ASSIGNMENT 2

AML-3204 Social Media Analytics

Submitted By:

Ma. Angelica D. Serrano C0785370

Instructor:

Debashish Roy

Lambton College

Submission date:

29-July-2021

Template Prepared by William Pourmajidi

Contents

Introduction

In this project, Topic modelling with *Gensim* was applied to create an intuitive model using LDA (Latent Dirichlet Allocation) algorithm. Dataset used was an extracted tweets from CityNews Toronto – a popular news program in Canada.

Methods

I. Data Collection procedure

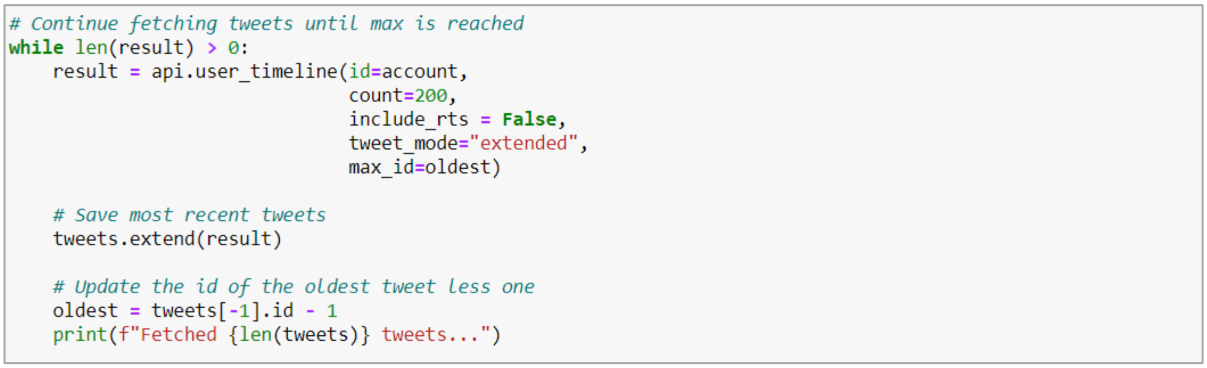
Tweepy was leveraged to extract data from Twitter. It is the most efficient way of accessing Twitter API using python. Developer credentials were defined primarily to connect and obtain data from Twitter which include the following:

- Access Token
- Access Token Secret
- Consumer API Key
- Consumer API Secret

Subsequently, Twitter authentication handler and API objects were created as well.

Since Twitter function allows only 200 maximum tweet extraction per timeline, a *while loop* was included on the code to expand the fetch results until the oldest tweet, excluding RTs. This process has accumulated more than 2K tweets.

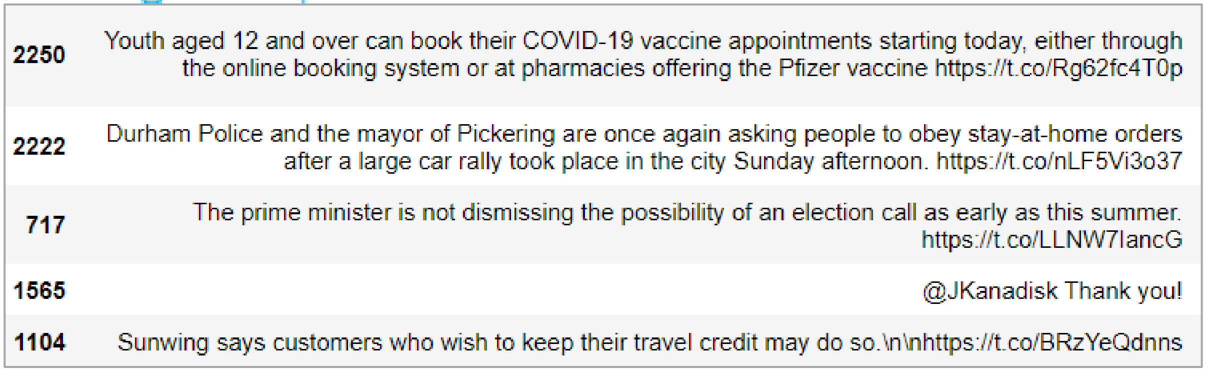
Code snippet:



```
# Continue fetching tweets until max is reached
while len(result) > 0:
    result = api.user_timeline(id=account,
                               count=200,
                               include_rts = False,
                               tweet_mode="extended",
                               max_id=oldest)

    # Save most recent tweets
    tweets.extend(result)

    # Update the id of the oldest tweet less one
    oldest = tweets[-1].id - 1
    print(f"Fetched {len(tweets)} tweets...")
```

