**LING 583 – PROJECT REOPRT**

PROJECT #1 – Conspiracy Theory

**Data:**

Shape - (17155, 2)

Columns – date (in type datetime[ns]), text (in type object)

Date Range - 01-01-2015 to 17-02-2023

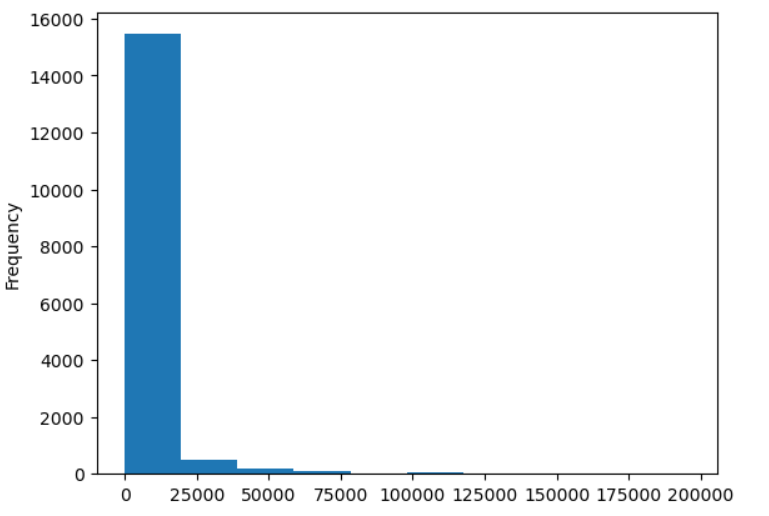
**Feature Engineering:**

len – calculates the length of string in text

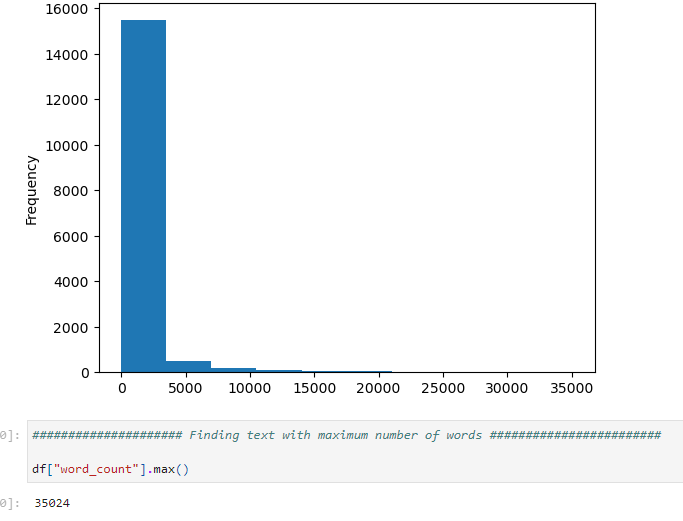
word\_count – calculates the number of words in text

**Observations:**

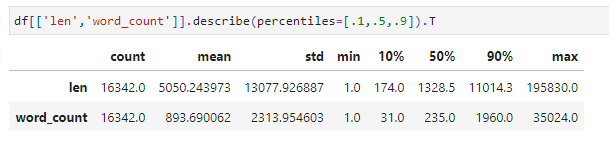
On plotting a *histogram of len,* we observe that length of text is mostly less than 25000. However, there are outliers that go as high as > 100000.



For word\_count, the histogram shows that the number of words in the text is mostly less than 2500. However, the text with maximum number of words has **35024 words.**

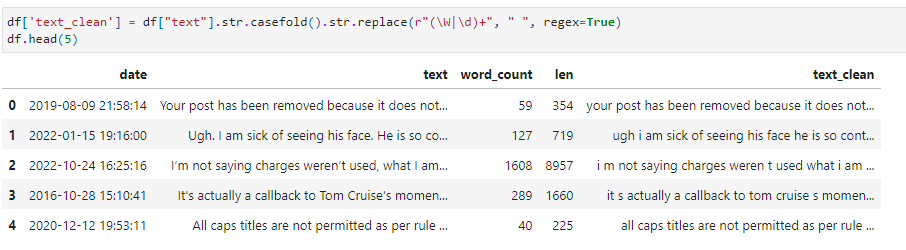
****

Only 10% of text has less than **31 words** and 90% of text has less than **1960 words.**



**Data Cleaning:**

* Removing leading spaces. 
* Removing emojis and any other component that is not a word or a number.

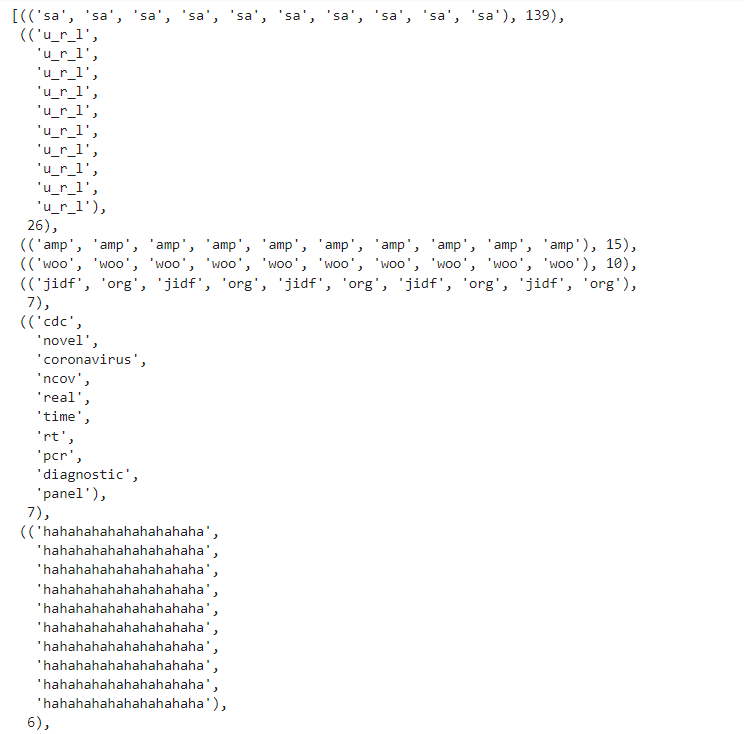


**Term Extraction:**

**Spacy** model "en\_core\_web\_sm" has been used for term extract along with the *Matcher* feature it provides.

We are trying to detect a pattern that begins with an adjective or a noun followed by singular/ plural common nouns or proper nouns along with hypen.

