PANDEMIC: USING TOPIC MODELLING TO DISCOVER WORLDWIDE HEALTH ISSUES OVER TIME

Contents

Introduction

In this report, we will be attempting to discover the evolution of worldwide health issues over time. Most news stations from all over the world have a section dedicated to health, as it would have dedicated to sports, lifestyle, business etcetera. To aid our discovery, we will be analysing the health section of several news stations from several countries. The title of news articles allows us to briefly understand major health topics from a certain time period.

We believe that through the use of dynamic topic modelling and LDA models, we can discover how health topics change over time.

Dataset

The link to the dataset is downloadable from References. A copy of it is also attached with the submission.

The dataset contains health news from over 15 major health news agencies such as BBC, CBC, and CNN. Most of the datasets consist of 3000 entries that span over a period of 3 years from 2013 to 2015.

Every file contains tweets from a single health news agency. Each file is formatted in a text document, containing three main columns separated by " | " with values: Reference number, DateTime and Tweet respectively. A snapshot of the data as shown below:

```
585947808772960257|Wed Apr 08 23:30:18 +0000 2015|GP workload harming care - BMA poll http://bbc.in/1ChTBRv
585947807816650752|Wed Apr 08 23:30:18 +0000 2015|Short people's 'heart risk greater' http://bbc.in/1ChTANp
585866060991078401|Wed Apr 08 18:05:28 +0000 2015|New approach against HIV 'promising' http://bbc.in/1E6jAjt
```

Figure 1: Raw Data

For this project, we will only be exploring three out of the 15 provided datasets. The news agencies that we chose are BBC, CBC, and CNN, due to the location of the news stations. We found that most of the 15 news stations are situated in the United States. As such, we chose BBC (UK), CBC (Canada), and CNN (US) to compare the various news reported amongst locations as we believe that it will provide us with more insights.

Tools and Packages

For this project, we will deploy the following tools and packages:

1) Pandas, Numpy, matplotlib – Standard packages in python
2) Gensim – For topic modelling (LDA model, Coherence model, dictionary, corpus)
3) Nltk – For pre-processing (lemmatization, stop words)
4) pyLDAvis – To help us visualise our topic-word distribution
5) seaborn – To plot topic over time
6) wordcloud – To visualise our top 10 words for each topic
7) re – Regular expression to facilitate data cleaning
8) logging – (Optional) Enabling logging allows us to check the progress of our LDA models as LDA models often take a very long time to train

Data Cleaning

To start off, we imported our datasets using latin encoding due to the nature of the text files. We also set error_bad_lines = False as some of our rows of data contains four columns instead of three. Such rows consist of a very small proportion of our dataset. Hence, we chose to eliminate those rows.

We then made use of regular expression to remove redundant information in our titles, such as the pointers below.

1) Links (e.g. http://bbc.in/1ChTBRv)
2) Mentions (e.g. @CNN, @RT)
3) Links to images (e.g. ebolaChart.jpg)

The 2nd column of our data is changed to a pandas datetime object. A pandas datetime object makes it much easier for us to manipulate and sort. We grouped all the news articles by their months and sorted them by the date. The reason for grouping by month is because we want to observe how the topic changes over time, and we defined our period to be by months. As such, we will be able to observe how the topic changes over a period of 24 months instead of 730 days (if the dataset is two years long). Our window is now much more concise, making it easier to analyse visually.

We then renamed our dataset and removed the first column which provides a unique ID for each title. The resulting data frame is as follows:

	datetime	title
0	2013-09	C. diff 'manslaughter' inquiry call VIDEO: 'I...
1	2013-10	Death home saw 'institutional abuse' Peek-a-b...
2	2013-11	Study links synaesthesia to autism 'I thought...
3	2013-12	Indian women who are choosing to be child-free...
4	2014-01	Obesity measure 'too high' for many Routine o...

Figure 2: Data Frame after Cleaning

Pre-Processing

We defined a function to pre-process any data provided as an input parameter. In the function, we deploy several pre-processing techniques to prepare the data for topic modelling.

Tokenizing

We want to change our string of text to a vector of text. For instance, 'Topic smog prompts health warning' will be tokenized to ['Toxic', 'smog', 'prompts', 'health', 'warning']. This is necessary such that the machine can determine the importance of each word based on its frequency.

