  2018       196       196          5
```

2.2 *Data preprocessing and analysis*

The `un_data` dataset is a data frame object which is not the best to analyse text data. The corpus data class is widely utilised in Natural Language Processing (NLP) therefore the dataset conversion is the first step. The preprocessing tasks in this study are:

1. Corpus conversion;
2. Lemmatisation;
3. Tokenization; and
4. Cleaning.

The steps are briefly described below:

2.2.1 Step 1: Corpus conversion

The conversion of data frame into corpus is done with the package `quanteda` with a simple line of code as demonstrated below.

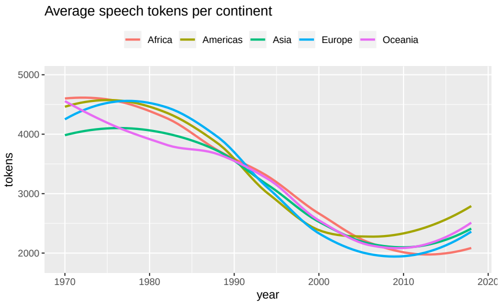
```
un_corpus <- corpus(un_data, text_field = "text")
class(un_corpus)
```

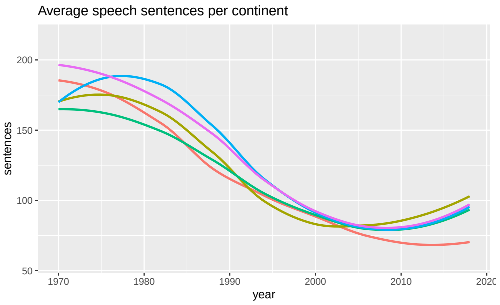
```
## [1] "corpus" "list"
```

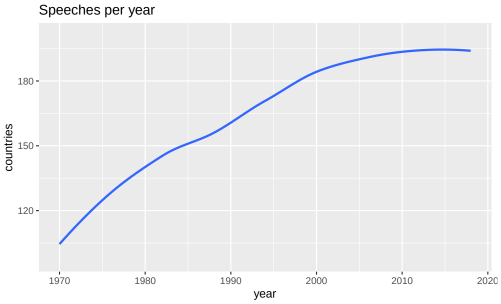
Now with a corpus object, the summary provides a lot more information, already containing data such as the number of tokens (words) and sentences per document.

```
## Corpus consisting of 8093 documents, showing 5 documents:
##
##         Text Types Tokens Sentences country country_name continent session year
##  ALB_25_1970  1728   9078       256     ALB      Albania    Europe      25 1970
##  ARG_25_1970  1425   5192       218     ARG    Argentina  Americas      25 1970
##  AUS_25_1970  1612   5690       270     AUS    Australia   Oceania      25 1970
##  AUT_25_1970  1340   4717       164     AUT      Austria    Europe      25 1970
##  BEL_25_1970  1288   4786       207     BEL      Belgium    Europe      25 1970
##
## Source: /Users/peterprado/data_projects/CYOP/* on x86_64 by peterprado
## Created: Wed Jan  8 20:08:16 2020
## Notes:
```

From the number of tokens and sentences it is possible to see how the length of the speeches changed over time. The plots below, a similar trend can be observed between the length and number of speeches delivered by world leaders in each year. With the increase of country members, the UN has likely implemented measures to limit the time that each country had to deliver their speech.


Average speech tokens per continent


Average speech sentences per continent



2.2.2 Step 2: Lemmatisation

The analysis of text data requires grouping words found in texts. However, grouping exact matches within a text would not yield good results in topic modelling due to the inflection of words (e.g. consulting, consultant, consultation). There are many approaches to solving this problem but the most popular ones are Stemming and Lemmatisation. In short, Stemming stands for "cutting" part of the word to reach its root. In this case, "consulting" and "consultant" would be reduced to "consult". On the other hand, Lemmatisation looks at the morphological meaning of the word, as defined in the Cambridge Dictionary:

> *the process of reducing the different forms of a word to one single form, for example, reducing "builds", "building", or "built" to the lemma "build":*
>
> - *Lemmatization is the process of grouping inflected forms together as a single base form.*
> - *In dictionaries, there are fixed lemmatization strategies.*

Each approach has advantages and disadvantages. Stemming is a faster process but the results may not be as realiable, for instance, "popular" and "population" would become "popula". Lemmatisation, on the other hand, preserves better meaning of the words albeit the processing time being extremely long. In this study, the latter has been used.