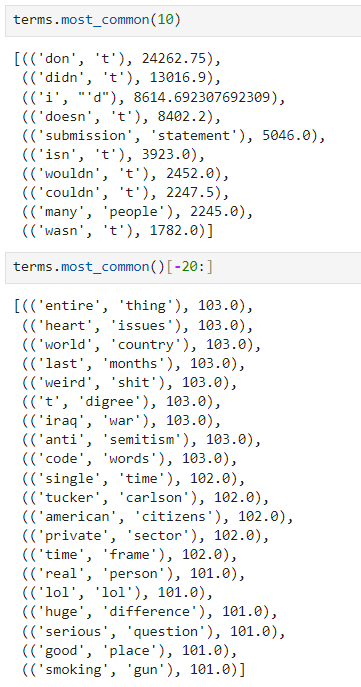
\*\* From the extracted terms it is noticed that there are terms in text than do not contribute to our analysis like



So, we invoke a second round of **Data Cleaning**  that removes words like

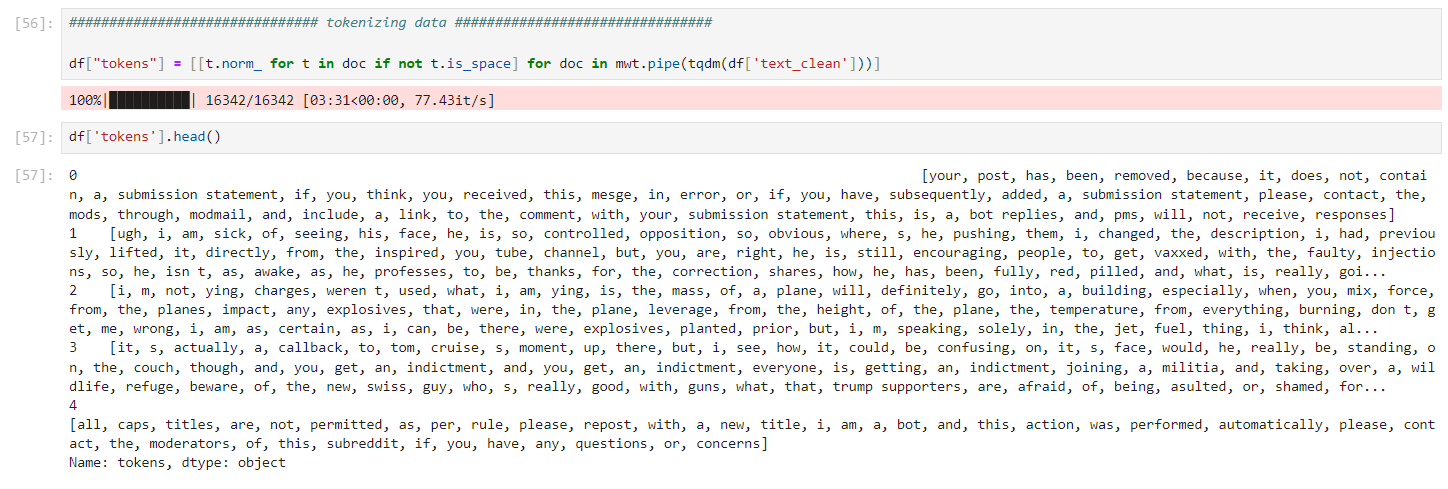


This is followed by term extraction using **C values** with **theta = 100.** The list of ten most common terms and 20 least common terms has been provided below.



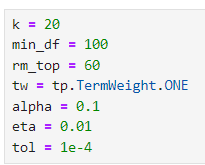
**Tokenization:**

We create tokens from text data using Spacy pipeline incorporating the terms as created above.