Stop words

We want to remove all the stop words that have no significant meaning. Nltk package very nicely provides us with a function that lists our all the stop words in the English dictionary. For instance, words like 'I', 'you', 'am', 'the' will be filtered out. Such words do not provide significant meaning to our topics and does not help us in inferring what our topics are.

Lemmatizing

This means any word would be in its root form. For instance, after lemmatizing, ['Playing', 'Plays', 'Played'] will be changed to 'Play'. This is to prevent several words of the same meaning to be deemed as different words. We chose lemmatizing over stemming even though lemmatizing generally takes longer than stemming, as we are not on a time crunch and value results over speed. Also, stemming simply strips the pre-fix and post-fix. This might cause some words to either lose its meaning (i.e. 'Witness' to 'Wit') or not exist in the English dictionary (i.e. 'Universal' to 'Univers'), whereas lemmatizing would guarantee a return of English words.

Lower Capitalization and Punctuation

All the words are changed to the lower capitalization form and punctuations are removed to ensure the smooth running of our LDA model.

Words less than 4 letters

We made sure that our dictionary consists of words that are of length four or higher. We believe that there are very few words in the English dictionary that are of length 3 and below that provide significant meaning.

Extra Pre-Processing

Extra stop words

After pre-processing, we went to look through our data again. We realised that there are certain words that occur very frequently and yet do not provide us with significant meaning. Such are words that are not detected by nltk stopword package. We extended this list of words to our stop words so that it will be filtered out during our pre-processing method. Example of such words are: 'video', 'audio', 'health', 'nh', 'com'.

Filter extremes

The purpose of our project is to understand how topics change over time. As such, we require our topics to be distinct and to have as little overlap as possible. We deployed a built-in function in Gensim to filter out extreme frequencies of our tokens. LDA is extremely dependent on the frequency of words used in the corpus and how frequently they show up. As such, we decided that tokens that appear in less than 20% and more than 50% of our total number of documents will be filtered out and not used to build our model. The 20% lower bound is to ensure that we do not allow words in too few documents to appear. The 50% upper bound is to reduce topics overlapping. For instance, 2013 to 2015 was the year where Ebola is heavily reported worldwide, and because we want to know other news that are happening besides Ebola, we removed words, like Ebola, that appeared in too many documents.

This is an example of how our data from BBC news channel looks like after pre-processing:

	datetime	title
0	2013-09	[diff, manslaughter, inquiry, call, thought, d...
1	2013-10	[death, home, institutional, abuse, peek, wind...
2	2013-11	[study, link, synaesthesia, autism, thought, t...
3	2013-12	[indian, woman, choosing, child, free, surgery...
4	2014-01	[obesity, measure, high, many, routine, north,...
5	2014-02	[child, adult, psychiatric, ward, drug, blind,...

Figure 3: Data Frame after Pre-Processing

Choosing Optimum Number of Topics

Coherence Model

We chose our optimal number of topics based on 2 principles.

Firstly, we want our models to be simple and not to have too many topics. For instance, having 20 topics would trigger topics overlapping, potentially producing inaccurate results.

Next, we aim for our model to have the highest coherence value. Coherence value measures the average of pairwise word similarity scores of words in a topic. The higher it is, the more accurate our model will be. As such, we tried nine different models, each with the number of topics set from two to ten respectively and plotted a graph of their coherence values, as shown in the figure below.