```
un_corpus_lemma <- # create a new corpus to preserve the original corpus
  corpus(un_data, text_field = "text")

start <- Sys.time()

un_corpus_lemma$documents$texts <-
  lemmatize_strings(un_corpus_lemma$documents$texts, dictionary = lexicon::hash_lemmas)

Sys.time()-start
```

```
## Time difference of 1.000218 hours
```

The processing time is indicated above as the "Time difference" between start and end of the Lemmatisation.

2.2.3 Step 3: Tokenization

Up to now, the text data is stored in documents as strings, that is, the text for each document is a single string. Tokenization - in NLP - is the process of splitting the strings into separate words (or tokens). This process will still keep the tokens allocated to each document in the corpus.

```
un_tokens <-
  quanteda::tokens(
    un_corpus_lemma
  )
```

A high level look into the results below already indicates why cleaning is important in this study. The tables below show the top 10 and bottom 10 terms.

```
## # A tibble: 10 x 2
##    term  count
##    <chr> <int>
##  1 ,      8093
##  2 .      8093
##  3 a      8093
##  4 and    8093
##  5 be     8093
##  6 for    8093
##  7 in     8093
##  8 of     8093
##  9 on     8093
## 10 that   8093
```

```
## # A tibble: 10 x 2
##    term  count
##    <chr> <int>
##  1 ¶         1
##  2 ˙         1
##  3 ˛         1
##  4 ∞         1
##  5          1
##  6 ¢         1
##  7 0.001     1
##  8 0.003     1
##  9 0.005     1
## 10 0.006     1
```

2.2.4 Step 4: Cleaning

In this step, the objective is to transform all tokens into lowercase and eliminate tokens that are symbols, URLs, punctuation, stopwords, numbers and hyphens. The `quanteda` package is used for lowercase transformation and removal of stopwords, however it must be done by transforming the object into a Data Feature Matrix. The result shows an improvement as noted below.

```
## # A tibble: 10 x 2
##    term     count
##    <chr>    <int>
##  1 much      8044
##  2 state     7998
##  3 also      7988
##  4 good      7972
##  5 general   7947
##  6 peace     7893
##  7 make      7886
##  8 people    7885
##  9 assembly  7871
## 10 can       7853
```

```
## # A tibble: 10 x 2
##    term      count
##    <chr>     <int>
##  1 00tmunity     1
##  2 00ü           1
##  3 04sj          1
##  4 06itte        1
##  5 06wie         1
##  6 0a0           1
##  7 0ctober7      1
##  8 0n            1
##  9 0n1ta         1
## 10 0nxc8f        1
```

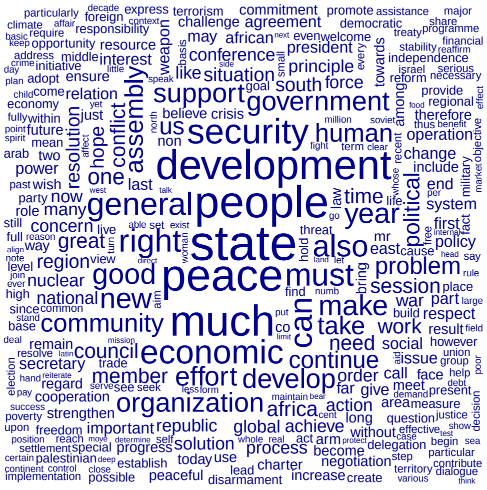
Cleaning can still be improved by (i) trimming the words rarely occurred and (ii) removing words that do not add value to topic modelling. In the first case (i), as seen below, a few words have numbers instead of letters such as “0ctober” and “0n” due to the processing of the texts stored as image files prior to 1992 (Baturo et al. 2017). In the second case (ii), the words “nation”, “unite”, “international”, “country” and “world” are amongst the most frequent ones as they directly relate to the United Nations, therefore also not adding value to topic modelling.

The trimming can be done to remove very low occurrences. Removing the bottom 0.002% yields the following result:

```
## # A tibble: 10 x 2
##    term     count
##    <chr>    <int>
##  1 much      8044
##  2 state     7998
##  3 also      7988
##  4 good      7972
##  5 general   7947
##  6 peace     7893
##  7 make      7886
##  8 people    7885
##  9 assembly  7871
## 10 can       7853
```

```
## # A tibble: 10 x 2
##    term         count
##    <chr>        <int>
##  1 1920s           17
##  2 abdallah        17
##  3 abjure          17
##  4 accentuation    17
##  5 acquis          17
##  6 agitate         17
##  7 alan            17
##  8 alba            17
##  9 annapolis       17
## 10 ant             17
```

A word cloud plot helps visualise the frequency of the top words comparatively. The size of the words represent their frequency.



2.3 *LDA*

There are various Machine Learning algorithms that can be applied in NLP. This study utilises the Latent Dirichlet Allocation (LDA) algorithm for topic modelling introduced by Blei et al. (2003). It was chosen due to its frequent use in unsupervised learning within NLP. There is extensive technical explanation about LDA in the referenced publication, therefore this study will briefly explain how fitting the model works.

A known limitation of the LDA algorithm is that the number of topics (or clusters) need to be specified upfront. In the case of the UN General Debates this can mean losing relevant insights. For now, the target will be to identify 20 topics.[2]

To fit the model, the parameters below are specified into the code that follows:

- *k* for the number of topics
- *seed* for the replication of the results
- *method* for the sampling method[3]

```
un_dtm <- convert(un_dfm, to = "topicmodels")
k <- 20 #number of topics
seed = 123 #necessary for reproducibility

start <- Sys.time() #calculating the total runtime
lda <- LDA(un_dtm, k = k, method = "GIBBS", control = list(seed = seed))
Sys.time()-start # total runtime between end and start of LDA processing
```