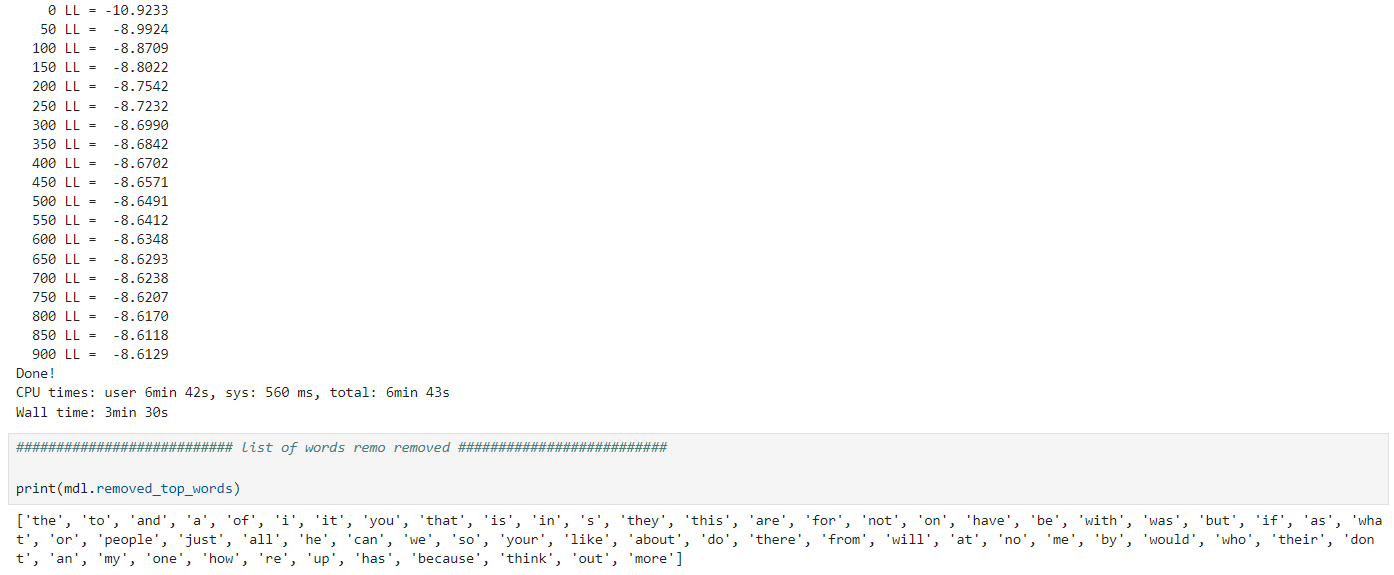


**Topic Modelling – LDA**:

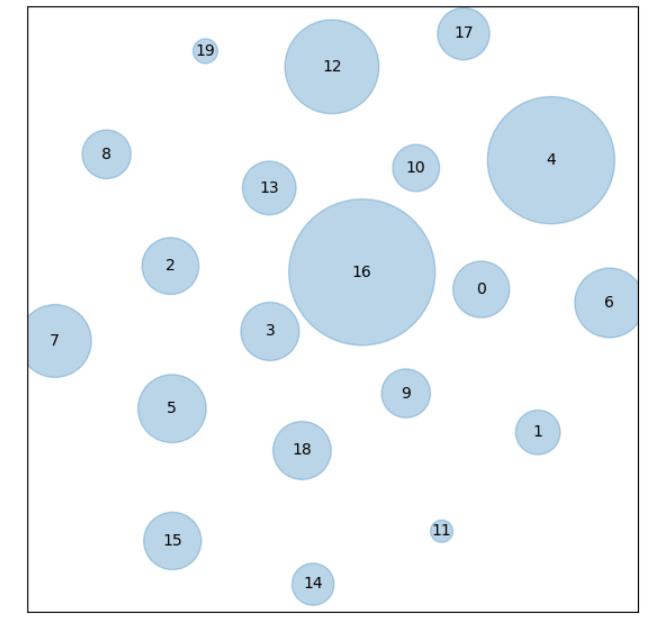
We have used the **LDA model** from **tomotopy** with the following features



We want the cleaned and tokenized texts to be modelled under **20 topics** where words occur in at **100 times** across the data and **remove the top 60 words** which do not make any significant contribution to our analysis.

The model training and the top words removed are shown below 

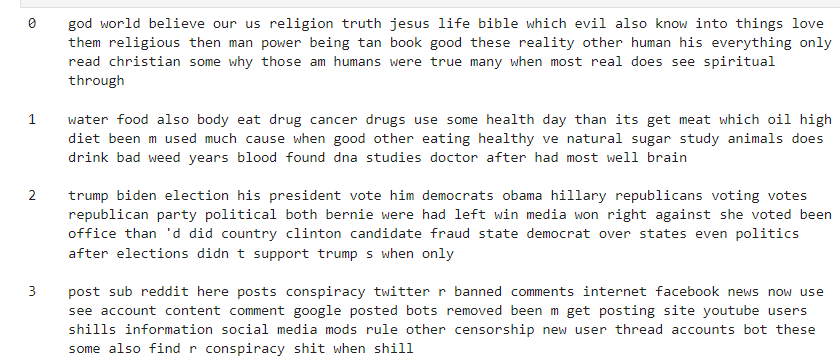
The twenty topics are plotted on a 2D space as below:

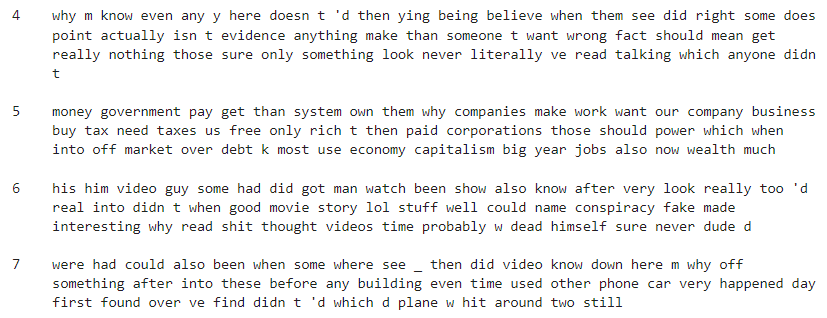


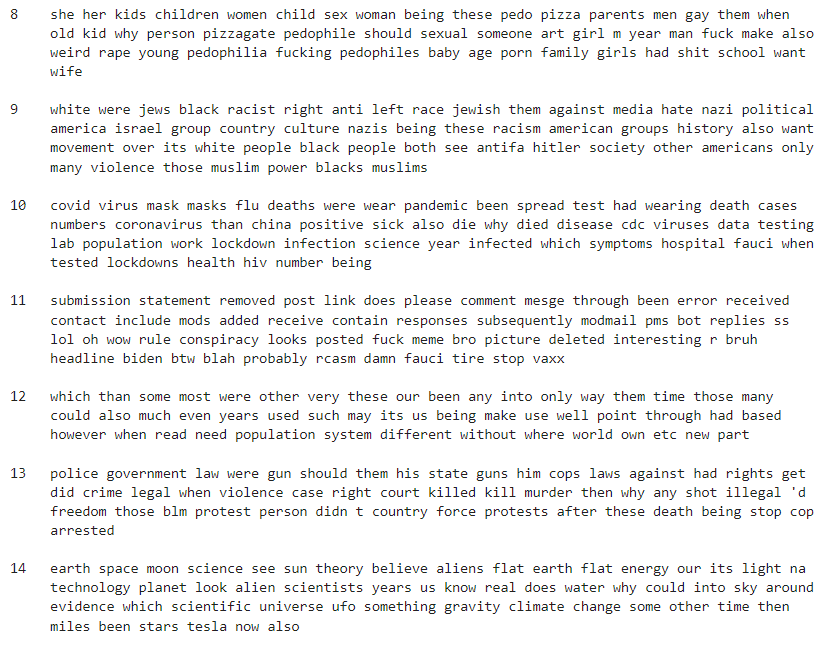
**Observations:**

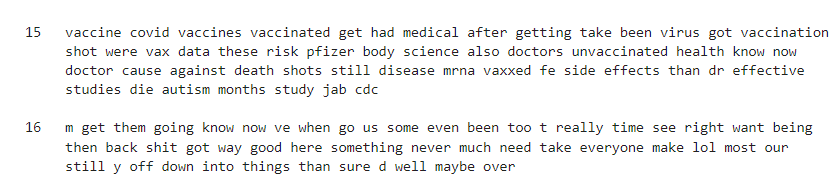
The topics are well distributed in the 2D space with no overlap between them.

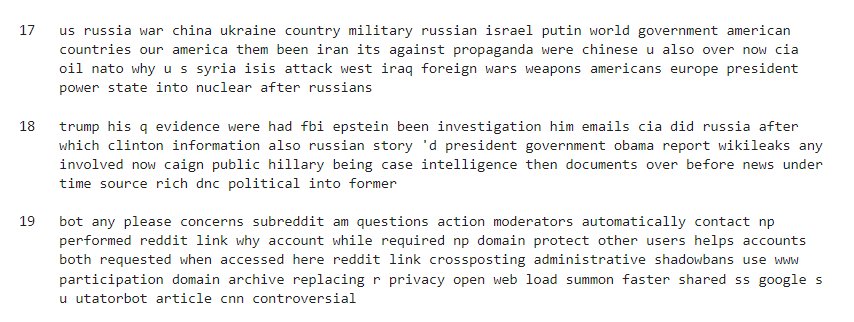
Topics 16, 4, 12 and 17 seems to have a higher share of text.







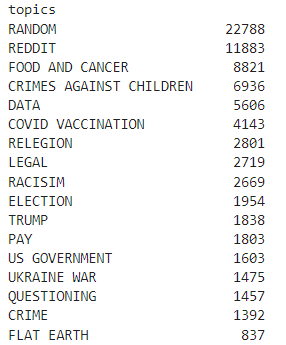




|  |  |  |
| --- | --- | --- |
| **TOPIC NUMBER** | **INTERPRETATION** | **LABLES** |
| 0 | Religion, God and Spirituality | Religion |
| 1 | Food Safety | Food and Cancer |
| 2 | Presidential Election | Election |
| 3 | Social Media Conspiracy | Data |
| 4 | Challenging Conspiracy Theories | Questioning |
| 5 | Government and Capitalism | US Government |
| 6 and 7 | Video (*Unclear*) | Random |
| 8 | Sexual Content | Crimes against Children |
| 9 | Racial Violence | Racism |
| 10 | Covid 19 – Test and Masks | Covid Vaccination |
| 11 | Reddit Submission Related (*Not Significant*) | Reddit |
| 12 | (*Unclear*) | Random |
| 13 | Police | Legal |
| 14 | Earth and Space | Flat Earth |
| 15 | Covid Vaccines | Covid Vaccination |
| 16 | (*Unclear*) | Random |
| 17 | Russia Ukraine War | Ukraine War |
| 18 | Trump and Election | Trump |
| 19 | Reddit Submission Related (*Not Significant*) | Reddit |

For more details, please see attached notebook.

Frequency of occurrence of each topic:

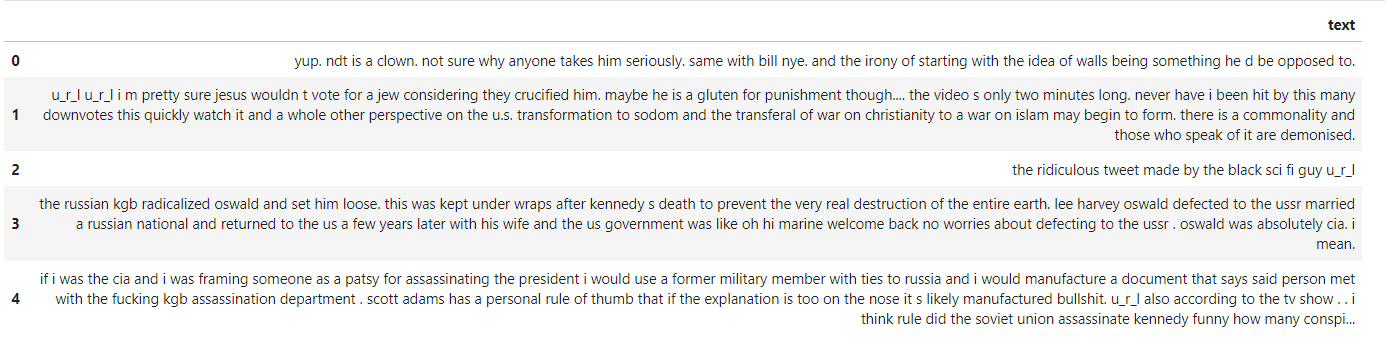


**Clustering:**

* The “text” data has been cleaned to remove emojis and unnecessary text and punctuation expect “.” as required for sentence tokenization.
* We have removed all data that has **less than 5 words.**
* We have removed data that begins with **"your post"** or **“please contact”** to remove reddit submission messages.