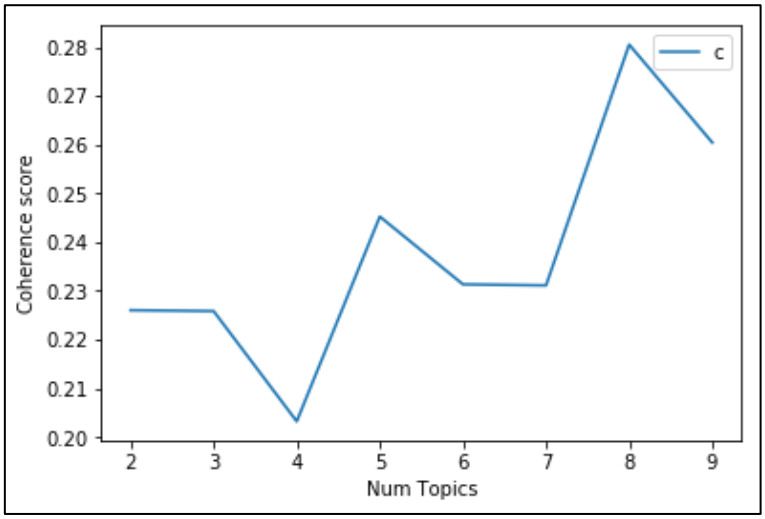

Figure 4: Coherence-Topic Graph

From the figure above, we observe that the optimal number of topics is eight, followed by five. We decided to choose five topics instead of eight despite a lower coherence value. This is because as we continued to build our model and infer our topics, we realise that some of the topics were not dominant topics in any document, and our algorithm is reliant on the fact that all the topics are dominant topics in at least one document. Also, if we were to set the number of topics to eight, there will be quite a few topics with overlaps. The topic similarity is rather low, and it is difficult for us to infer the topics based on the words they provide in each topic-word distribution.

LDA Hyperparameters

We also made 2 additional decisions in the parameters of our LDA model.

Firstly, it is the passes. Passes refer to how many times the model will go through the data to construct the topics. During the passes, the model will find the best *N* groups of word that occur together frequently and seem separate from other group of words. Such groups are our topics. We set it to 100 to balance between the quality of our model and the speed of our training.

Next, we set our random_state to 100. Random_state acts like a seed. This is to ensure that our model can be easily replicable in the future. Setting the random state is necessary due to the random state of LDA models.

Visualising Our LDA Model

We deployed a package – pyLDAvis – to help us visualise our model. An interactive copy of the visualisation is provided in the attached code. A snapshot of the visualisation is as follows:

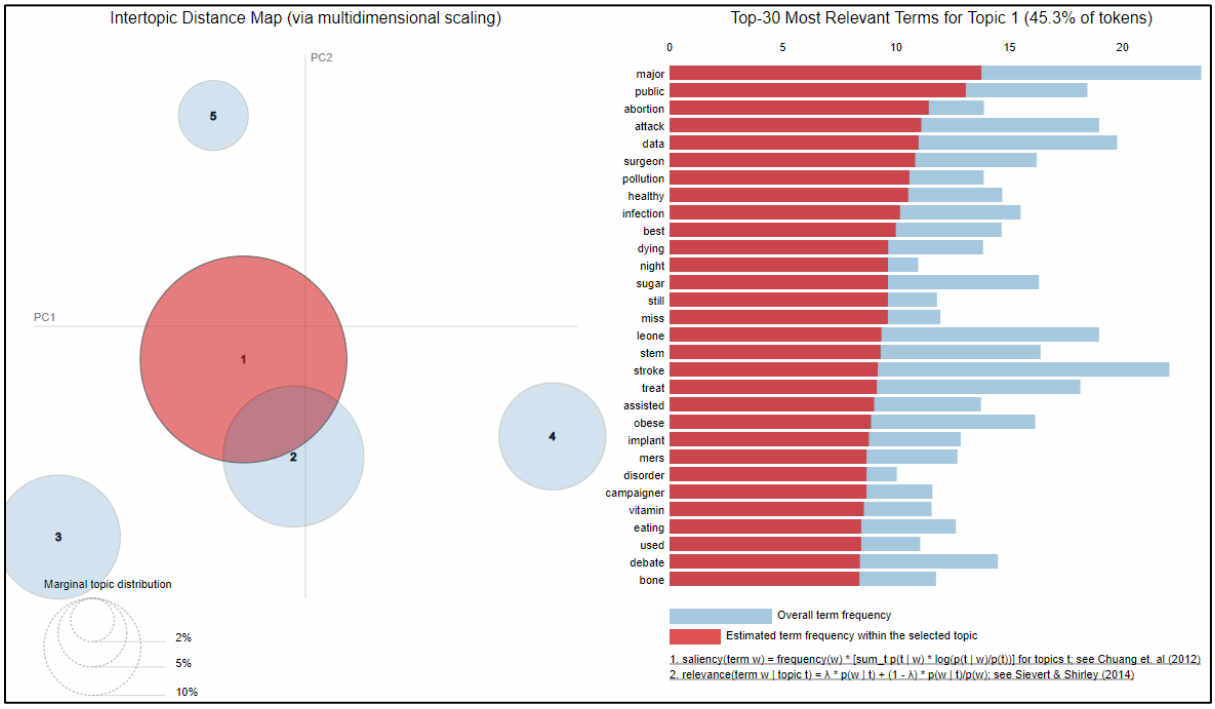

Figure 5: Visualisation with pyLDAvis

The three key information this visualisation provides are as follows.

Firstly, each circle represents a topic. The area of the circle is proportional to the percentage of contribution of each topic to the entire corpus. The larger the area, the more prominent the topic is in our corpus.

Next, the distance between circles is directly proportional to the topic similarity. We want to ensure that our topics are spaced out as much as possible and, also, to reduce the number of circles that are overlapping.

Lastly, the right side of the visualisation tells us the top 30 words that occur in each topic. The red bar represents the term frequency within the selected topic while the blue bar represents the overall term frequency. The bar allows us to ensure that each word not only occurs frequently in a single document, but also throughout the entire corpus, a similar concept as tf-idf values.

Topic Inference

This section describes how we came about inferring our topics.

Word Cloud

We generated word clouds to display the top ten words of each topic. Word clouds are very helpful as they are visually appealing and provide us with a rough idea of the topic with a simple glance.

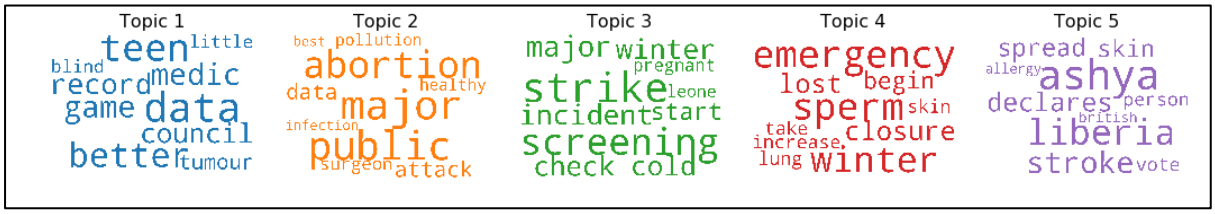

Figure 6: Word Cloud

Top 30 words

Using the top 30 words provided to us by pyLDAvis, we explored more of the top contributing words in each of the topics.

Context

Our data is pre-processed in such a way that it is easy for the machine to infer the topics, but difficult for us, as humans, to infer with tokenized words, while sentences provide us with more context, increasing its interpretability. As such, we went back to search for the documents that contributed the most to each topic and read the documents again. We also researched on some of the events, to figure out what some of the incidents or illnesses were about, to better infer our topics.

With a combination of the above 3 methods, we were able to infer the following topics:

```
Topic 1: Problematic teenagers

Topic 2: Controversy on Abortions

Topic 3: NHS Strike | Ebola Screening

Topic 4: Frozen sperm fight case (2012-2014)

Topic 5: Ashya King case (Proton Therapy) (2014-2015)
```

Figure 7: Topic Inference

Dominant Topics

For every document, we devised a method to figure out its dominant topic. A dominant topic refers to the topic with the highest percentage contribution to a document. Attached below is our data frame:

	date	Dominant_Topic	Topic_Perc_Contrib	keywords	title
0	2013-09	1.0	0.7319	major, public, abortion, attack, data, surgeon...	C. diff 'manslaughter' inquiry call VIDEO: 'I...
1	2013-10	1.0	0.9969	major, public, abortion, attack, data, surgeon...	Death home saw 'institutional abuse' Peek-a-b...
2	2013-11	3.0	0.9968	emergency, sperm, winter, closure, lost, begin...	Study links synaesthesia to autism 'I thought...
3	2013-12	1.0	0.6394	major, public, abortion, attack, data, surgeon...	Indian women who are choosing to be child-free...
4	2014-01	1.0	0.9963	major, public, abortion, attack, data, surgeon...	Obesity measure 'too high' for many Routine o...

