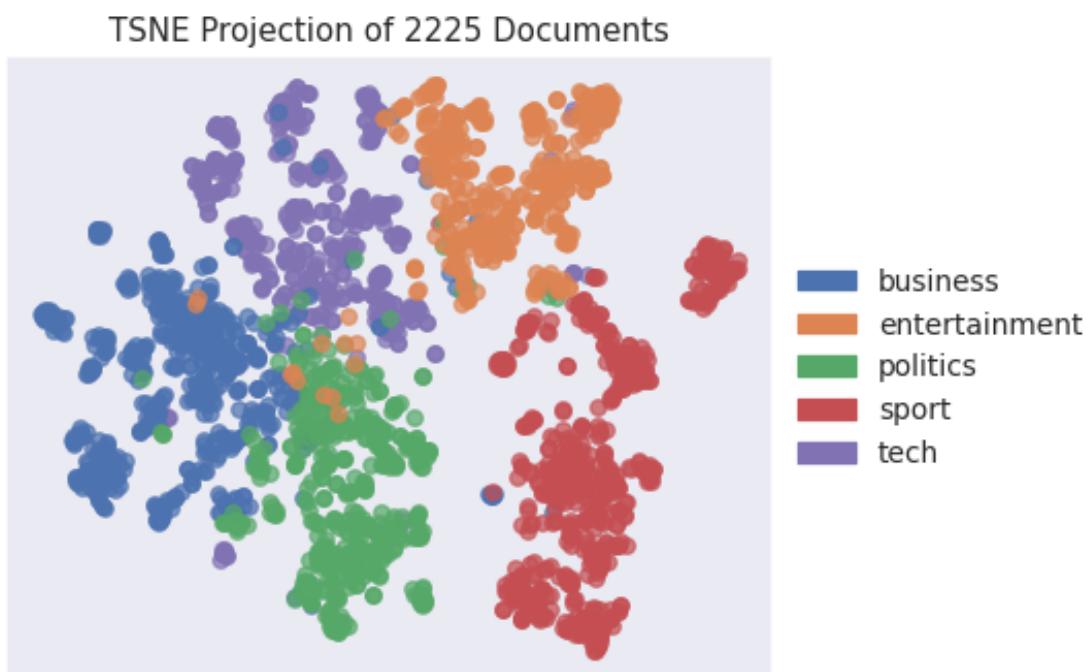


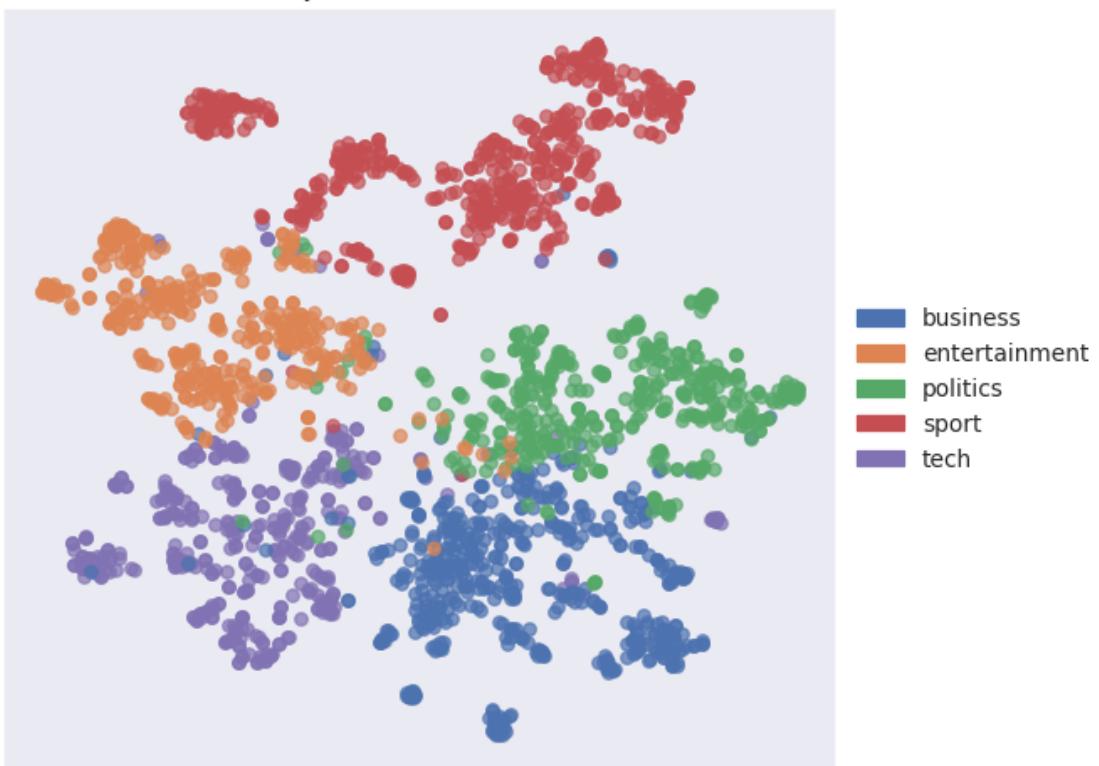
bbc

July 6, 2022

## 1 Article Similarity Distribution