```
## Time difference of 33.37474 mins
```

The “time difference” shown above indicates, again, how long it took for the LDA function to be processed.

There are many other parameters that can be adjusted within the LDA function of the code, those include the samples to discard, iterations, etc. Those were not adjusted in this study in order to assess the performance of a standard application of the algorithm to the UN General Debates dataset.

[2] 20 was arbitrarily determined.

[3] This LDA application utilised the method Gibbs for sampling (Resnik and Hardisty 2010).

3 Results

3.1 Topics

The following is a result of the 20 topics found within the 8,093 documents withe the 15 most relevant words that form the topic.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
african	people	people	israel	state	security	peace	right	government	right
peace	war	republic	peace	island	state	continue	human	state	human
africa	life	peace	arab	small	peace	issue	council	america	peace
government	human	state	palestinian	government	region	much	security	american	development
community	terrorism	people's	state	pacific	effort	develop	must	latin	new
republic	right	struggle	resolution	develop	council	south	member	people	must
conflict	child	democratic	people	new	people	year	general	political	people
general	state	korea	right	development	iraq	global	much	peace	economic
state	terrorist	national	security	people	stability	economic	peace	central	social
take	can	war	territory	caribbean	also	can	state	president	respect
development	million	independence	israeli	change	call	assembly	organization	make	political
organization	one	government	lebanon	continue	support	remain	conflict	support	order
support	freedom	force	east	support	achieve	power	secretary	respect	community
president	year	viet	middle	member	resolution	far	need	social	freedom
people	fight	power	palestine	community	arab	indeed	work	must	principle

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
development	european	nuclear	people	much	development	develop	problem	problem	must
reform	security	state	africa	can	global	economic	general	south	us
security	state	weapon	south	one	sustainable	development	session	effort	much
cooperation	cooperation	disarmament	african	power	challenge	per	assembly	development	can
effort	region	arm	independence	state	climate	much	co	economic	year
terrorism	conflict	security	economic	organization	goal	economy	operation	co	new
organization	europe	treaty	state	may	change	trade	government	operation	one
general	republic	peace	namibia	time	security	cent	conference	solution	make
challenge	process	soviet	regime	us	agendum	resource	development	general	time
council	community	non	struggle	even	support	increase	hope	hope	now
global	support	use	session	fact	also	need	effort	community	need
millennium	regional	force	organization	interest	commitment	growth	develop	concern	work
summit	union	military	right	problem	address	financial	organization	peace	good
also	new	conference	support	without	achieve	year	concern	continue	many
support	member	relation	apartheid	become	effort	market	great	situation	today

As seen above, a few topics do not provide much meaning by looking at the first 15 words. This is because, certainly, a lot of the documents make reference to the General Assembly of the UN and some general purposes of the UN as an organisation itself. Therefore, a few topics are separated below to proceed further in the study, those are:

- Topic 1: Peace in Africa
- Topic 2: War & Terrorism
- Topic 3: Korea
- Topic 4: Israel & Palestine
- Topic 6: Peace in Iraq
- Topic 8: Security Council
- Topic 10: Human Rights
- Topic 12: European Conflicts
- Topic 13: Nuclear Weapons
- Topic 14: South Africa & Namibia
- Topic 16: Climate Change
- Topic 17: Economic Development

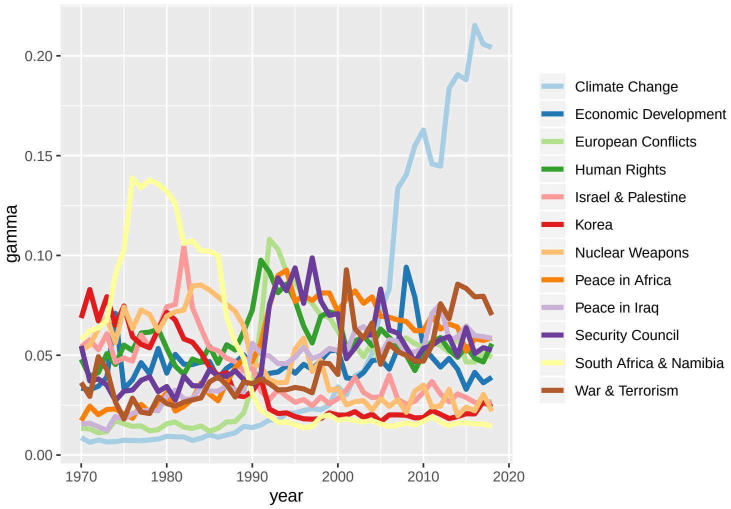
The reduction from 20 to 12 topics shall facilitate the visualisation of topic trends and topic relationships.

