**IT 6903 Assignment 1  
Topic Modeling With Latent Dirichlet Allocation (LDA)**

Elaine Harris

Kennesaw State University

MSIT Program – Fall 2020

Contact: [eharr147@students.kennesaw.edu](mailto:eharr147@students.kennesaw.edu)

Table of Contents

[Abstract 3](#_Toc51327131)

[**IT 6903 Assignment 1 Topic Modeling With Latent Dirichlet Allocation (LDA)** 4](#_Toc51327132)

[Real Life Application 4](#_Toc51327133)

[Social Media Posts Format 5](#_Toc51327134)

[Proposed Solution 6](#_Toc51327135)

[Topic Modeling with Latent Dirichlet Allocation 7](#_Toc51327136)

[LDA Practical Example 8](#_Toc51327137)

[Planned Execution and Expected Results 9](#_Toc51327138)

[Additional Readings 10](#_Toc51327139)

[References 11](#_Toc51327140)

[Figure 1 LDA Model Code Snippet 9](#_Toc51327177)

[Figure 2 Coherence Model Code Snippet 9](#_Toc51327178)

Abstract

This report describes a proposed practical application of Natural Language Processing: Topic Modeling using LDA - Latent Dirichlet Allocation. LDA is one of the most popular algorithms for discovering topics in a large body of documents. In this assignment, a practical application of LDA for the analysis of Social Media activity during the COVID-19 Pandemic will be proposed.

**IT 6903 Assignment 1  
Topic Modeling With Latent Dirichlet Allocation (LDA)**

This paper describes a proposed use of Natural Language Processing to discover the topics addressed in the Singapore Health Ministry social media accounts during the COVID-19 Pandemic. The author is also taking the 7993 Capstone class with the MSIT program, and the project topic is *Analyzing Risk Communication and Behavioral Change During COVID-19 Pandemic*. A fundamental objective of the project is to learn what topics government officials choose to address in their social media accounts, and whether or not their choice of topics has an influence in the public’s behavior. The author proposes that an LDA algorithm can be used to scan social media posts and discover the topics contained in Twitter and Facebook posts.

# Real Life Application

The following is an excerpt from the Capstone Project Requirements by Dr Shirley Tian:

***Project Title: Analyzing Risk Communication and Behavioral Change During COVID-19 Pandemic***

*Introduction: We are particularly interested in how to track, model, understand and predict the spread of COVID-19 through data mining and visualization methods. We will address the following questions: What social behaviors should policymakers and government leadership take into consideration when engaging in public health messaging? How can messaging better promote facts and counter misinformation? We will compare both national and international public health messaging and their effects by analyzing and visualizing social media data.*

## Social Media Posts Format

This project only collects posts from the Singapore Ministry of Health, and the text is very repetitive. We need to analyze the text and attempt to identify topics related to social distancing, facial coverings and hygiene, Government-imposed restrictions, at a minimum, and any other topics that may surface after analysis.

Below are some examples of real posts collected from Twitter and Facebook.

|  |  |  |
| --- | --- | --- |
| **Username** | **Text** | **Date** |
| sporeMOH | There are currently 117 confirmed cases who are still in hospital. Of these, most are stable or improving, and none is in the intensive care unit. | 2020-08-01 |
| sporeMOH | As of 1 August 2020, 12pm, 249 more cases of COVID-19 infection have been discharged from hospitals or community isolation facilities. In all, 46,740 have fully recovered from the infection and have been discharged. | 2020-08-01 |
| sporeMOH | Of the new cases, 99% are linked to known clusters, while the rest are pending contact tracing. | 2020-08-01 |
| sporeMOH | As of 1 August 2020, 12pm, we have confirmed and verified an additional 307 cases of COVID-19 infection in Singapore. Breakdown: 5 imported, 1 case in the community & 301 cases residing in dorms. | 2020-08-01 |
| sporeMOH | As of 1 August 2020, 12pm, we have preliminarily confirmed an additional 307 cases of COVID-19 infection in Singapore. | 2020-08-01 |
| sporeMOH | There are currently 136 confirmed cases who are still in hospital. Of these, most are stable or improving, and none is in the intensive care unit. | 2020-07-31 |