Visualization via t-SNE



TSNE Projection of 2225 Documents

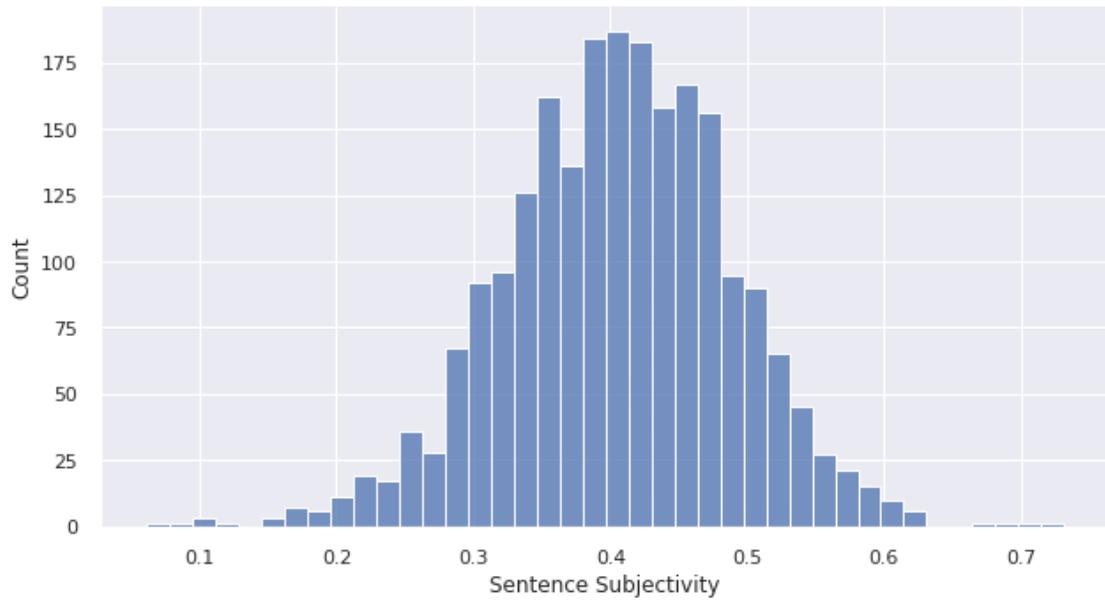
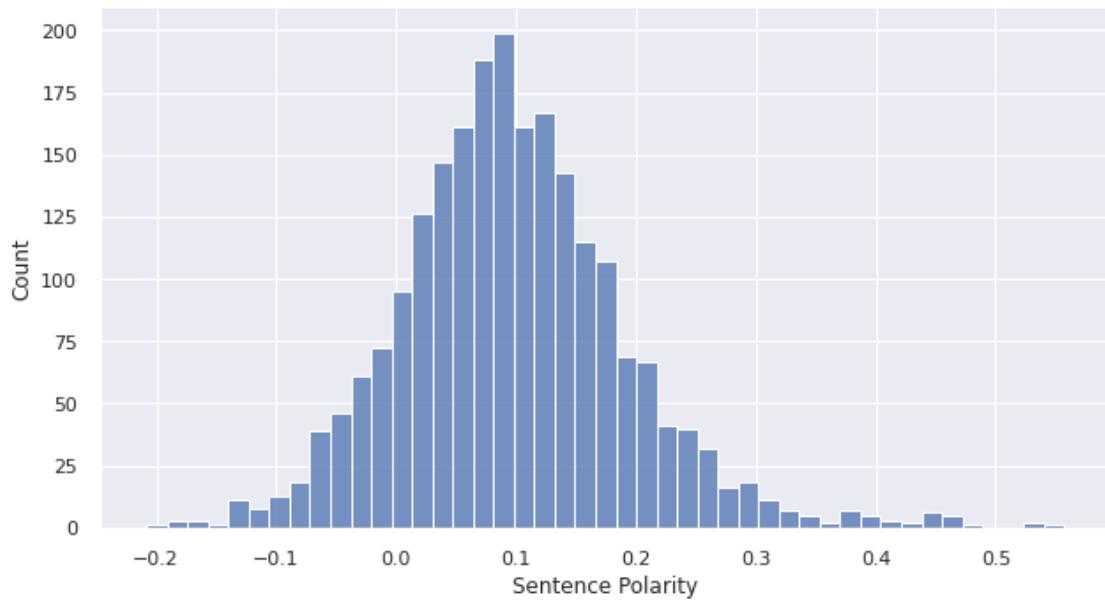