Figure 8: Dominant Topic Data Frame

For instance, the first row of the data represents the document from 2013 September, whose most prominent topic is the controversy on abortions, topic 1, with the top ten words listed under keywords, and the original text listed under title.

Document-Topic Matrix

	date	keywords	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
0	2013-09	major, public, abortion, attack, data, surgeon...	0.06668	0.73193	0.06804	0.06667	0.06668
1	2013-10	major, public, abortion, attack, data, surgeon...	0.00078	0.99688	0.00078	0.00078	0.00078
2	2013-11	emergency, sperm, winter, closure, lost, begin...	0.00079	0.00080	0.00080	0.99681	0.00080
3	2013-12	major, public, abortion, attack, data, surgeon...	0.00085	0.63941	0.00085	0.35805	0.00085
4	2014-01	major, public, abortion, attack, data, surgeon...	0.00094	0.99625	0.00094	0.00094	0.00094
5	2014-02	data, teen, better, game, medic, record, counc...	0.99645	0.00089	0.00089	0.00089	0.00089

Figure 9: Document-Topic Matrix

Deriving the document-topic matrix allows us to observe the topic contributions for each document. In our context, each document represents each time frame. Going down the rows, we can see how each topic evolves over time. For instance, in figure 9, topic 1 showed very low significance in the corpus until 2014 February when it suddenly represents 99.6% of the corpus. We can repeat the observations for all the topics for the entire duration of the data collected to plot the evolution of topics over time.

Topic Over Time – BBC

We made use of a seaborn function – factorplot to plot several colour coded time series, where one colour represents one topic. The top ten words of each topic is provided in the legend, with each topic being aligned with the colours. The topic inferences have been conveniently pasted below the plot for reference. For instance, 'Problematic teenagers' is represented by blue, 'Controversy on Abortions' is represented by orange etc. The result is as follows:

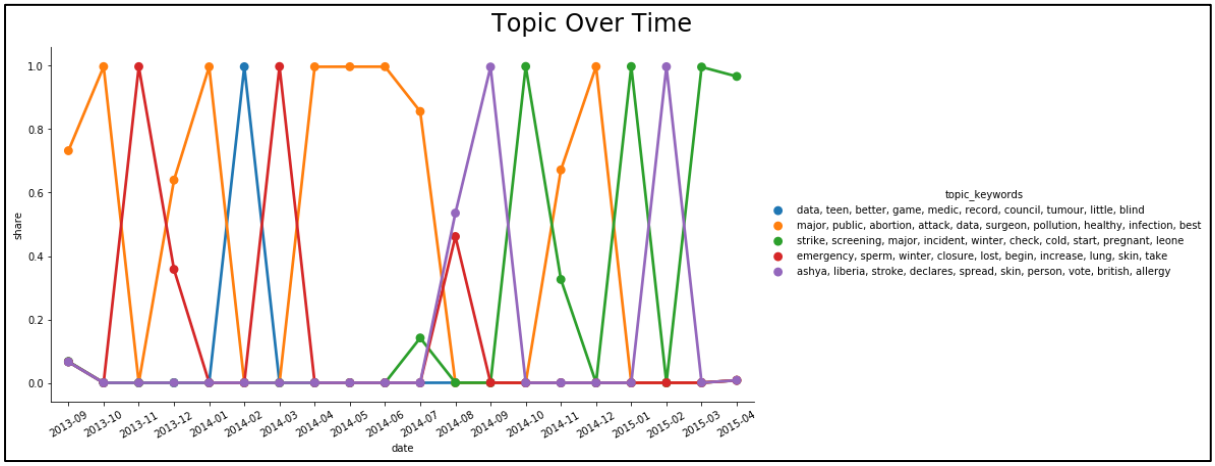

Figure 10: Topic over Time (BBC)

```
Topic 1: Problematic teenagers

Topic 2: Controversy on Abortions

Topic 3: NHS Strike | Ebola Screening

Topic 4: Frozen sperm fight case (2012-2014)

Topic 5: Ashya King case (Proton Therapy) (2014-2015)
```

Figure 11: Topic Inference (BBC)

From the above plot, we can derive several interesting insights.

