Unsupervised Learning of News Articles using a Custom Topic Modelling Method

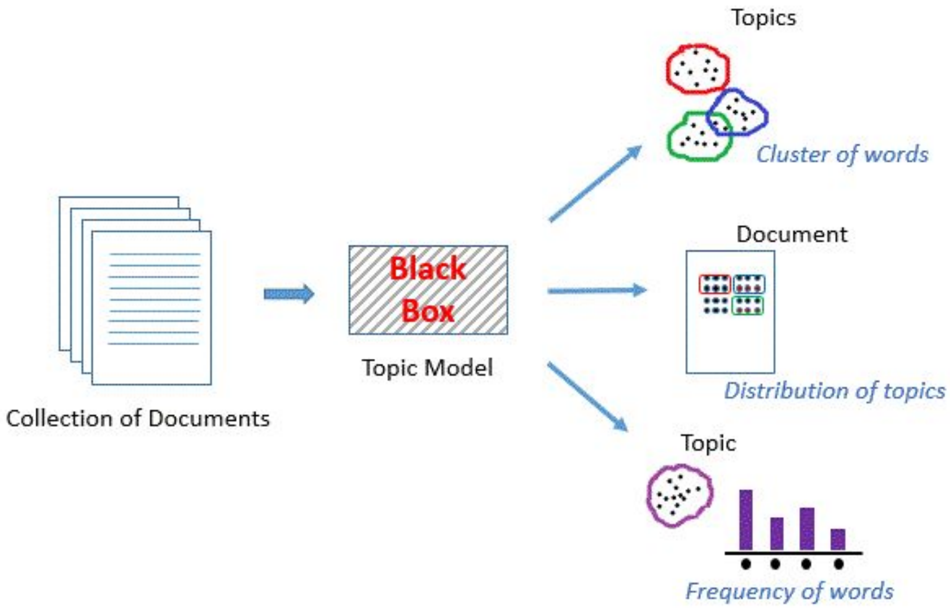
BrainStation Capstone Project
By: Leanna Lo

Problem Statement

Using **Unsupervised Natural Language Processing**, how might I conduct **topic modelling** on an article database from a news network to optimize current classification?

Background

What is Topic Modelling?



Source: https://devopedia.org/topic-modelling

Context

In the age where many read the news digitally, **topic modelling** can help a business:

- Inform website hierarchy, increasing user experience for site visitors
- Determine the most relevant tags to use, improving SEO

Project Workflow

Data

Dara Pre-processing

- EDA
- Cleaning/Wrangling: Data type conversion,, Text cleaning (special character + stopword removal, Lemmatization)

Modelling

Word Vectorization and Clustering

- Word2vec on Article text
- HDBSCAN clustering and visualization

Topic Labelling

Cluster Identification

- Bag of Words on each cluster, Cluster topic identification using top 30 words

Evaluation

Topic/Cluster Evaluation

- Cosine similarity score, using article 'Section' as a reference
- Visual scan

Data Collection

- Dataset from Kaggle
- Articles from CNN website from 2011-2022:
- 38,000 rows, 11 columns:
    - Author
    - Date Published
    - Category
    - Section
    - Headline, Description, Keywords, Article Text

Results

HDBSCAN Cluster Results

- 60 valid clusters (excluding noise cluster)
  - 28.2% of the data in the noise cluster
  - Second largest cluster slightly smaller at 27%
- Validity score after running pipeline= 37.8%