II. Data Cleansing steps

Mainly, the extracted dataset has native twitter data format which consist of sentences, fragments, and twitter URLs. Thus, data cleaning is required before utilizing it on the later steps.



2250	Youth aged 12 and over can book their COVID-19 vaccine appointments starting today, either through the online booking system or at pharmacies offering the Pfizer vaccine https://t.co/Rg62fc4T0p
2222	Durham Police and the mayor of Pickering are once again asking people to obey stay-at-home orders after a large car rally took place in the city Sunday afternoon. https://t.co/nLF5Vi3o37
717	The prime minister is not dismissing the possibility of an election call as early as this summer. https://t.co/LLNW7lancG
1565	@JKanadisk Thank you!
1104	Sunwing says customers who wish to keep their travel credit may do so.\n\nhttps://t.co/BRzYeQdnns

Table 1 *- Sample extracted tweets*

A. **Removal of null and duplicate values –** Null, empty strings, blank values and duplicates may generate errors in creating any model. Hence, elimination of these entries from our samples is a vital step.

B. Removal of URLs and non-alphanumeric characters – Characters such as emoticons and strings like shortened URLs will have no relevance in topic modelling. So, to optimize the inputs these items were eliminated from the dataset.

C. **Removal of stopwords -** In natural language processing (NLP), stopwords are common English words that can be safely ignored or removed from a sentence without sacrificing the quality of data to be analyzed. Some examples of stopwords are "*has, have, I, is, are".*

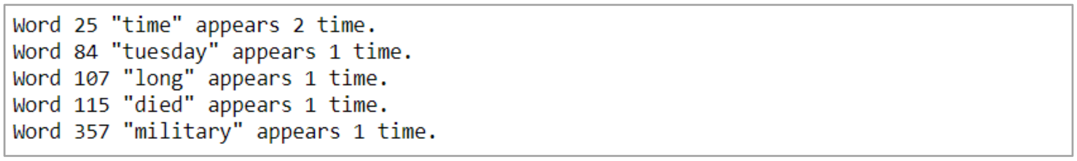
D. **Lemmatization** – is another NLP process that is used to normalize words in a document. Lemmatization produces the root forms of derived words. This is very similar to stemming process but generates more coherent output. For example: *vaccination* become *vaccine.*

III. Topic Modelling

Topic modelling is a process of determining the conceptual "Topic" of a document or sentence. LDA is one of the popular statistical models that are used for this process. In this section, the following steps were performed to determine the associated *Topic* and *word tags* of each extracted CityNews tweets.

A. Bag of words

- Bag of words is a method that was used to classify the frequency or occurrence of each word in a tweet. The generated frequency was then used as feature in the successive steps.
- Gensim Corpora Dictionary was used to generate the dictionary. This dictionary captures the mapping of words and their integer id. In this project, tokens that appear in less than 15 documents were filtered out and then kept only the first 100K most frequent tokens.
- Gensim doc2bow (document to bag-of-words) was used to generate the word and their corresponding frequency.

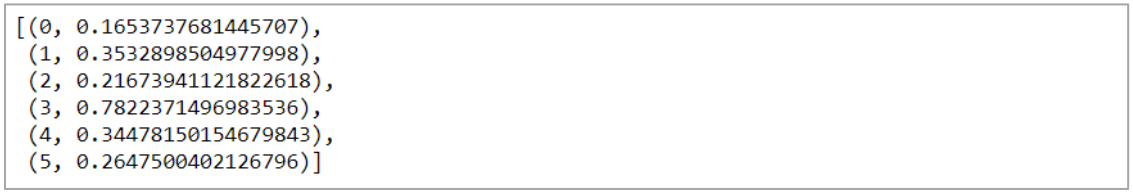


```
Word 25 "time" appears 2 time.
Word 84 "tuesday" appears 1 time.
Word 107 "long" appears 1 time.
Word 115 "died" appears 1 time.
Word 357 "military" appears 1 time.
```

Figure 1 *– Sample record that shows doc2bow output*

B. TF-IDF model

TFIDF or Term Frequency-Inverse Document Frequency measured the relevance of each word in a document. It worked by using two matrices in the equation (a) frequency of word and (b)inverse frequency of word. If the word appears in many documents, then the closer its value to 0 otherwise, closer to 1.