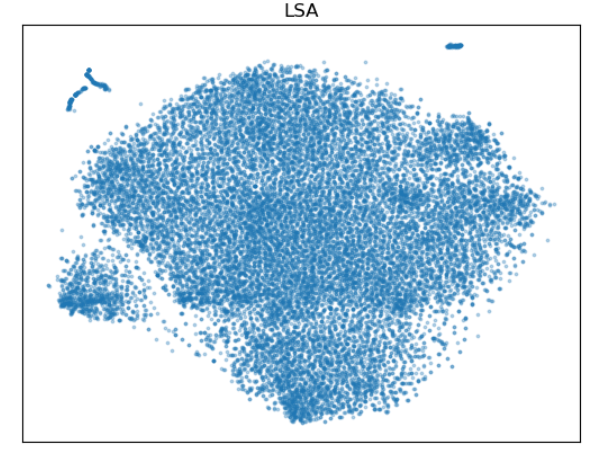
**Sentence Splitter:**

We have used sentence splitter from Spacy to take a **batch of 5 sentences** at a time from each post.



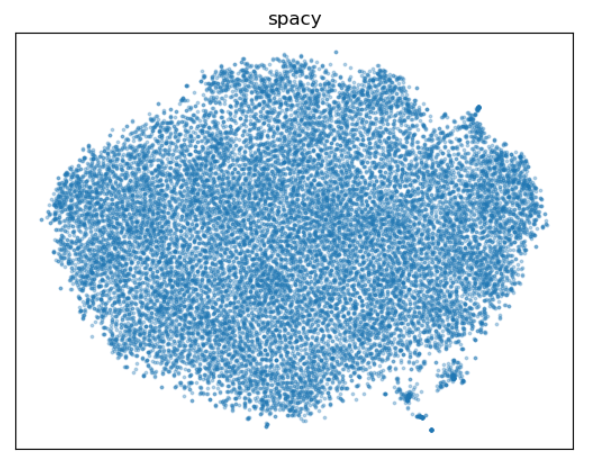
**SVD:**

We have used **TruncatedSVD** to transform the data into a **20 dimensional vector.**



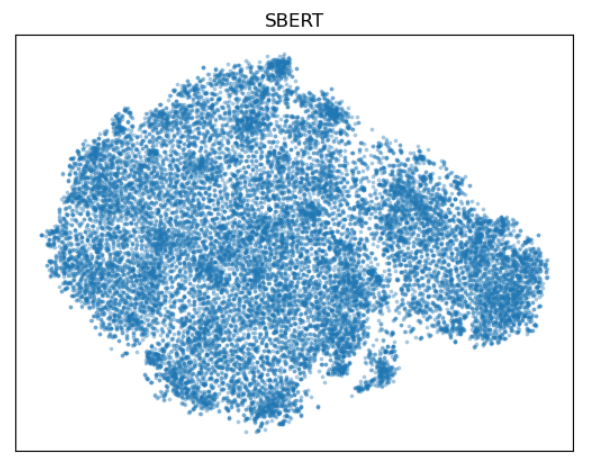
**SPACY:**

We have used the **"en\_core\_web\_lg"** from **Spacy** to transform the cleaned text into a **300-dimension vector**. Further dimension reduction has been done using TruncatedSVD to bring it down to 20 dimension.



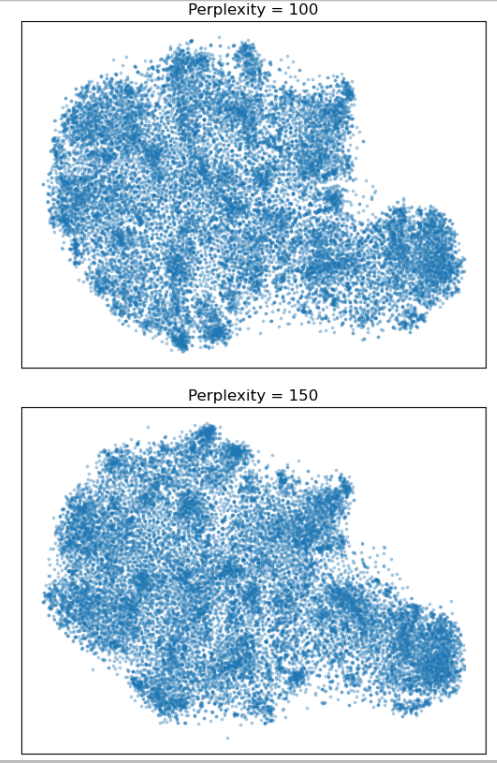
**SBERT:**

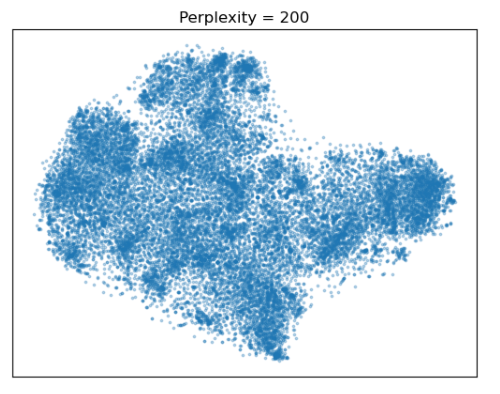
Dynamic embedding was done using SBERT sentence transformer from **"all-mpnet-base-v2"** toobtain a **768 dimension vector.** . Further dimension reduction has been done using TruncatedSVD to bring it down to 20 dimensions.



**PERPLEXITY:**

As seen above, with Perplexity=50, the SBERT model clearly produces better results as compared to LSA and Spacy. Thus, we tried to plot the vector projections with perplexity 100, 150 and 200, to compare results and select best model for clustering.

