

COMPSCI5078

Web Science

Topic Modelling Coursework Report

Submitted by:

Abhishek Sinha

GUID 2705491S

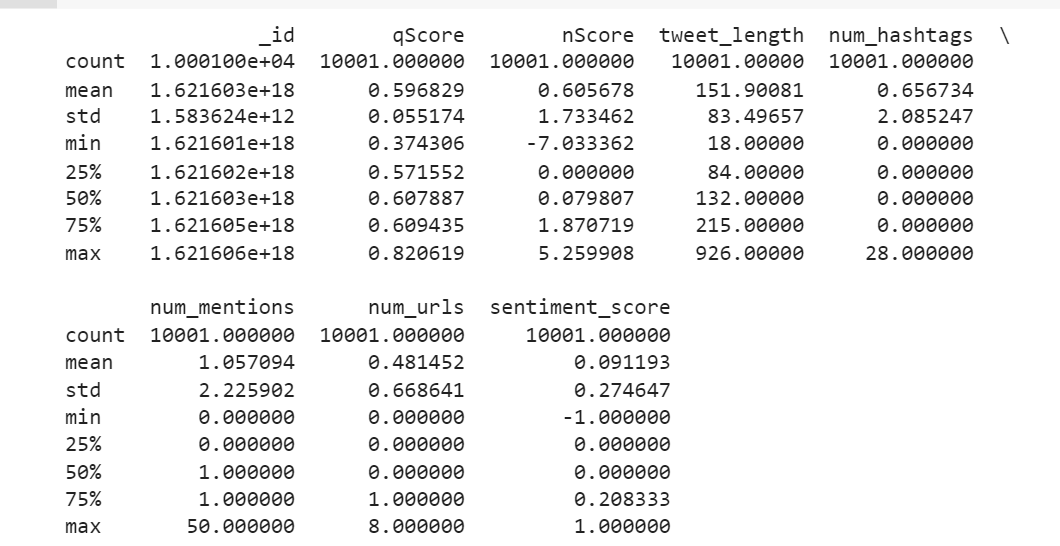
# Statistical Analysis of Tweet Data

* The number of items in the data is 10001
* Number of tweets: 10001
* Minimum tweet length: 1
* Maximum tweet length: 46
* Average tweet length: 12.13138686131387
* To perform the statistical analysis on the data i used number of statistical approaches such as calculating the number of hash tags , length of each tweet, number of mentions , number of urls and the sentiment analysis of tweets.As per the below table the detailed description of each variable is displayed.

| **Column Name** | **Description** |
| --- | --- |
| \_id | Unique ID of the tweet |
| qScore | Quality score of the tweet |
| nScore | Normalized score of the tweet |
| tweet\_length | Length of the tweet (in characters) |
| num\_hashtags | Number of hashtags used in the tweet |
| num\_mentions | Number of mentions used in the tweet |
| num\_urls | Number of URLs or links included in the tweet |
| sentiment\_score | Sentiment score of the tweet using a NLP tool (range -1 to 1) |

f**ig 1.1**

* After adding these particular columns to my original dataset describe function was used of the python to to get a better understanding of each added characteristics
* This statistical ensures the preprocessing step to be perfectly precise and help in tokeinsation of the text as only the weighted words will be kept after removing every other unnecessary character.
* Also the sentiment analysis of the tweet gives us a general idea about the the positive negative of neutrality of text.



**From the updated data frame,following information can be looked at:**

* tweet\_length: The average tweet length is 152 characters, with a standard deviation of 83 characters. The shortest tweet is 18 characters long, while the longest tweet is 926 characters long.
* num\_hashtags: On average, tweets contain 0.66 hashtags, with a standard deviation of 2.09 hashtags. The minimum number of hashtags used in a tweet is 0, while the maximum number of hashtags used in a tweet is 28.
* num\_mentions: The average number of mentions in a tweet is 1.06, with a standard deviation of 2.23. The minimum number of mentions used in a tweet is 0, while the maximum number of mentions used in a tweet is 50.
* num\_urls: The average number of URLs or links in a tweet is 0.48, with a standard deviation of 0.67. The minimum number of URLs or links used in a tweet is 0, while the maximum number of URLs or links used in a tweet is 8.
* sentiment\_score: The average sentiment score of the tweets is 0.09, with a standard deviation of 0.27. The sentiment score ranges from -1 to 1 that is from negative to positive , with 0 indicating a neutral sentiment.