General cases

Abortion is one of the topics that is consistently mentioned in BBC (UK news). Given that it is a controversy that has been around for ages, we would expect that this topic to not die down so easily. We can also see that it is of paramount importance to the UK health system since it appeared as the top topic consistently from 2013 to 2015. This consistent mentioning of the topic is different from topic 4 and 5, which are very specific cases that are not related to general health issues. As such, we will explore this in the next section.

Specific Cases

The Frozen Sperm Fight Case is a case that has been ongoing since 2012. It was concluded in 2014. The case describes a widow's legal battle to keep the frozen sperm of her dead husband, with ongoing debates within that period and it was frequently documented by BBC. From figure 10, following closely with the red line (topic 4), we see that the spikes are from 2013 to 2014, which is within the time frame of this entire saga.

The Ashya King Case is a case that has been ongoing since 2014. It was concluded around 2016. This case describes a parent struggle to treat their son. Ashya King is a boy with a brain tumour, whose parents wanted him to receive a treatment known as Proton Therapy. However, NHS (health system in UK) did not provide that treatment during that period. As a result, the parents flew their son to France for the treatment without the approval of the UK medical team. Consequently, there was a manhunt and the parents were sentenced to prison. This sparked several issues, one of which was the importance of communication between the health system and the parents. If we were to extend this model to the present, we can see if there has been another spike in this topic, which might possibly mean a similar case happening, and this might urge the NHS to improve on its communications with the public, if not already done so.

Topic Over Time – CBC

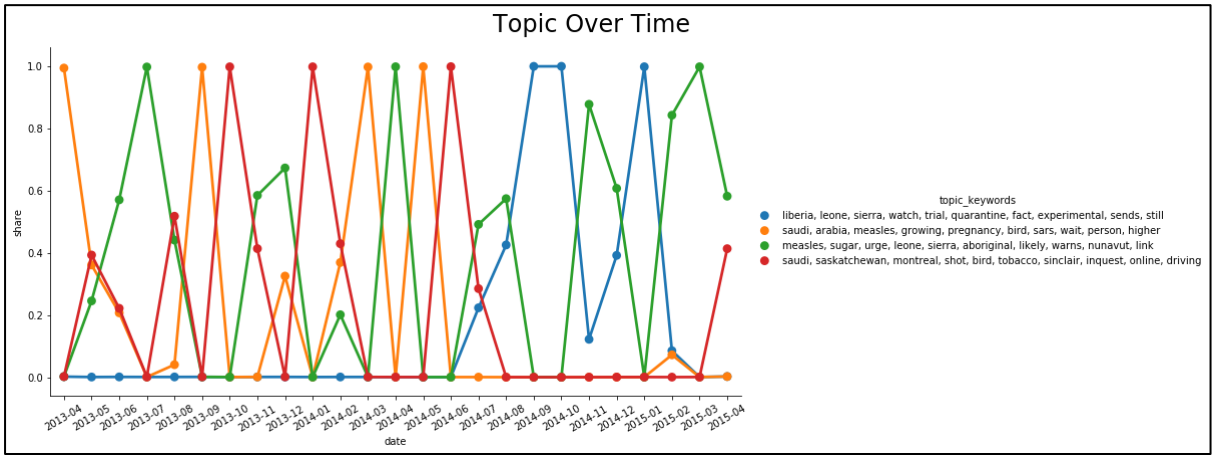

Figure 12: Topic over Time (CBC)

```
Topic 1: Providing aid to Africa

Topic 2: MERS virus | Measles

Topic 3: Concerns about Sugar | Measles

Topic 4: Sasakatchewen (province in Canada)
```

Figure 13: Topic Inference

A similar process was applied to the dataset on CBC. Though, this time, we have only four topics, with our insights below.

Worldwide Cases

Topic 1 and topic 2 mainly discussed about worldwide issues, with topic 1 mentioning the provision of aid from Canada to Africa – Liberia, Leone, and Sierra – due to Ebola. Also, further investigation of the original text allowed us to understand that Canada was providing aid throughout the duration of Ebola, which is represented by the blue line. If an outbreak were to hit Africa again, we can expect Canada to provide aid four months after the outbreak, as shown in figure 12.