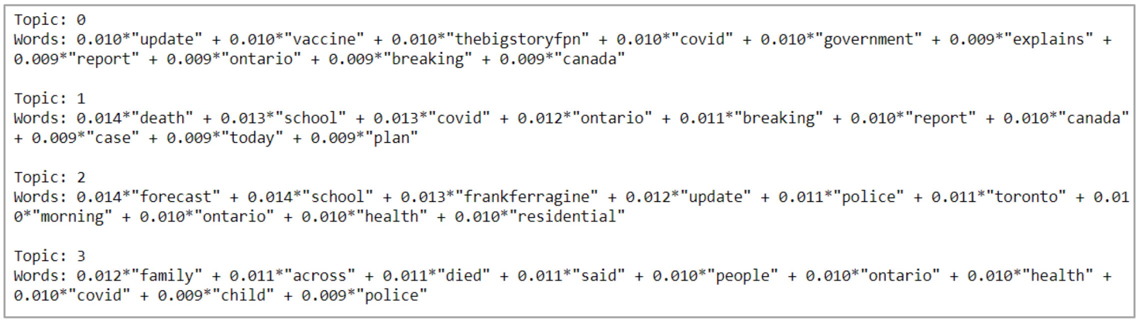


```
[(0, 0.1653737681445707),
 (1, 0.3532898504977998),
 (2, 0.21673941121822618),
 (3, 0.7822371496983536),
 (4, 0.34478150154679843),
 (5, 0.2647500402126796)]
```

Figure 2 *– Preview of TF-IDF score for the first document*

C. LDA

LDA or Latent Dirichlet Allocation was used to generate and classify which topic a tweet belongs to.

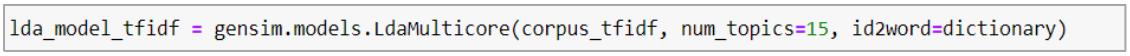


```
Topic: 0
Words: 0.010*"update" + 0.010*"vaccine" + 0.010*"thebigstoryfpn" + 0.010*"covid" + 0.010*"government" + 0.009*"explains" +
0.009*"report" + 0.009*"ontario" + 0.009*"breaking" + 0.009*"canada"

Topic: 1
Words: 0.014*"death" + 0.013*"school" + 0.013*"covid" + 0.012*"ontario" + 0.011*"breaking" + 0.010*"report" + 0.010*"canada"
+ 0.009*"case" + 0.009*"today" + 0.009*"plan"

Topic: 2
Words: 0.014*"forecast" + 0.014*"school" + 0.013*"frankferragine" + 0.012*"update" + 0.011*"police" + 0.011*"toronto" + 0.01
0*"morning" + 0.010*"ontario" + 0.010*"health" + 0.010*"residential"

Topic: 3
Words: 0.012*"family" + 0.011*"across" + 0.011*"died" + 0.011*"said" + 0.010*"people" + 0.010*"ontario" + 0.010*"health" +
0.010*"covid" + 0.009*"child" + 0.009*"police"
```

Figure 3 *– Preview of topics with keywords and their relative weight*

LDA worked by going through each tweet and randomly assign each word to one of the defined Topic #. In this project, number of topics was set to 15 and used the tf-idf score of each word from the dictionary generated previously.