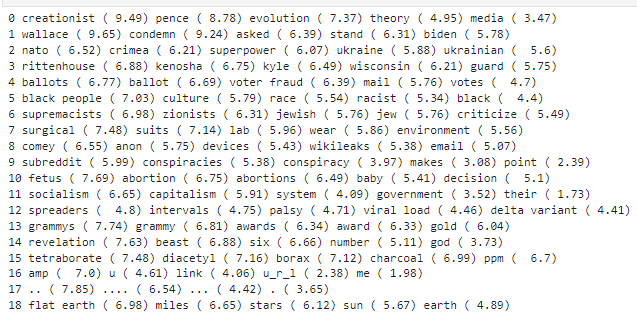




Perplexity of 200 seems to provide optimal results for the analysis.

**HDBSCAN CLUSTERING:**

We have considered a **cluster size of 20** and fitted the SBERT vectors.



Highlighted terms in each cluster after incorporating c-values gives us the following topics as focus

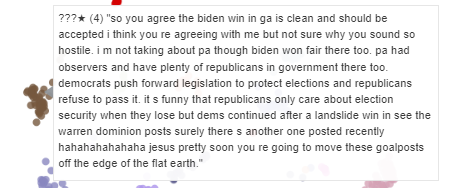
|  |  |
| --- | --- |
| **CLUSTER** | **TOPIC** |
| 0 | Evolution theory |
| 1 | Presidential Election |
| 2 | Ukraine War |
| 3 | Racism (Black Lives Matter) |
| 4 | Voter Fraud (Presidential Election) |
| 5 | Racism |
| 6 | Israel Crisis |
| 7 | Lab? |
| 8 | James Comey(wikileaks and Clinton emails) |
| 9 | Reddit and Conspiracy |
| 10 | Abortion Laws |
| 11 | Government (Capitalism vs Socialism) |
| 12 | Covid 19 (Delta Varient) |
| 13 | Grammy Awards |
| 14 | Religion - Christianity |
| 15 | Chemicals? |
| 16 | Random |
| 17 | Random |
| 18 | Falt Earth |



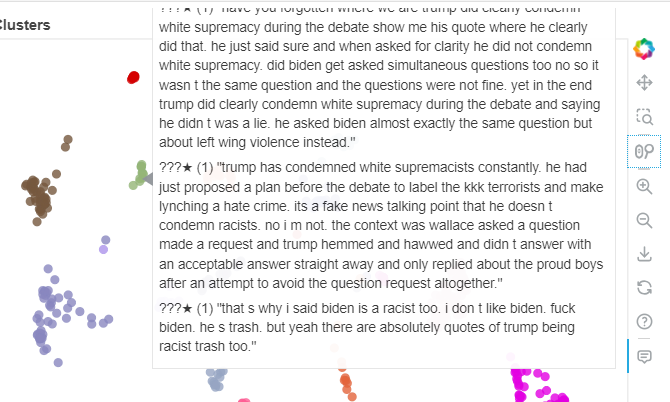
**Results :**

Through the above analysis the key topics identified for the conspiracy theories are as follows

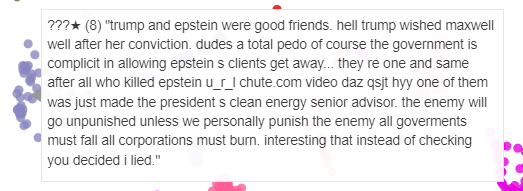
1. **Presidential Election Voter Fraud**