# Topic modelling

**Data Analysis:**

* The first step is procuring the given data and perform a statistical analysis on the data to better understand the nature of data I.e in the given case analysing the tweet data.

**Pre-Processing and Tokenization :**

* After successfully analysing the data the next step is to perform pre processing the text such as stop words removal , removing special characters , url mentions and after successfully cleaning the data converting each individual text into tokens and representing them as individual terms.

**Document Term Matrix:**

converting each row of the tweets and column of word or phrases into a matrix that describes the number of word that appears in each tweet.

**Evaluation Metrics** :

Many evaluation metrics can be used such as coherence kl divergence and perplexity , these valuation metrics can be better used to analyse the tweets moreover The level of "uncertainty" in a model's prediction outcome is taken into consideration by the perplexity value, which is a confusion measure. The coherence score, in contrast, reveals the degree of semantic similarity across terms related to a subject

Visualization:

This is one of the most important step in topic modelling as it helps us to analyse the data in a more visual manner such as using tag clouds also graph can be a good way to analyse the topics occurrence and their usage.

Result Analysis :

Analysis of the obtained result

# Preprocessing Tweet Text :

Text preprocessing ensures the data contains only the words that have weights and can add to the over all text analysis, the tweet data has been a tough challenge to process as for some entries “ **@ ,# and /n** “ and the keyword token had to be removed for which I used a custom function that removes the @ symbol only without removing the important key word attached to it.

In order to achieve this task the NLP library was used that uses **t.text.**



The major task provided in the preprocessing are listed below that ensures the data to be in a tokenised form and be used for topic modelling in further sections :

* This function lower-cases every character in the text.
* Removes any preceding or trailing white space from the text using this function.
* Removes any digits (0–9) from the text using the function.
* Punctuation removal: The function takes away all punctuation from the text.
* Stop words are words that frequently appear in sentences, such as "the," "and," "is," etc. The function removes any stop words from the text.
* Lemmatization: This function lemmatizes the text, reducing words to their dictionary or basic forms. Running, for instance, would be shortened to "ran". command at this stage to limit the language to only nouns, adjectives, verbs, and adverbs.

**Evaluation Metrics :**

In natural language processing, evaluation metrics are used to measure the effectiveness of models or algorithms on a particular task. The task and the model's goal will determine the evaluation measures to use.

For example, in topic modelling, commonly used evaluation metrics include coherence, perplexity, and topic diversity. Coherence is a metric used to gauge how coherent the topics produced by a model are. It is frequently calculated as the average pairwise coherence of the top terms in each topic. Lower numbers indicate greater performance when it comes to perplexity, which assesses how well the model guesses a set of documents that have been held out. The degree to which the generated subjects are unique and do not overlap is measured by topic diversity.

**Perplexity**:

* Perplexity is a statistical measure of how effectively a language model predicts a text sample in natural language processing. It is frequently employed as a language model evaluation metric.
* Perplexity assesses how effectively a probability distribution or language model predicts a set of samples. The model does better at predicting the data when the perplexity score is lower. For instance, a perplexity score of 2 indicates that the model is equally perplexed as if it were required to uniformly select from two potential outcomes.
* Perplexity is a metric for how effectively a topic modeling with LDA model can forecast the distribution of terms in the documents..

**Coherence**:

* A topic model's output topics' coherence is a metric for how comprehensible and significant they are. It assesses the degree of semantic similarity between the high probability terms in a particular topic. A topic with a higher coherence score has words that are more tightly related to one another, making it more comprehensible and meaningful.
* There are various forms of coherence metrics, such as c v coherence, u mass coherence, and cosine similarity coherence between topic vectors. The specific use case and the type of data can influence the coherence metric that is chosen

**Kullback-Leibler divergence (KLD)**:

* Kullback-Leibler divergence (KLD) is a metric for comparing how two probability distributions differ from one another. KLD can be used in topic modeling to compare a model's topic distribution to a reference distribution, such as a topic distribution created by a human or a topic distribution from a previous study.
* One can compare the topic distribution of various models for various numbers of subjects when using KLD for topic modeling. If and only if two topics are identical, the KLD is non-negative and equal to zero.
* KLD can be a valuable metric for topic modeling, as it can provide a detailed measure of how well the model matches the reference distribution.