Topic 2 mentioned about another worldwide virus, MERS. This is represented by the yellow line, which is in line with the period of MERS from 2013 to 2014. From figure 12, we see that the Ebola outbreak hit after MERS died down, which may urge health agencies to investigate any causal relations between the two viruses.

General cases

Topic 3, albeit similar to topic 2, mentioned issues concerning sugar consumption. As a result, we can see that the green line appeared quite consistently from 2013 to 2015 where we assumed that sugar consumption is a serious issue in Canada which received substantial coverage. Should the green line not tend to zero in the long run, suggestions to curb this issue will have to be made to health agencies.

Specific Cases

Topic 4 is about Sasakatchewen, a province in Canada. Several issues came up in that province, such as a professor discovering a superbug and the death of a resident due to a type of flu. We can see that it is a specific case as the coverage seems to have died down from 2014 august.

Topic Over Time – CNN

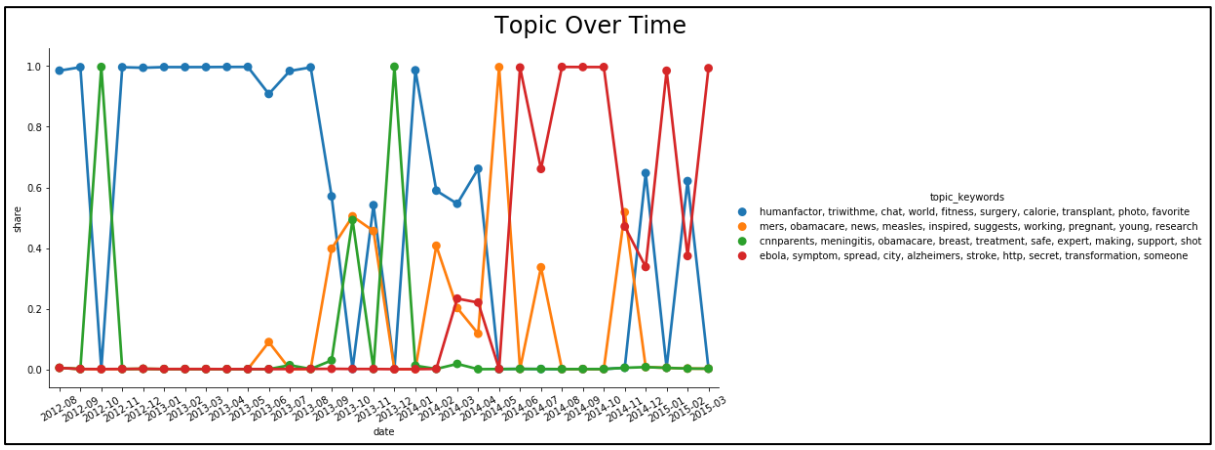

Figure 14: Topic over Time (CNN)

```
Topic 1: Getting Fit (Triathlon, Human Factor)

Topic 2: Obamacare | MERS virus

Topic 3: Mental Illness

Topic 4: Ebola
```

Figure 15: Topic Inference (CNN)

Lastly, we have our final news channel, CNN, situated in United States, Atlanta. Below are our insights:

Worldwide Cases

We would like to first mention about MERS. MERS is represented by topic 2, the yellow line. The period of outbreak is from 2013-07 to 2014-06, which is in line with our dataset for CBC. We can see how this is a worldwide issue as two major news channels are heavily mentioning MERS at the same time period.

Topic 4, the red line, belongs to Ebola. We previously mentioned that we want to remove major topics such as Ebola. However, we have to keep in mind that we did not specifically remove the topic on Ebola by adding Ebola as a stop word. Instead, we only filtered out extreme values. This means that the topic on Ebola is still within our desired frequency range, which is between 20% to 50% of our corpus. This might mean that there were very few mentions about Ebola as compared to other countries, where the mention on Ebola tend to exceed 50%.