## 2 Word Cloud by Each Article Topic

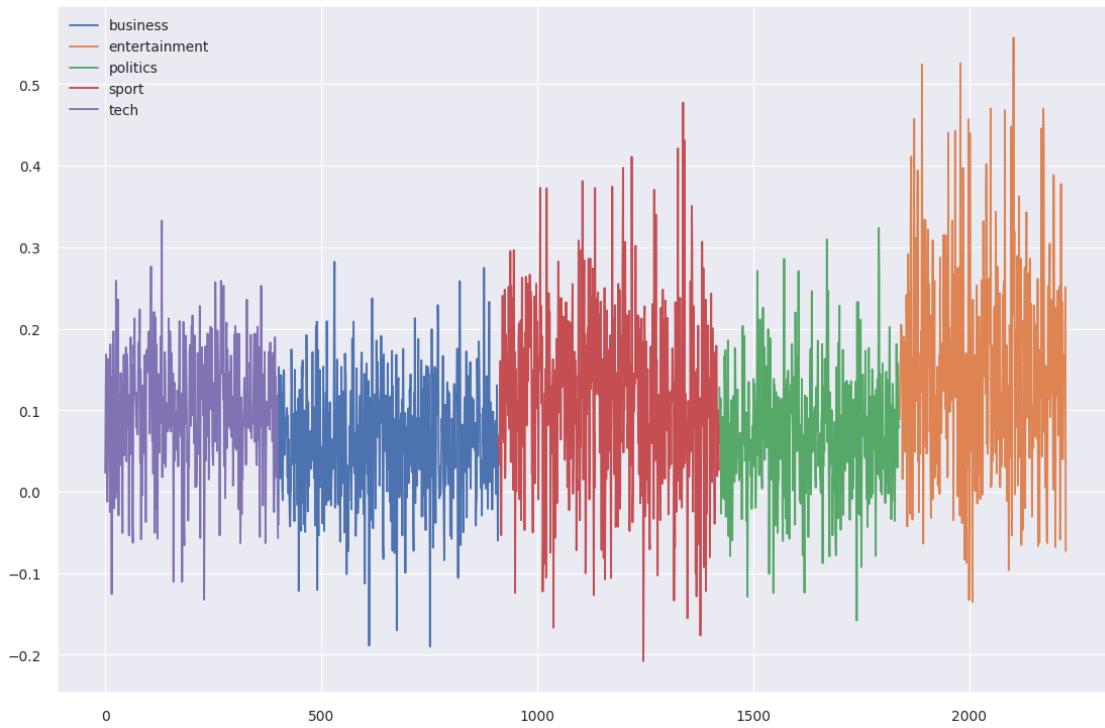


### 3 Phrase Cloud for Each Article Topic

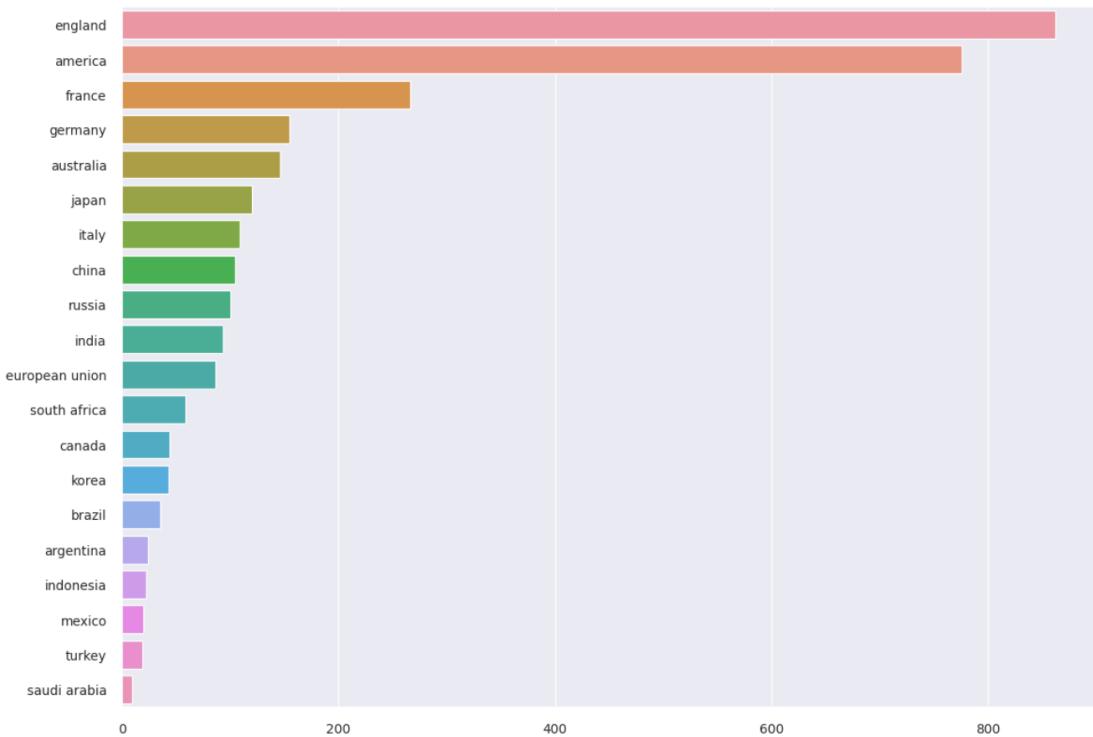