**Visualisation :**

Word Cloud :



The word could clearly highlights the majority of most important words or the top words in the dataset with larger font and the less effective words in shorter fonts

This depiction helps us to recognize the topics and their meaening by just looking at them.

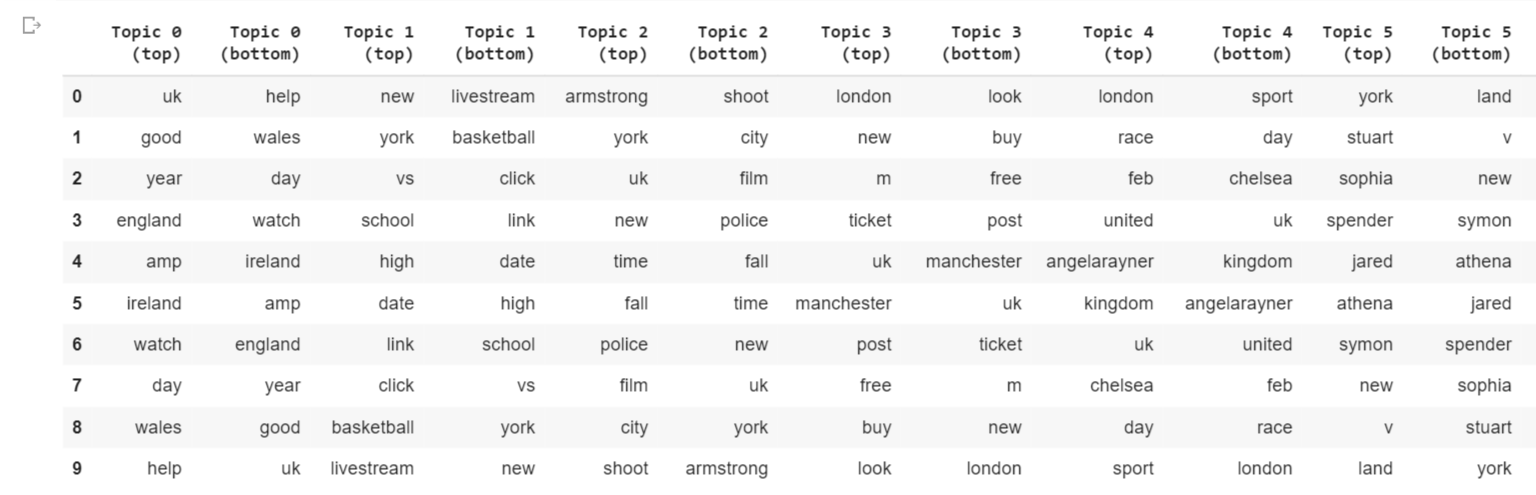
For example the topic 0 : england , good year and watch are clearly reognizable and helps to get the overall sense of the topic.

Metrics studied & discussion

**Top and bottom words**

The top and bottom words in topic modeling aid in understanding the content and overall idea of each topic. The most significant and frequently occurring words in a topic appear at the top, while the least significant and infrequently occurring terms appear at the bottom.

We may learn more about the issue and the keywords that are most pertinent to it by glancing through the top terms. Yet, the bottom words may indicate less significant terms that aren't necessarily as pertinent to the subject.

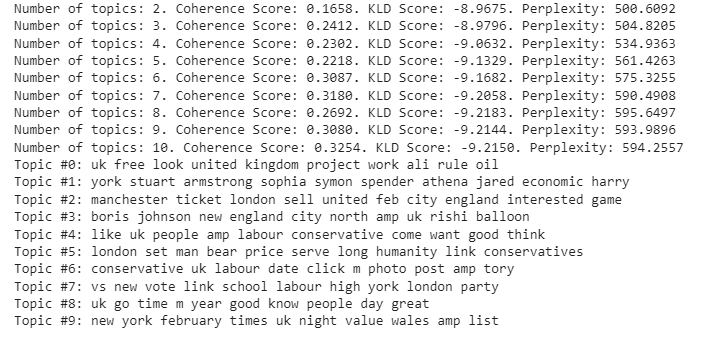


**Topic Selection**

Topic selection for selecting the number of topics based on the evaluation metrics of the tokenised tweets obtained generally higher coherence score and lower KLD score is desirable , it was noticed that at the time of training of LDA model the value of KLD score remains relatively stable for all values of k, so it doesn't provide a clear indication of the best number of topics.