3.2 *Topic trends*

With the topics already generated, it is possible to start analysing relationships between topics, year, continents and even countries.[4]

This is possible because LDA not only generated the topics found in the documents (based on the combination of words) but also created the probabilities of each topic within each document.

This probability is referred to as gamma γ.



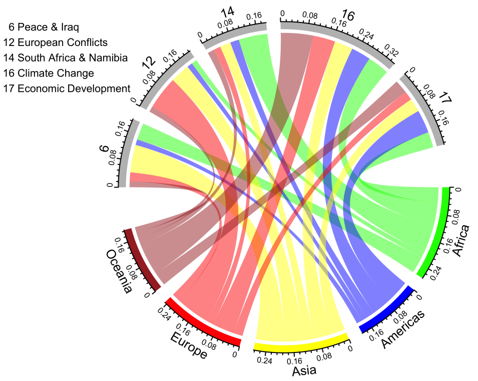
From the plot above, the probability of the topics being debated by the UN can be directly attributed to historical facts, such as:

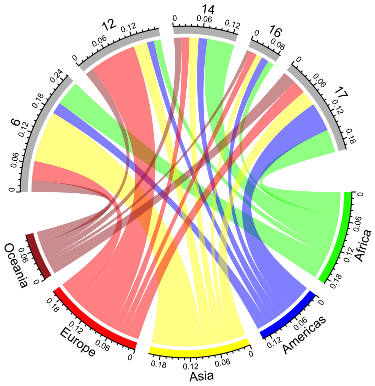
- Topic 17: Economic Development peaked between 2005-2010, correlating to the Global Financial Crisis of 2007-2008
- Topic 12: European Conflicts peaked between 1990-1995, correlating to the Yugoslav wars between 1991-2001
- Topic 14: South Africa & Namibia peaked between 1975-1980 remaining high during 1980s, correlating to the South African Border War between 1966-1989 and Namibia's Independence in 1990
- Topic 6: Peace in Iraq peaked between 2010-2015, correlating to the Iraq War between 2003-2011

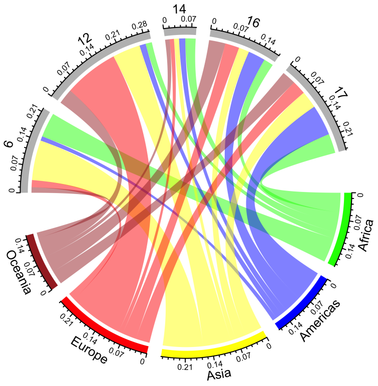
Those are just a few. As observed in the plots above, the Topic 16: Climate Change, has the most notable increase over recent years. Based on this data it is possible to conclude that Climate Change has been prevailing in the UN General Debates since 2010.

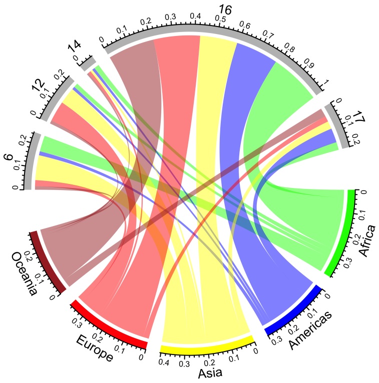
Representing 12 topics visually is still a lot, therefore, the continent-topic relationship is shown below only in relation to the 5 topics discussed above.

[4] Country analysis is excluded from this study.


All Time Topic and Continent Relationship


Topic and Continent Relationship in 1990


Topic and Continent Relationship in 2000


Topic and Continent Relationship in 2018

While the continents' relationship with the Climate Change topic is somewhat homogeneous (particularly in 2018), other topics have a clear stronger relationship to the continent where such topic has initially emerged. The relationships are:

- Economic Development and the Americas
- European Conflicts and Europe
- Peace in Iraq and Asia
- South Africa & Namibia and Africa.

4 Conclusion

The UN General Debates dataset is extremely valuable in providing insights about international politics. In this study, the application of Unsupervised Machine Learning via the LDA algorithm proved effective in uncovering the main topics and their trends. Some of the identified topics related to historic events of regional wars, economic development and other international issues. According to those results, the most prevalent topic of the of the past decade is Climate Change. It reached the highest probability - over 20% - recorded for the analysed period, followed by the topics related to South Africa & Namibia and European Conflicts.

In terms of expanding this study, possible improvements include the benchmarking of other topic modelling algorithms such as PCA and K-means. This work can be further developed via the application of supervised learning techniques and the addition of sentiment analysis which produce invaluable insights in understanding the countries perspectives on certain matters. Lastly, such understanding of countries perspectives would allow to conclude whether countries within the same continent discuss the same topics.

5 References

Alexander Baturo, Niheer Dasandi, and Slava Mikhaylov, “Understanding State Preferences With Text As Data: Introducing the UN General Debate Corpus” Research & Politics, 2017.

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research. 3 (4–5):993–1022.

Resnik P, Hardisty E. 2010. Gibbs sampling for the uninitiated. Technical Report UMIACS-TR-2010-04, University of Maryland. http://drum.lib.umd.edu//handle/1903/10058.