Specific Cases

We found that there are several case-specific hashtags that appeared in our major topics. Two of which were #triwitme and #humanfactor that belong to topic 1, the blue line. It is a hashtag where a girl was trying to inspire people to join more triathlon in a pursuit to become healthier. Hashtags are often used in twitter in an attempt to make it into a trend such that people will follow the actions and re-tweet it, further contributing to the trend. Should we continue following this topic to the present, and notice an upward trend, we can also conclude that hashtag is an effective way to gain influence.

#cnnparents is a hashtag about parents being concerned about the mental health of their children. This is represented by topic 3, the green line. We tried to search online for a specific period where the hashtag was very popular. However, we couldn't come up with an exact timeframe. If we were to assume that our model is accurate, which it should be (since according to the above three topics over time plots, we found that most of the timeline coincides with real world events), we can infer this issue occurs every winter (September to December). We can see that the spikes usually occur towards the end of the year, nearing the holidays where children would be potentially stressed out by family gatherings, which is evident in the spikes in 2012 and 2013. Although there is no visible spike in 2014, we could always monitor the years after to look out for spikes during the winter season, and to also keep a closer eye on the children during winter.

Extensions and Limitations

Due to several limitations such as our knowledge, time and computational power, we were not able to do extensive research. However, below are some of the areas that could possibly be explored:

Full dataset

We could further explore the full 12 datasets instead of just three of them. If possible, we could also deploy Twitter API to collect more data from more news agencies. We can then explore the relationships between location of news agencies and topics over time.

Narrow datetime

In our project, we grouped our data by the month. However, health is a very volatile topic. A virus could easily emerge and might be cured within a month. As such, by narrowing our time period to weeks instead of months, we can perhaps derive even more hidden insights to account for the drastic change in topics.

Sentiment analysis

Other than simply deploying LDA to count the frequency of words, we can deploy sentiment analysis using techniques and models such as word2vec to analyse the tone of sentences in the tweet.

Conclusion and Further Thoughts

In this report, we explored the usage of topics over time with LDA to come up with insights. In the process, we went through data cleaning and pre-processing such that our model will cater to our model as closely as possible. A lot of visualisations were also deployed in the process as we feel that visualisations greatly boost the ability to infer trends and insights.

Through this project, we realised that LDA is not a magical tool. It is difficult to train and requires lots of hyperparameter tuning. Lots of manual work are involved as seen by the need to assess the model, interpret the topics, determine hyperparameters and filter out noisy data. Our interpretation of the topics is not exhaustive and anyone else could easily derive another set of topics. However, from our comparison of plots across 3 news channels, we can conclude that our method is sound, models are accurate and that there are no major discrepancies with real world contexts.

Nevertheless, there are several meaningful insights we obtained from LDA. From this project, in addition to our objective of simply discovering how worldwide health topic evolves over time, we also discovered several insights that was out of our initial scope. We discovered the pattern in how each news channel from different part of the world tend to report their news article. For instance, we realised that:

1) BBC tends to follow a saga closely, as seen by the frozen sperm fight case and Ashya King case,
2) CBC tends to report more on general health issues such as sugar consumption and worldwide issues such as major virus,
3) CNN tends to follow closely on trends. By trends, we are referring to hashtags deployed by netizen in an attempt to spread influence.

We also discovered that it is possible for us to determine the time period in which events are happening, even when there is no specific period mentioned online.

Inspiration

We want to thank Dr Qiao for introducing us to topic modelling. As Year Two students, we were not exposed to topic modelling before this module. Little did we know, it opens doors to many possibilities.

We were also inspired by "That's Mental!" Using LDA Topic Modeling to Investigate the Discourse on Mental Health over Time where they used another dataset to explore how mental health issues evolved over time.

References

1. Inspiration: https://towardsdatascience.com/thats-mental-using-lda-topic-modeling-to-investigate-the-discourse-on-mental-health-over-time-11da252259c3
2. Dataset: http://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter
3. Pre-processing: https://www.programcreek.com/python/example/107282/nltk.stem.WordNetLemmatizer
4. Dominant topic: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
5. Word cloud: https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/
6. TOT: http://jeriwieringa.com/2017/06/21/Calculating-and-Visualizing-Topic-Significance-over-Time-Part-1/
7. Conclusion: https://medium.com/@alexisperrier/topic-modeling-of-twitter-timelines-in-python-bb91fa90d98d