## 4 Overall Article Polarity and Subjectivity



## Polarity by Topic

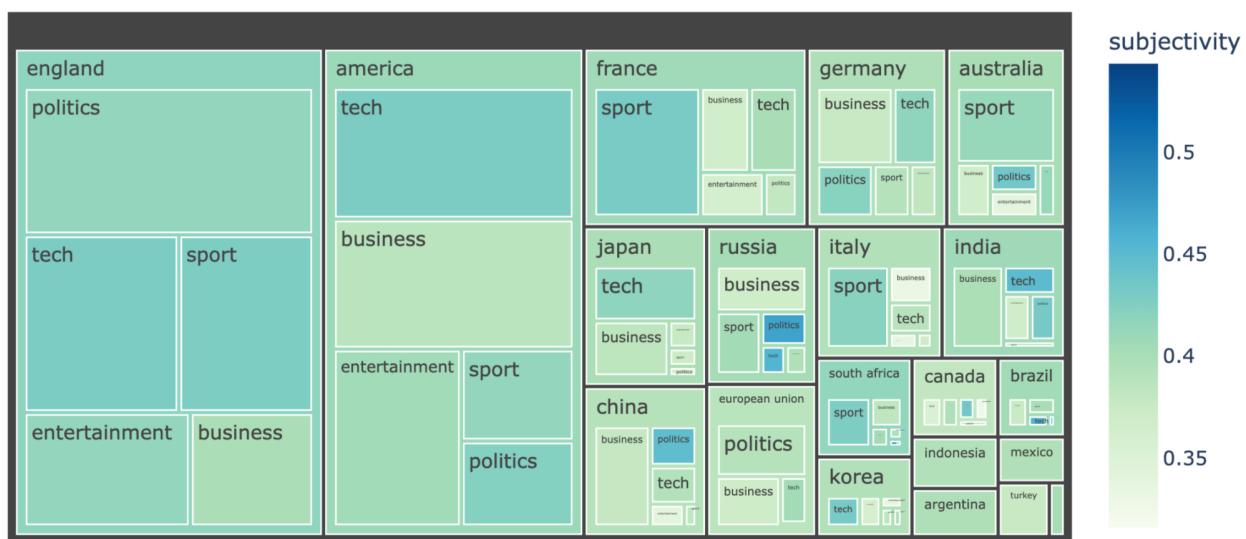
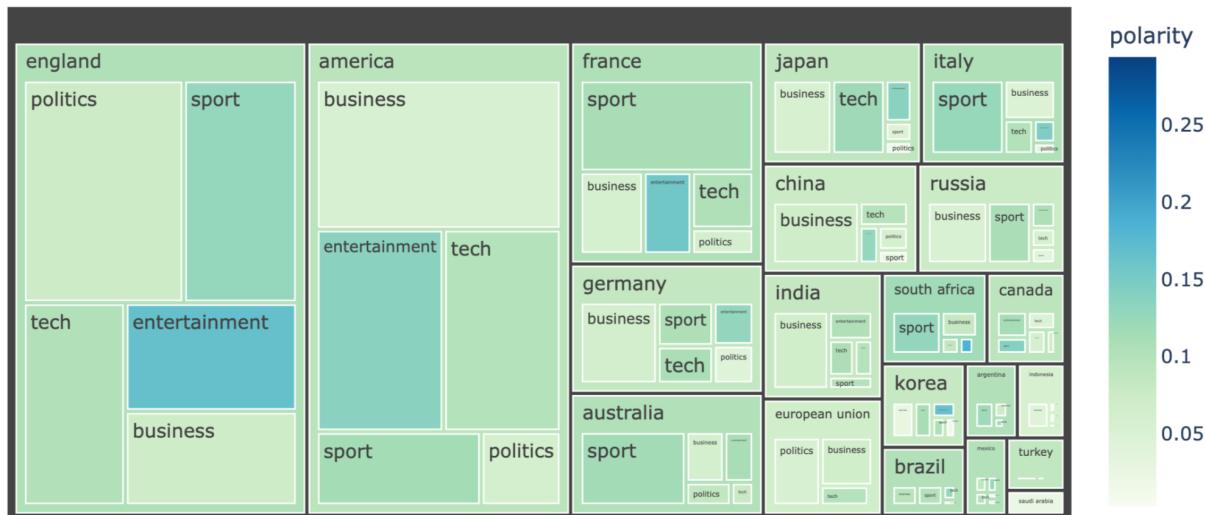