Table Sample Twitter Posts

|  |  |
| --- | --- |
| Date | Post |
| 5/1/2020 | As of 1 May 2020, 12pm, we have confirmed and verified an additional 932 cases of COVID-19 infection in Singapore. The breakdown is as follows:  a) Imported cases: 0 b) Cases in the community: 11 (5 Singaporeans/Permanent Residents, 6 Work Passes) c) Work Permit holders (residing outside dormitories): 16 d) Work Permit holders (residing in dormitories): 905 Of the new cases, 70% are linked to known clusters, while the rest are pending contact tracing.  24 more cases of COVID-19 infection have been discharged from hospitals or community isolation facilities. In all, 1,268 have fully recovered from the infection and have been discharged from hospitals or community isolation facilities.  There are currently 1,764 confirmed cases who are still in hospital. Of these, most are stable or improving, and 23 are in critical condition in the intensive care unit. Read more in the press release: https://www.moh.gov.sg/news-highlights/details/24-more-cases-discharged-932-new-cases-of-covid-19-infection-confirmed |
| 5/1/2020 | As of 1 May 2020, 12pm, we have preliminarily confirmed an additional 932 cases of COVID-19 infection in Singapore, the vast majority of whom are Work Permit holders residing in foreign worker dormitories. Five cases are Singaporeans/ Permanent Residents. We are still working through the details of the cases, and further updates will be shared via the MOH press release that will be issued tonight. https://www.moh.gov.sg/news-highlights/details/932-new-cases-of-covid-19-infection |
| 5/2/2020 | <One More Week of Tightened Circuit Breaker Measures; Gradual Easing Thereafter> We tightened the circuit breaker measures two weeks ago and said we would review the situation by 4 May. To date, we have seen a reduction in our daily community infection numbers. But we are still not in single digits. So aside from some minor adjustments (for TCM practitioners and what one can do within the grounds of strata-titled residential developments), we will have to continue with the tight measures for at least another week. From 12 May, we can allow some gradual easing of these tightened measures. In particular, we will allow selected activities and services to resume operations, including home-based businesses, selected food manufacturing and retail outlets; laundry and barber services, and pet supplies stores.  We will also phase in the reopening of the economy, and allow more people to resume going to work. The key precondition for this to happen is stricter requirements in the workplace for safe distancing, with employers to take ownership and responsibility of these measures. The specific measures are being worked out, and government agencies will be engaging industry associations, business chambers and firms in the coming days.  One specific requirement is that all businesses must put in place the SafeEntry app to log the entry and exit of their staff and visitors. This information will enable us to speed up contact tracing.  Further down the road, from 19 May, we will start to bring back students in small groups for face-to-face lessons. We will focus on the graduating cohort who are taking national exams, and amongst them, priority will be given to those who require school facilities for coursework and practical sessions, and those who need additional support during the school vacation period. The Institutions of Higher Learning especially the ITEs will also bring back small groups of students on campus for critical consultations, projects or practicums We have provided a broad outline of how we intend to adjust the measures in the coming weeks so that everyone knows what to expect. But the situation is fluid, and the measures may have to be further adjusted along the way.  Most importantly, remember that the circuit breaker is still in place till 1 June. This is not the time to slacken and let our guard down. We must continue to stay disciplined and vigilant. Stay home as much as possible. Go out for essential activities only, and when you have to do so, try to go out one person at a time. We still have a long fight ahead of us, and the virus can flare up again anytime. Let’s stay focused, work together and win this fight against the virus. |

Table Sample Facebook Posts

# Proposed Solution

We will use **Topic Modeling** to detect topics in the social media posts. *Topic Modeling* is a “type of statistical modeling for discovering the abstract topics that occur in a collection of documents.” (Li, 2020)

Topic Modeling will be performed using an unsupervised Natural Language Processing algorithm to detect topics in the collected posts. *Natural language processing (NLP)* is a “branch of artificial intelligence that helps computers understand, interpret and manipulate human language.” ("What is Natural Language Processing?", 2020). *Unsupervised learning* is a “type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision.” ("Unsupervised learning", 2020)

A very popular solution for topic detection applications is the *Latent Dirichlet Allocation(LDA)* algorithm. Implementations of the algorithm are available in commercial Statistical Modeling Packages such as SAS, and as open source libraries in popular Data Science programming languages such as Python and R. For this project, we will use the **Gensim** library for Python.