****

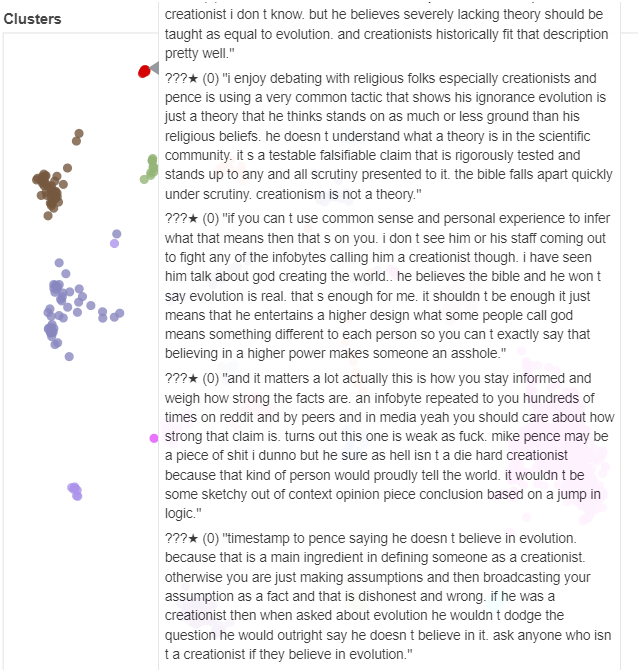
1. **Racism in America**



1. **Trump Controversaries – Jeffery Epstein Scandal**

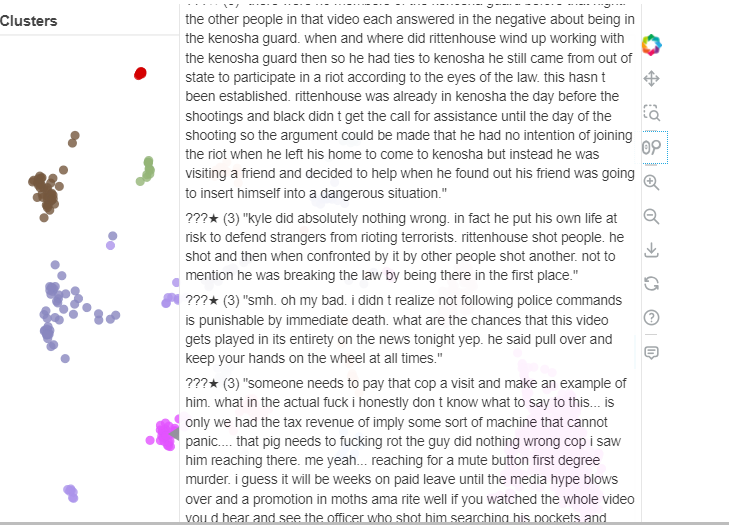


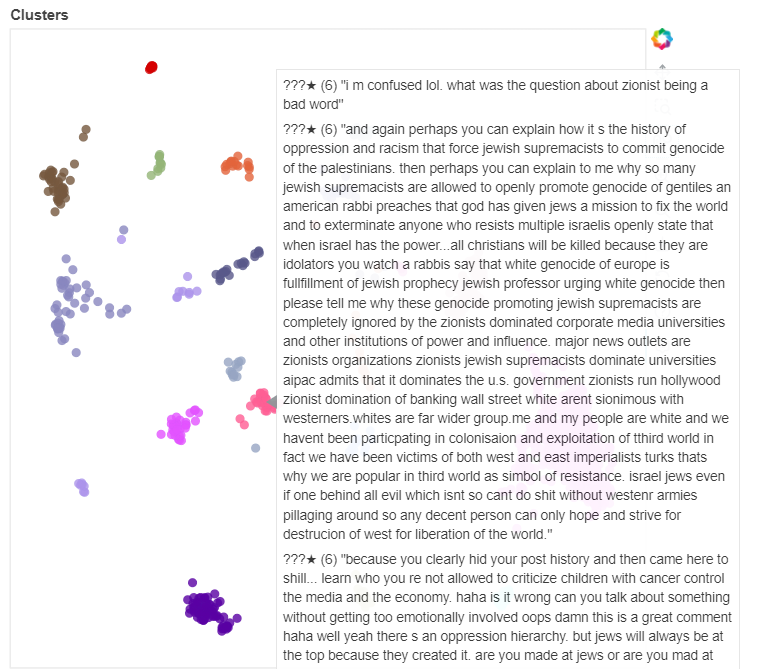
1. **Evolution Theory vs Religious Beliefs**



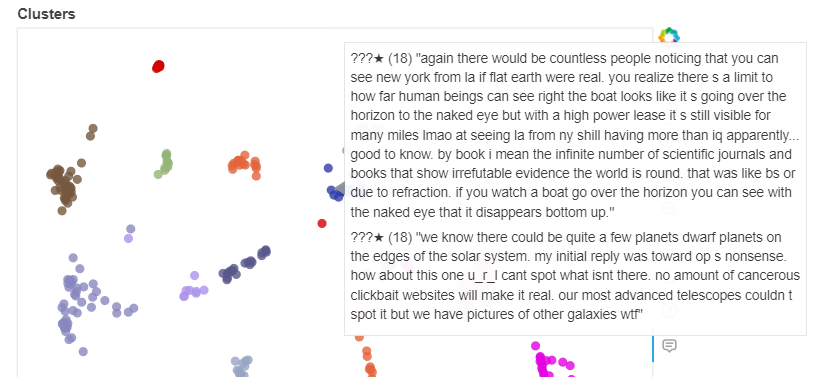
1. **Government – Socialism vs Capitalism**



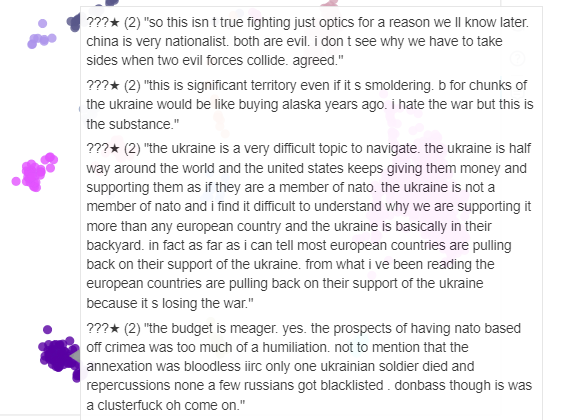
1. **Kyle Rittenhouse**
2. **Zionists – Israel and Palestine Crisis**