Code Snippet:

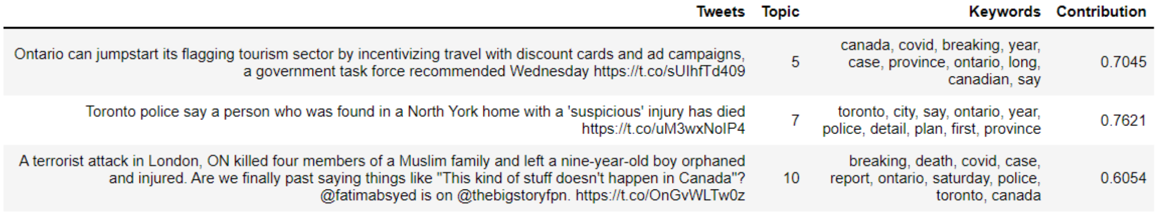


```
lda_model_tfidf = gensim.models.LdaMulticore(corpus_tfidf, num_topics=15, id2word=dictionary)
```

IV. Visualization

In this section, top five topics were calculated and visualized. The following method and steps were performed to generate the most accurate and logical result.

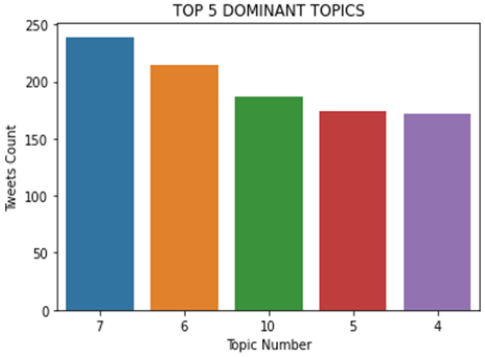
A. **Generate the df_topics table** – In this step, the trained LDA model was ran against the extracted tweets to identify their corresponding Topic. Keywords and their Contribution percentage across the documents were concatenated on the table.



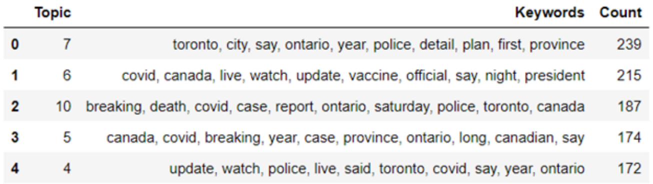
	Tweets	Topic	Keywords	Contribution
	Ontario can jumpstart its flagging tourism sector by incentivizing travel with discount cards and ad campaigns, a government task force recommended Wednesday https://t.co/sUlhfTd409	5	canada, covid, breaking, year, case, province, ontario, long, canadian, say	0.7045
	Toronto police say a person who was found in a North York home with a 'suspicious' injury has died https://t.co/uM3wxNoIP4	7	toronto, city, say, ontario, year, police, detail, plan, first, province	0.7621
	A terrorist attack in London, ON killed four members of a Muslim family and left a nine-year-old boy orphaned and injured. Are we finally past saying things like "This kind of stuff doesn't happen in Canada"? @fatimabsyed is on @thebigstoryfpn. https://t.co/OnGvWLTw0z	10	breaking, death, covid, case, report, ontario, saturday, police, toronto, canada	0.6054

Table 2 – *Three sample records of generated df_topics table*

B. Top 5 topics based on the dominance to the whole document

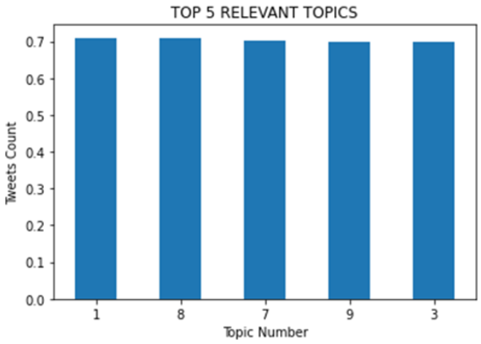


Top 5 topics based their value counts across the tweets dataset.