The coherence value is highest for k equals to 10 It's also worth noting that the perplexity decreases as the number of topics increases. However perplexity is not always a reliable metric for topic modelling.

Coherence score with value 0.3254 gives a good analysis of the topic value to be 10 and use it as the appropriate number of topics to further derive meaning from them



Based on the coherence score, the best number of topics appears to be around 7 to 10. However, we should also consider the interpret ability and relevance of the topics in order to determine whether they are good or bad.

**Topic 0** relates to the UK's policies and regulations, with keywords such as "united kingdom", "rule", and "project work". This topic could be relevant for people who are interested in UK politics or businesses.

**Topic 1** relates to economics, with keywords such as "economic" and "spender". However, some of the keywords like "stuart" and "sophia" are not very informative, and it is not entirely clear what this topic is about.

**Topic 2**  relates to sports and events in Manchester and London, with keywords such as "ticket", "manchester", and "game". This topic could relevant for people who are interested in sports particularly football events in these cities.

**Topic 3** relates to UK politics and its new prime minister and old prime minister, with keywords such as "boris johnson", "england", and "rishi". This topic could potentially be relevant for people who are interested in UK politics or government.

**Topic 4** related to UK politics and opinions, with keywords such as "labour", "conservative", and "like". This topic could potentially be relevant for people who are interested in UK politics or public opinions.

**Topic 5** related to London and its values, with keywords such as "price", "humanity", and "conservatives". This topic could potentially be relevant for people who are interested in London's values or politics.

**Topic 6** related to UK politics and parties, with keywords such as "conservative", "labour", and "tory". This topic could potentially be relevant for people who are interested in UK politics or parties.

**Topic 7** related to UK politics and education, with keywords such as "vote", "school", and "labour". This topic could potentially be relevant for people who are interested in UK politics or education.

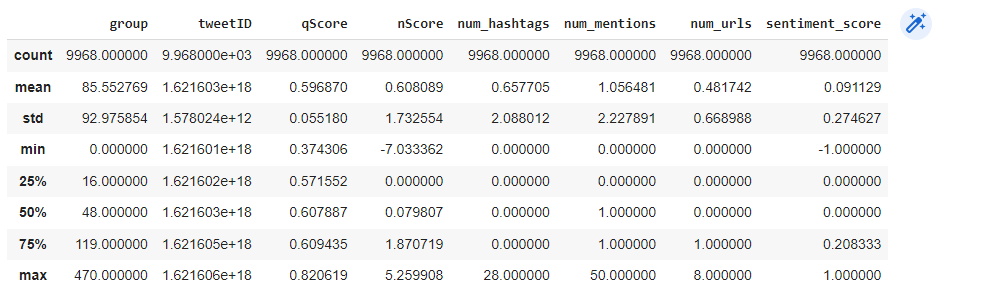
Conclusion:

The model overall gives a good picture of topic modelling as the topic and their meaning are easily interpretable.

Grouped Tweets

# Statistical Analysis

For the statistical analysis of the data num\_hashtags , num\_mentions , num\_urls have been added in the dataframe to more pro actively study the data wand with that different statistics measure is calculated in order to get better understanding of the data.