## Topic Modeling with Latent Dirichlet Allocation

According to Wikipedia, “the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.” ("Latent Dirichlet allocation", 2020).

Topic Modeling with LDA “builds a topic per document model and words per topic model, modeled as Dirichlet distributions [[1]](#footnote-1)” (Li, 2020), and the algorithm works in two steps:

First, the algorithm detects a set of topics across all documents. Each topic consists of a set of words selected based on the calculated probability of those words appearing together. In a large corpus, the top *x* words are selected to represent a topic.

Second, the algorithm evaluates each *document* (in our case, each post is a document) against the list of topics. A probability distribution is calculated that results in a match score to each topic. For example, if we had 3 topics, document A might be scored as 60% topic 1, 35% topic 2 and 5% topic 3.

For our project, we are interested in detecting the scores for each topic, and identifying both the most and least dominant topic for each post.

## LDA Practical Example

This real-world example is taken from Edwin Chen’s blog *Introduction to Latent Dirichlet Allocation* (Chen, 2020)*.* Chen applied LDA to a set of Sarah Palin’s emails. Here are some of the topics that the algorithm learned:

**Trig/Family/Inspiration**: family, web, mail, god, son, from, congratulations, children, life, child, down, trig, baby, birth, love, you, syndrome, very, special, bless, old, husband, years, thank, best, …

**Wildlife/BP Corrosion**: game, fish, moose, wildlife, hunting, bears, polar, bear, subsistence, management, area, board, hunt, wolves, control, department, year, use, wolf, habitat, hunters, caribou, program, denby, fishing, …

**Energy/Fuel/Oil/Mining**: energy, fuel, costs, oil, alaskans, prices, cost, nome, now, high, being, home, public, power, mine, crisis, price, resource, need, community, fairbanks, rebate, use, mining, villages, …

**Gas**: gas, oil, pipeline, agia, project, natural, north, producers, companies, tax, company, energy, development, slope, production, resources, line, gasline, transcanada, said, billion, plan, administration, million, industry, …

**Education/Waste**: school, waste, education, students, schools, million, read, email, market, policy, student, year, high, news, states, program, first, report, business, management, bulletin, information, reports, 2008, quarter, …

**Presidential Campaign/Elections**: mail, web, from, thank, you, box, mccain, sarah, very, good, great, john, hope, president, sincerely, wasilla, work, keep, make, add, family, republican, support, doing, p.o,

## Planned Execution and Expected Results

Execution the Gensim implementation of LDA is as simple as cleaning up the text and instantiating the model in Python:

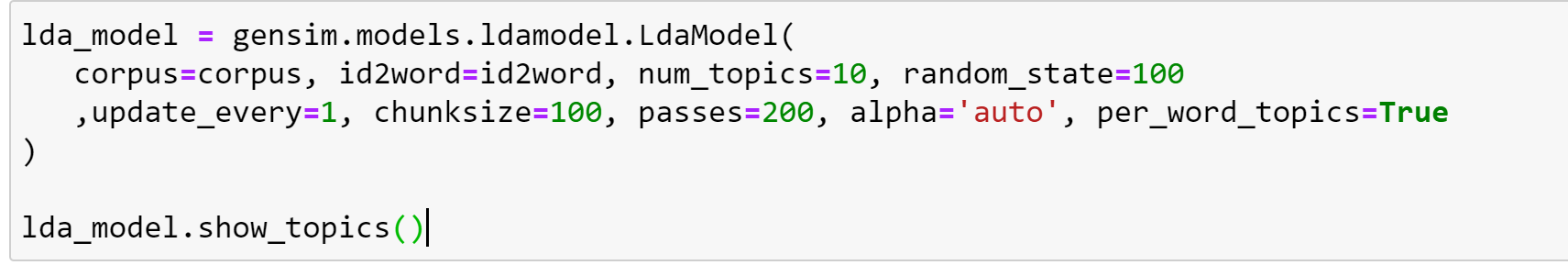


Figure LDA Model Code Snippet

The performance of the model is measured by its *Coherence Score*. Generally speaking, the higher the score, the better. Conveniently, Gensim also provides a Coherence Model to measure the performance of the LDA model:

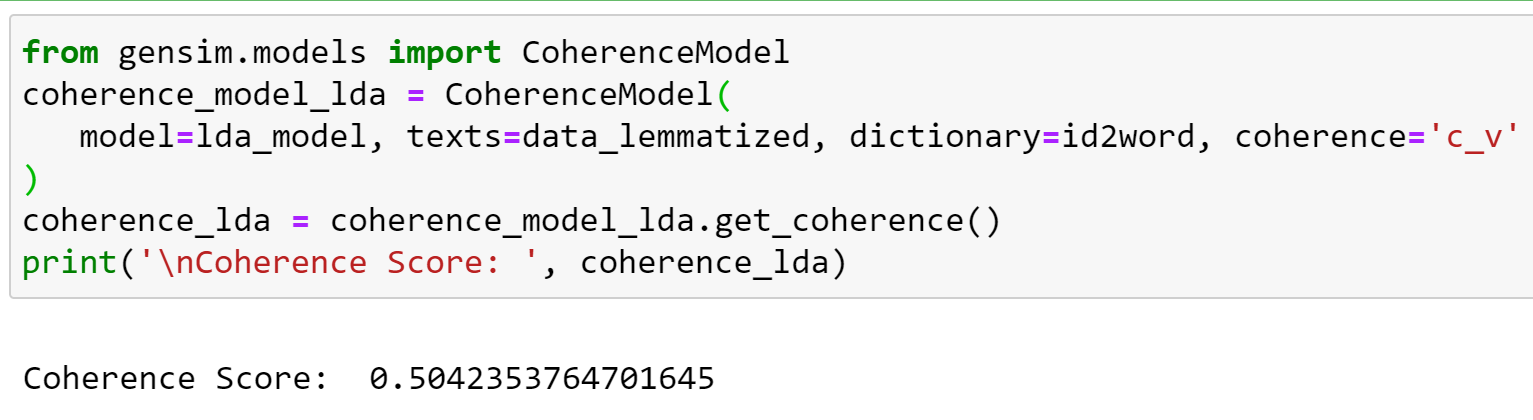


Figure Coherence Model Code Snippet

The expected results are a list of topics that will allow the identification of posts related to behavior education such as hand washing, face coverings, social distancing measures, gathering restrictions, etc. Further regression analysis will be performed on the identified topics, but such analysis is not the subject of this paper.

Additional Readings

**SAS® Visual Text Analytics 8.4: Programming Guide**

LDA Topic Modeling Action Set

<https://documentation.sas.com/?docsetId=casvtapg&docsetTarget=p0rjgv6q2ez91fn1a06crcjoq8lq.htm&docsetVersion=8.4&locale=en>

**Latent Dirichlet Allocation in Gensim**

This module allows both LDA model estimation from a training corpus and inference of topic distribution on new, unseen documents. The model can also be updated with new documents for online training.

<https://radimrehurek.com/gensim/models/ldamodel.html>

**Towards Data Science article on the gory details of Dirichlet distribution**

<https://towardsdatascience.com/dirichlet-distribution-a82ab942a879>

References

Chen, E. (2020). Introduction to Latent Dirichlet Allocation. Retrieved 18 September 2020, from <https://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/>

Latent Dirichlet allocation. (2020). Retrieved 18 September 2020, from <https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation>

Li, S. (2020). Topic Modeling and Latent Dirichlet Allocation (LDA) in Python. Retrieved 18 September 2020, from <https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24>

Liu, S. (2020). Dirichlet distribution. Retrieved 18 September 2020, from <https://towardsdatascience.com/dirichlet-distribution-a82ab942a879>

Unsupervised learning. (2020). Retrieved 18 September 2020, from <https://en.wikipedia.org/wiki/Unsupervised_learning>

What is Natural Language Processing?. (2020). Retrieved 18 September 2020, from <https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html>

**END OF DOCUMENT**

1. Textbook definition: The Dirichlet distribution Dir(α) is a family of continuous multivariate probability distributions parameterized by a vector α of positive reals. It is a multivariate generalisation of the Beta distribution. Dirichlet distributions are commonly used as prior distributions in Bayesian statistics. (Liu, 2020) [↑](#footnote-ref-1)