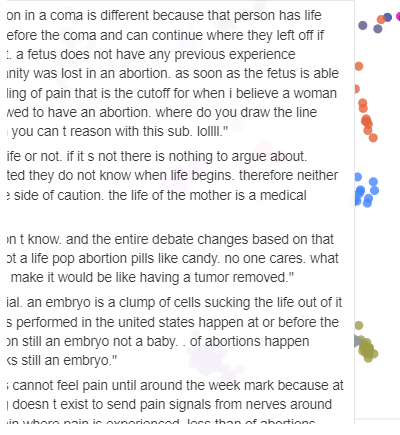
1. **Flat Earth**



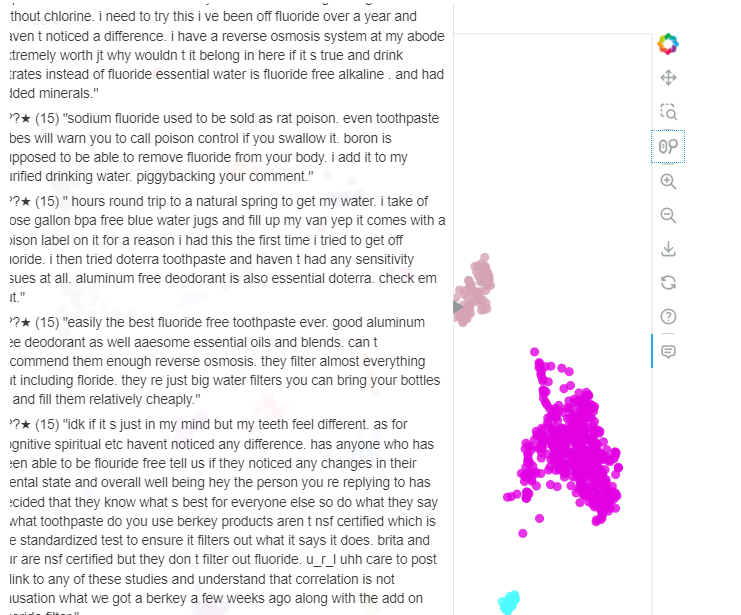
1. **Ukraine War**



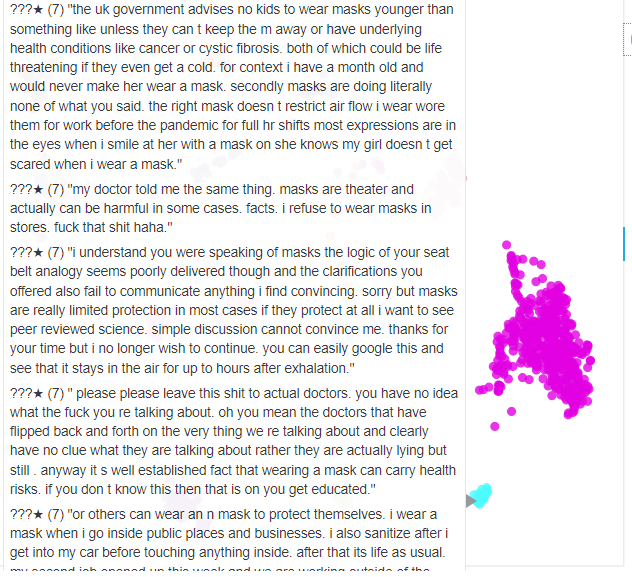
1. **Abortion Laws**



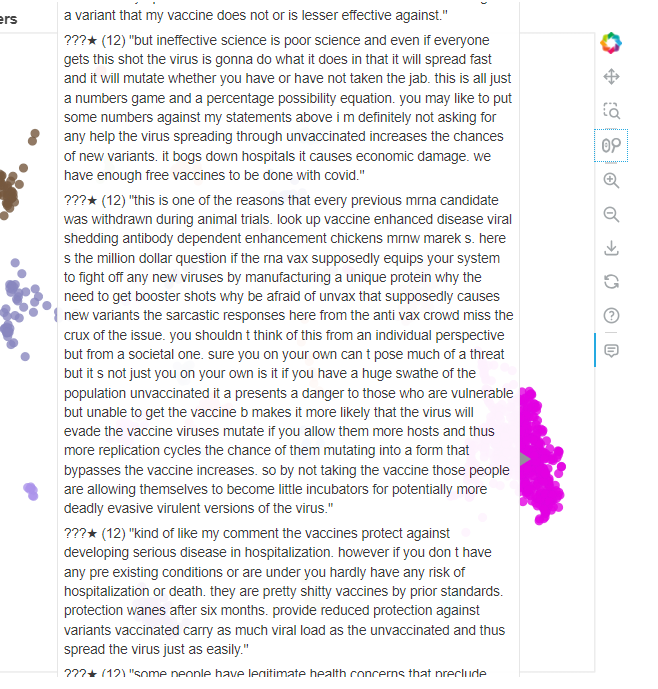
1. **Food Safety**



1. **Anti – Maskers (COVID 19)**



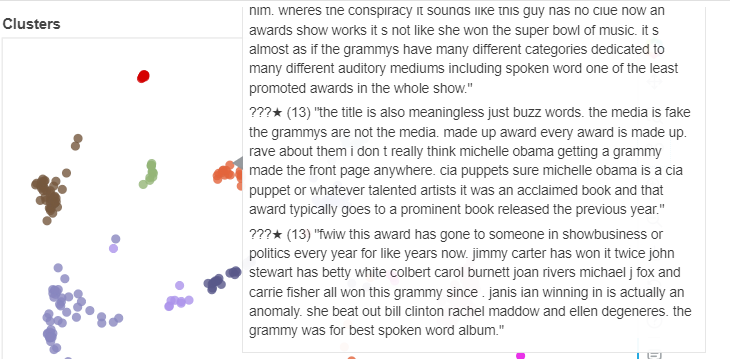
1. **COVID 19 – Vaccines**



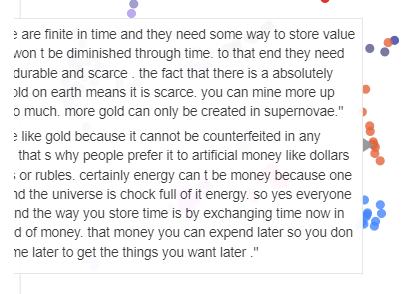
1. **Black Lives Matter**



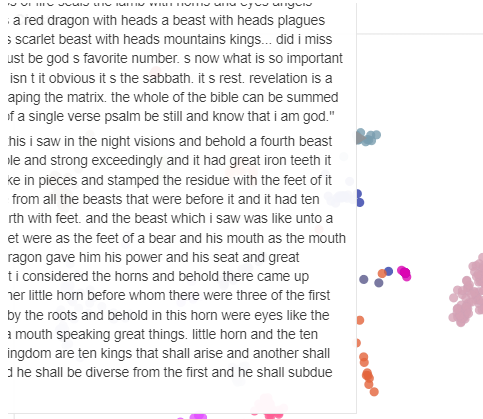
1. **Grammys**



1. **Gold**



1. **Religion – Unclear**



Relations:

The cluster projection plots show the following patterns:

* Voter Fraud and Trump Controversaries share a proximity.
* Kyle Rittenhouse and Zionism share a very close proximity
* Flat Earth and Religion(Unclear) cluster are very closely linked
* Covid 19- Vaccine is very closely linked with Anti-maskers (Covid 19)

and Food Safety

* Kyle Rittenhouse is closely related to Black Lives Matter