The statistical analysis of the dataset shows that:

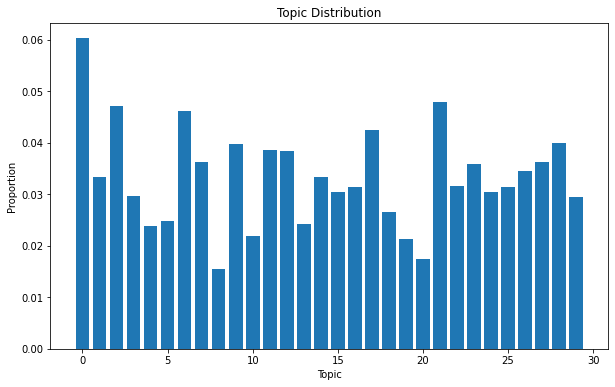
* The dataset contains 9968 tweets that have been grouped based on a specific topic.
* The mean value of the 'qScore' column, which provides the quality score of the tweets, is 0.596870. The standard deviation is 0.055180.
* The mean value of the 'nScore' column, which provides the novelty score of the tweets, is 0.608089. The standard deviation is 1.732554.
* The mean value of the 'num\_hashtags' column, which provides the number of hashtags used in each tweet, is 0.657705. The standard deviation is 2.088012.
* The mean value of the 'num\_mentions' column, which provides the number of mentions used in each tweet, is 1.056481. The standard deviation is 2.227891.
* The mean value of the 'num\_urls' column, which provides the number of URLs used in each tweet, is 0.481742. The standard deviation is 0.668988.
* The mean value of the 'sentiment\_score' column, which provides the sentiment score of the tweets, is 0.091129. The standard deviation is 0.274627.
* The minimum and maximum values for the 'qScore' column are 0.374306 and 0.820619, respectively.
* The minimum and maximum values for the 'nScore' column are -7.033362 and 5.259908, respectively.
* The minimum and maximum values for the 'num\_hashtags' column are 0 and 28, respectively and they help us to recognize the subsequesnt distribution of hashtags across the tweet data.
* The minimum and maximum values for the 'num\_mentions' column are 0 and 50, respectively as it can be seen there are tweets with withi no mention this can help to neglect or minimize the tweets on the basis of mentions.
* The minimum and maximum values for the 'num\_urls' column are 0 and 8, respectively. So there tweets with no urls.the analysis shows the tweets that may or may not contain the hyperlink
* The minimum and maximum values for the 'sentiment\_score' column are -1.0 and 1.0, respectively.which gives highly negative and positive tweets and can be refrained from certain set of audience.

**Visualization of the topics**

Visualization helps us to effectively and isntantly understand the topic nature and what the topic wants to convey here we use two sets of statistical analysis namely bar chart and word cloud where bar chart is used to get an overall pictire of the model nd what it is conveying

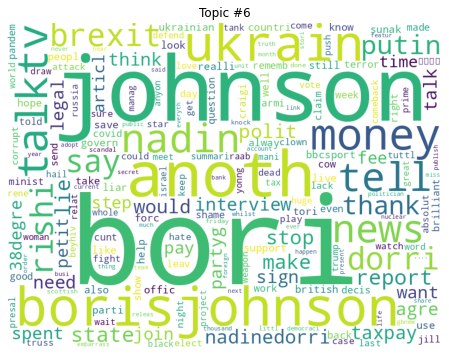
Word loud helps us to properly identify the meaning behind each topics and the context of it.

Bar Chart:



Bar chart clearly shows the measure in topic and their proportion.

**Word Cloud :**



Topic Selection

The program uses Latent Dirichlet Allocation for topic modeling (LDA). The three evaluation metrics used are coherence score, Kullback-Leibler Divergence (KLD) score, and perplexity score. It involves training an LDA model on a range of topic numbers (5 to 30 in steps of 5) and assessing the models. Based on the highest average score across the three measures, the model with the most subjects is picked.

The coherence score evaluates the topic' semantic coherence. It evaluates the degree to which words within a topic are related to one another and the degree to which topics are distinct from one another. Better topics are those with higher coherence scores.

By comparing the probability distribution of the original dataset and the distribution of the model's topics, the KLD score evaluates the model's quality. Better models are indicated by lower KLD scores.

The model's ability to predict the held-out data is indicated by the perplexity score. Better models have lower confusion scores.

The code selects the optimal amount of subjects based on the highest average score across the three criteria after analyzing the models. Lastly, it prints the topics produced after training an LDA model with the optimal amount of topics.