Topic Labelling

Cluster Identification

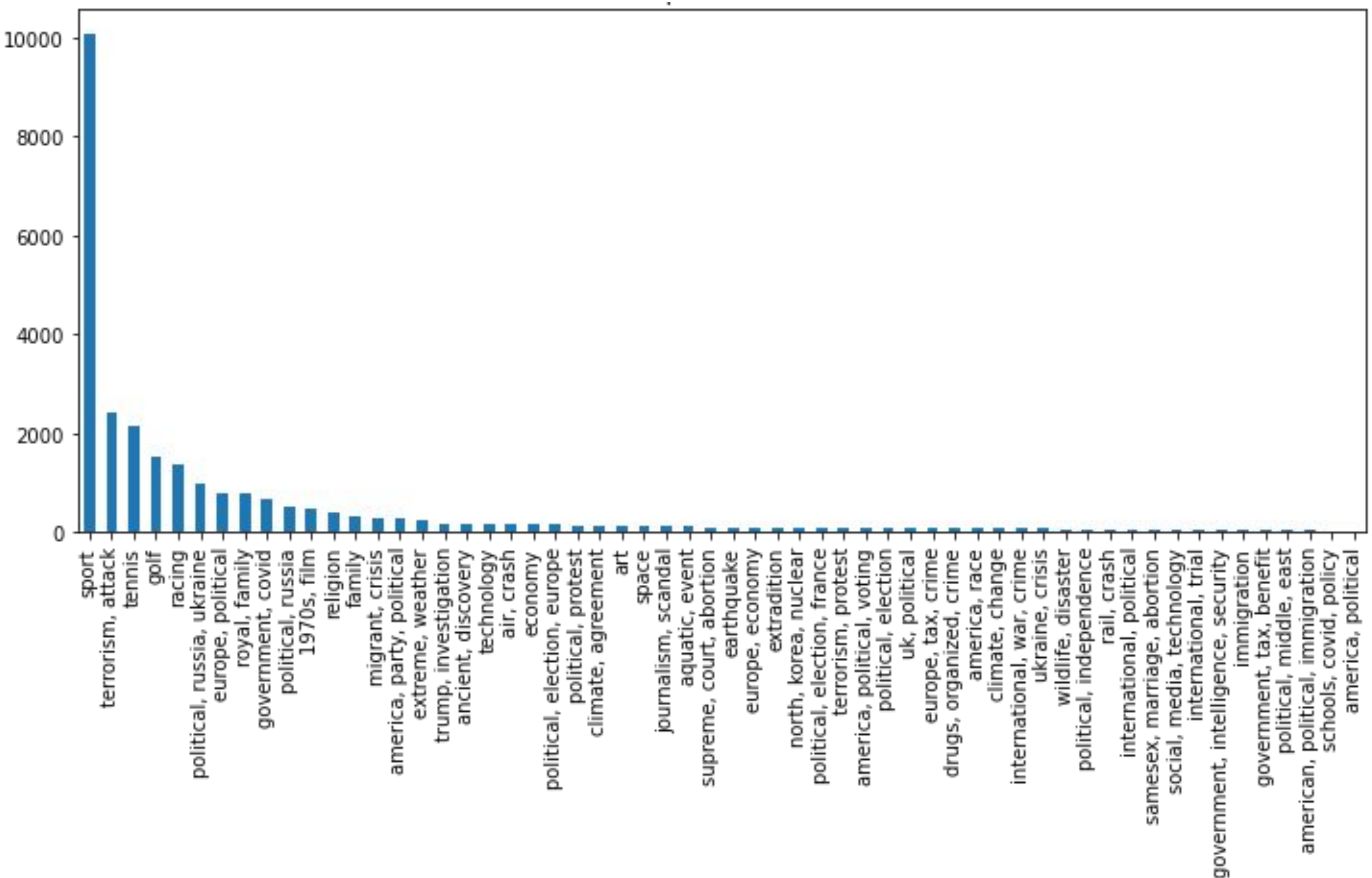
Cluster Identification - Labelling

cluster	top30words
-1	[best, world, new, state, president, country, 2012, woman, day, coronavirus, government, like, trump, could, russia, according, right, many, police, family, ukraine, get, life, match, video, child, home, group, may, city]
0	[ukraine, crisis, 168, ukrainian, march, prorussian, may, april, russian, building, slovyansk, donetsk, 132, guard, police, military, near, stand, soldier, crimea, kiev, outside, activist, armed, force, front, regional, government, protester, troop]
1	[golf, open, best, wood, round, shot, master, ryder, major, hole, pga, cup, world, win, tour, tiger, championship, course, player, back, 2012, day, play, second, mcilroy, tournament, video, must, british, final]
2	[open, tennis, match, slam, grand, set, final, win, world, williams, djokovic, player, title, federer, nadal, wimbledon, french, australian, must, video, court, champion, play, tournament, murray, second, serena, game, back, round]
3	[news, hacking, murdoch, phone, police, world, british, newspaper, former, brook, inquiry, editor, corp, must, international, video, scandal, medium, journalist, tabloid, public, investigation, minister, sun, cameron, voice, rupert, charge, uk, mail]
4	[pope, francis, vatican, church, catholic, abuse, cardinal, benedict, priest, new, bishop, sexual, child, must, world, st, report, john, video, peter, rome, visit, 44, day, victim, xvi, holy, mass, papal, meeting]
5	[race, f1, driver, team, formula, car, hamilton, grand, season, world, prix, champion, vettel, win, title, second, championship, new, racing, sport, bull, must, red, mercedes, video, ferrari, lewis, point, back, track]
6	[world, team, game, player, league, football, club, cup, sport, win, season, champion, goal, new, day, match, final, fan, like, second, video, best, must, city, back, 10, woman, olympic, home, play]