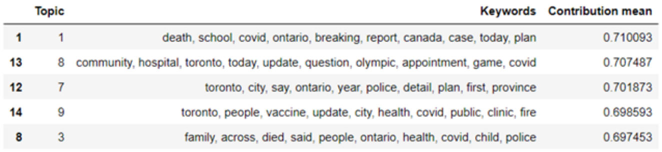


	Topic	Keywords	Count
0	7	toronto, city, say, ontario, year, police, detail, plan, first, province	239
1	6	covid, canada, live, watch, update, vaccine, official, say, night, president	215
2	10	breaking, death, covid, case, report, ontario, saturday, police, toronto, canada	187
3	5	canada, covid, breaking, year, case, province, ontario, long, canadian, say	174
4	4	update, watch, police, live, said, toronto, covid, say, year, ontario	172

C. Top 5 topics based on the average Contribution or weight to the whole document



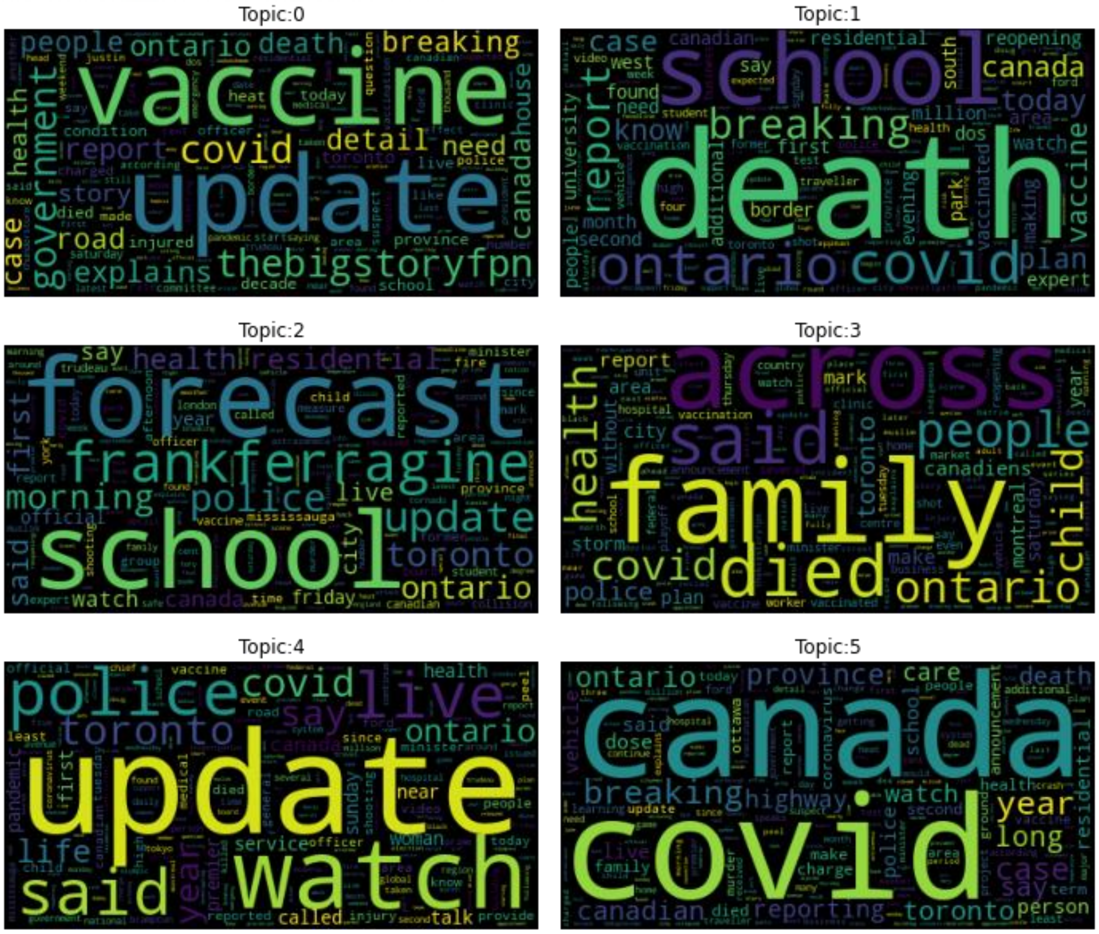
Top 5 topics based their average contribution or weight across the tweets dataset.