The topic selection can be detailed as follows :

* **Transportation**: flight, airport, London, Heathrow, link, plane, callsign : gives us breif about transportation and flight for london heathrow airport
* **Work**: medical job, care, assist, worker, visa : Gives us a analysis of the medical jobs and requirement of visa.
* **Politics**: conservative, liberal, country, right, news: it gives overal a topic discussion for poilitcal news.
* Unknown: call, answer, seen, precqp, pitchero, preston: unclear on whatthe topic states
* **Entertainment**: play, Dublin, award, radio, watch : gives us a breif on the play or award in dublin and is being played on radio
* **Social Interaction:** well, mate, luck, women, kid : gives us family oriented analysis of the topic.
* **Politics**: race, great, minister, prime, power, year: Talks about the prime minister and race towards political power.
* **Unknown**: free, work, govern, fall, Kabul, done, guardian, service, sack.
* **Music**: right, Tory, totp
* **City**: city, London, Manchester, Bristol, transport, transit : gives us a breif idea about the transit in london and tranportation from bristol to manchester.
* **Business**: ticket, sell, night, interest, people, 2023 : presents with the business of ticket being sold for the year 2023
* **Sports**: race, China, balloon, open, service, week, bull
* **Economy**: good, really, best, live, would, make
* **Politics**: need, Rishi, Sunak, help, interview, bank, West, cut, Johnson, make: gives us many idea about the ex prime minter boris johnson and current rishi sunak.

It is unclear what the unknown topics (topics 4 and 8) are about, as the associated words do not seem to be strongly associate with other words but rest of the topics paint a very good picture on what topic analysis look like.

**Comparison Of the Two models**

In the first model, when the number of topics rises, the coherence score and KLD score typically rise as well. The coherence score, however, only significantly rises up to three issues before fluctuating within a similar range for the remaining topics. When the number of topics increases, the perplexity score declines, as expected.

* The 10 subjects in the first model don't seem to have a consistent theme or coherence. Although several of the subjects share terms or ideas, such as "conservative," "labour," "uk," and "vote" in topic 6 and "new york," "february," and "times" in topic 9, the themes do not appear to be particularly well-defined or significant overall.
* The coherence score and KLD score in the second model both peak at about 2-3 topics and then start to decline as the number of subjects rises. When the number of subjects increases, the perplexity score similarly rises, suggesting a lesser quality in the model. However compared to the previous model, the coherence scores are typically higher, indicating that the themes are more meaningful and coherent.
* The 30 issues in the second model appear to have more focused and consistent themes. Topic 19 appears to be about Bristol medical careers, subject 17 appears to be about Dublin entertainment, and topic 24 appears to be about personal experience and guidance. Several of the subjects, though, continue to be less meaningful, such as topic 28 which seems to be a mix of different concepts. Overall, the second model seems to have produced more coherent and meaningful topics than the first model.

**Topic Modelling:**

To find themes within a corpus of text, such as news stories or social media messages, utilize the first model of topic modeling. The topics "versus new vote link school labor high york london party" and "uk free look united kingdom project work ali rule oil," for instance, can be connected to a project the UK government is working on about free trade with other nations. Companies or organizations trying to understand public opinion or follow trends about a specific subject may find this information beneficial.

**Text classification:**

The second model, which was improved for 30 subjects and had a higher coherence score, may be employed. For instance, a company may use the themes the model discovered to train a classifier to categorize social media messages as favorable, bad, or neutral. The topic "england east medicaljob sale fan care bristol brivsal assist english" may be about healthcare or job prospects, while the theme "good realli greatest live mate luck woman kid would make" may be suggestive of a positive mood.

**Content recommendation:**

Depending on the user's interests, both models could be used to suggest related content to them. An article or post on "race amazing happi minist prime power year birthday project pakistan" can be suggested to a user who enjoys sports, for instance. Similar to how articles or posts about "ticket pleas london sell night interest peopl 2023 anyon look" could be suggested if a person has an interest in travel.

**Conclusion :**   
  
In the first model output, the coherence score and KLD score are quite low for all the number of topics tried. This suggests that the subjects are not well segregated and that the model is unable to identify coherent topics. The model performs poorly at predicting the words in the corpus, as seen by the high perplexity score.

In comparison, the second model's output displays stronger coherence and KLD scores for each topic's number, demonstrating the model's superior ability to recognize coherent and clearly delineated topic. Also, the model performs better at predicting the terms in the dataset because of the low perplexity score.

Hence, the model output with the relevant topic is the second one.