Cluster 0 : ‘ukraine, crisis’
Cluster 1: ‘golf’
Cluster 2: ‘tennis’
Cluster 3: ‘journalism, scandal’
Cluster 4: ’religion’
Cluster 5: ‘racing’
Cluster 6: ‘sport’

….

Histogram of Topic Clusters Determined by My Model



Evaluation

Topic/Cluster Evaluation

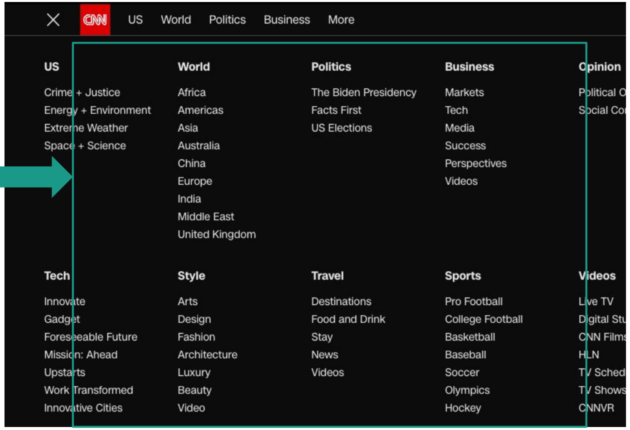
Cluster Performance -
Cosine Similarity Score of My Label vs CNN Section Label