	Topic	Keywords	Contribution mean
1	1	death, school, covid, ontario, breaking, report, canada, case, today, plan	0.710093
13	8	community, hospital, toronto, today, update, question, olympic, appointment, game, covid	0.707487
12	7	toronto, city, say, ontario, year, police, detail, plan, first, province	0.701873
14	9	toronto, people, vaccine, update, city, health, covid, public, clinic, fire	0.698593
8	3	family, across, died, said, people, ontario, health, covid, child, police	0.697453

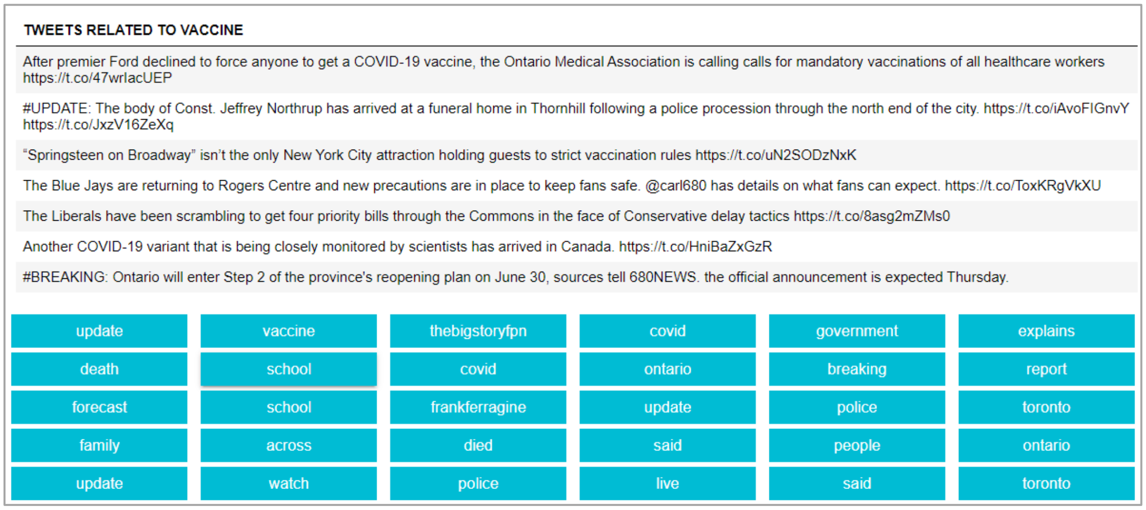
V. Word Cloud

Word cloud or tag cloud is a visualization of set of words based on their corresponding weight. In this project, WordCloud API was utilized to generate graphs per topic. The font of each word corresponds to the weights defined by the LDA model. Sample below -



A. Finding related tweets for each word

In this section, ipywidgets was applied to create an interface that consist of clickable and functional buttons. GridspecLayout allowed flexible layout which in this case 5x6 grid of 'info' buttons that pertains to each word per topic.



TWEETS RELATED TO VACCINE

After premier Ford declined to force anyone to get a COVID-19 vaccine, the Ontario Medical Association is calling calls for mandatory vaccinations of all healthcare workers https://t.co/47wrlacUEP

#UPDATE: The body of Const. Jeffrey Northrup has arrived at a funeral home in Thornhill following a police procession through the north end of the city. https://t.co/iAvoFIGnvY https://t.co/JxzV16ZeXq

"Springsteen on Broadway" isn't the only New York City attraction holding guests to strict vaccination rules https://t.co/uN2SODzNxK

The Blue Jays are returning to Rogers Centre and new precautions are in place to keep fans safe. @carl680 has details on what fans can expect. https://t.co/ToxKRgVkXU

The Liberals have been scrambling to get four priority bills through the Commons in the face of Conservative delay tactics https://t.co/8asg2mZMs0

Another COVID-19 variant that is being closely monitored by scientists has arrived in Canada. https://t.co/HniBaZxGzR

#BREAKING: Ontario will enter Step 2 of the province's reopening plan on June 30, sources tell 680NEWS. the official announcement is expected Thursday.

update	vaccine	thebigstoryfpn	covid	government	explains
death	school	covid	ontario	breaking	report
forecast	school	frankferragine	update	police	toronto
family	across	died	said	people	ontario
update	watch	police	live	said	toronto

***Table 3** – Shows the related tweets for each button pressed.*

References

- https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/
- https://ipywidgets.readthedocs.io/en/latest/