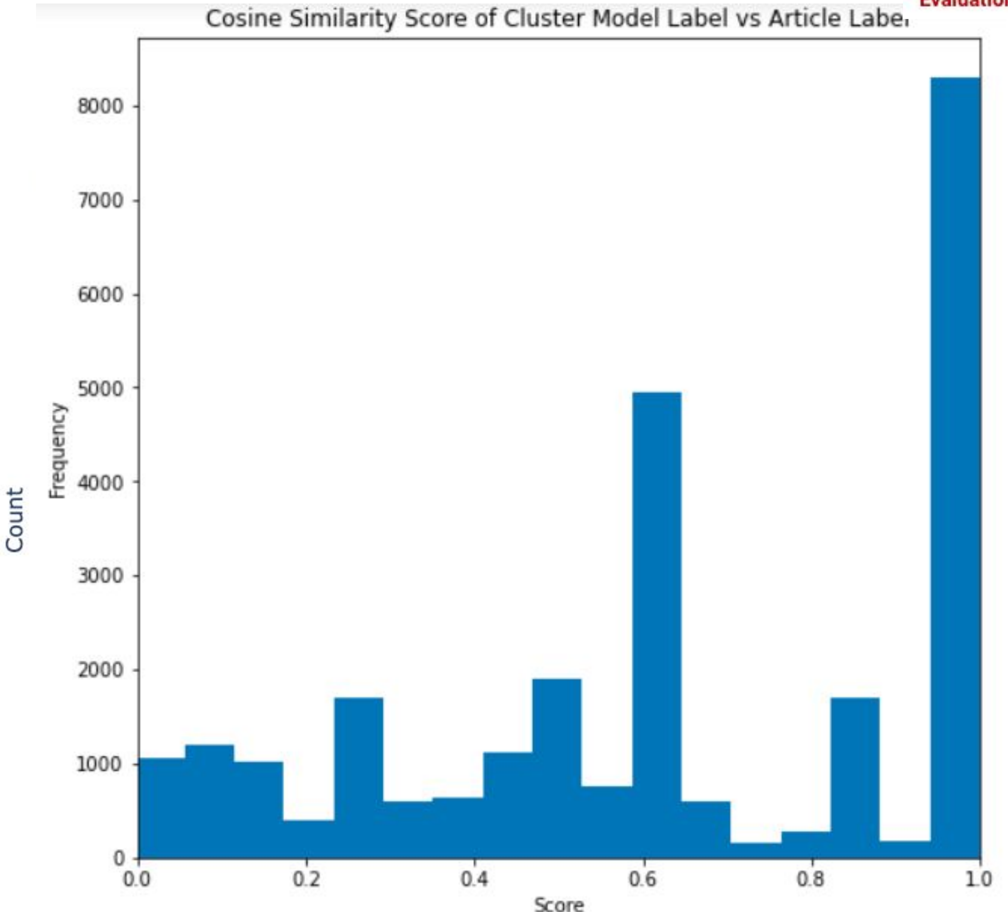
Evaluation

Topic/Cluster Evaluation

- 100% similarity score for 30.9% of the documents!
- Average score was 62.7%
- Over ⅔ of the articles had a score higher than 50%
- 50th percentile at a score of 64%, 25th percentile at a s of 40% or below

CNN section label





Cluster Model / Topic Performance - Visual Scan

Most cluster labels captured the topic quite well and in many cases had more granularity than Section
This also revealed some weaknesses in the Word2vec vectors used to determine similarity score

Evaluation

Topic/Cluster Evaluation

Section label | My model label

	Year published	Month_year published	Category	Section	Article text	cluster ID	cluster category	cosine_similarity_score_rounded
33169	2020	2020-08	sport	sport	(CNN)Serena Williams came back from the brink...	2	tennis	0.50
29491	2019	2019-06	news	europe	Moscow (CNN)Russian President Vladimir Putin h...	33	political, russia	0.50
33224	2020	2020-08	news	australia	(CNN)A light aircraft overloaded with cocaine...	41	drugs, organized, crime	0.17
25992	2018	2018-05	news	europe	Rome (CNN)A victim of clerical sexual abuse ha...	4	religion	0.15
5513	2019	2019-08	news	world	(CNN)After months of record temperatures, sci...	28	extreme, weather	0.13
27619	2018	2018-11	news	uk	(CNN)Nervous fliers, stop reading now.A Japan...	23	air, crash	0.12
34491	2021	2021-02	news	europe	(CNN)Prince Philip has spent a second night i...	7	royal, family	0.11
5529	2016	2016-01	news	us	(CNN)The Rev. Martin Luther King Jr. was a Re...	55	america, race	0.11

Cluster Model / Topic Performance - Visual Scan

Other cluster labels (a smaller amount) did not seem the most accurate

Evaluation

Topic/Cluster Evaluation

Section label

My model label

	Year published	Month_year published	Category	Section	Article text	cluster ID	cluster category	cosine_similarity_score_rounded
10188	2016	2016-08	politics	politics	(CNN)Filmmaker Spike Lee said Monday Donald T...	6	sport	0.26
26600	2018	2018-07	news	europe	Rome (CNN)George Clooney has been released fro...	48	terrorism, attack	0.29
30809	2019	2019-10	sport	sport	(CNN)The New Orleans Saints got some unexpect...	4	religion	0.20

Next Steps

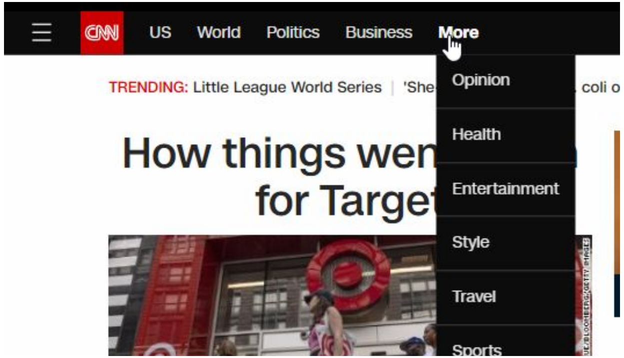
Model Improvement

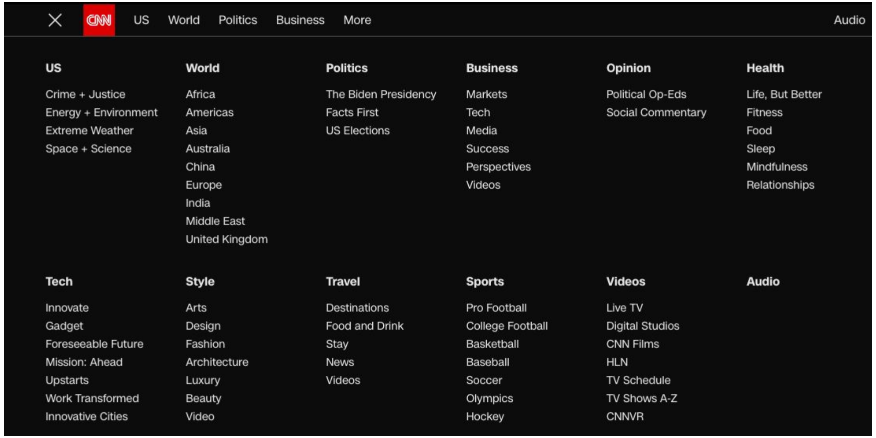
- **Word model:**
  There may be advantages in exploring another word model, such as BERT (which takes context into account)

- **Clustering model:**
  Find ideal balance between a high validity score for HDBSCAN model, smaller noise cluster, and cluster number

Practical Applications

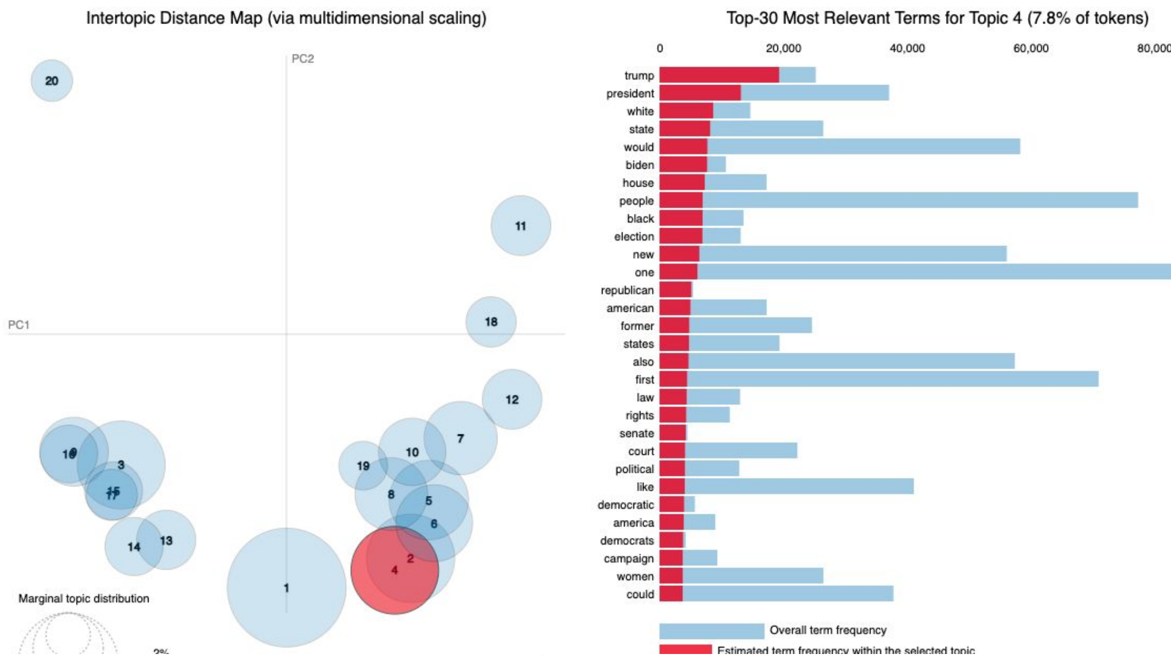
- Topics can be added as **tags** to current articles; CNN appears to not use meta tags, which will improve SEO
- Topics that were more granular than the current CNN section label can be added to the site **online navigation**





Future Work

Carry out other common topic modelling methods (i.e. LDA) and compare its performance/findings to my model



The end.
Thank you!