## 5 Most Mentioned G20 Countries

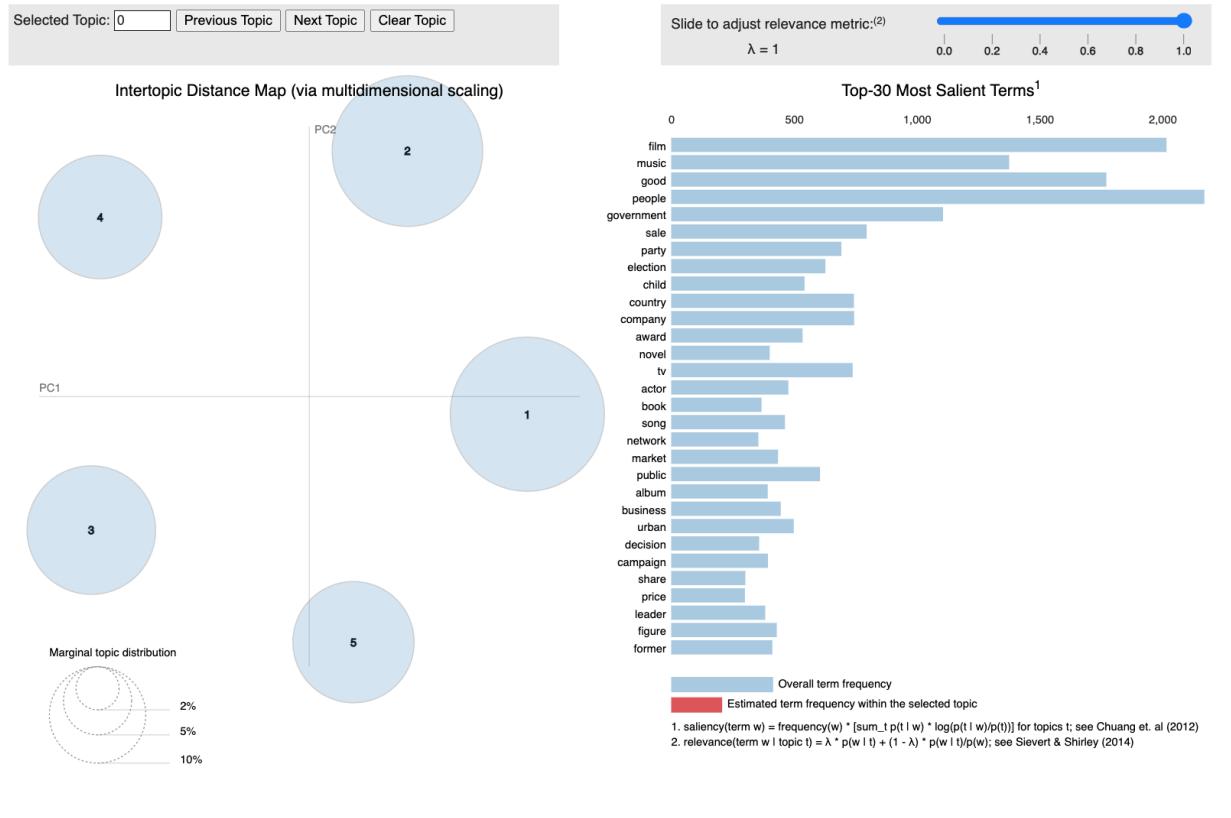


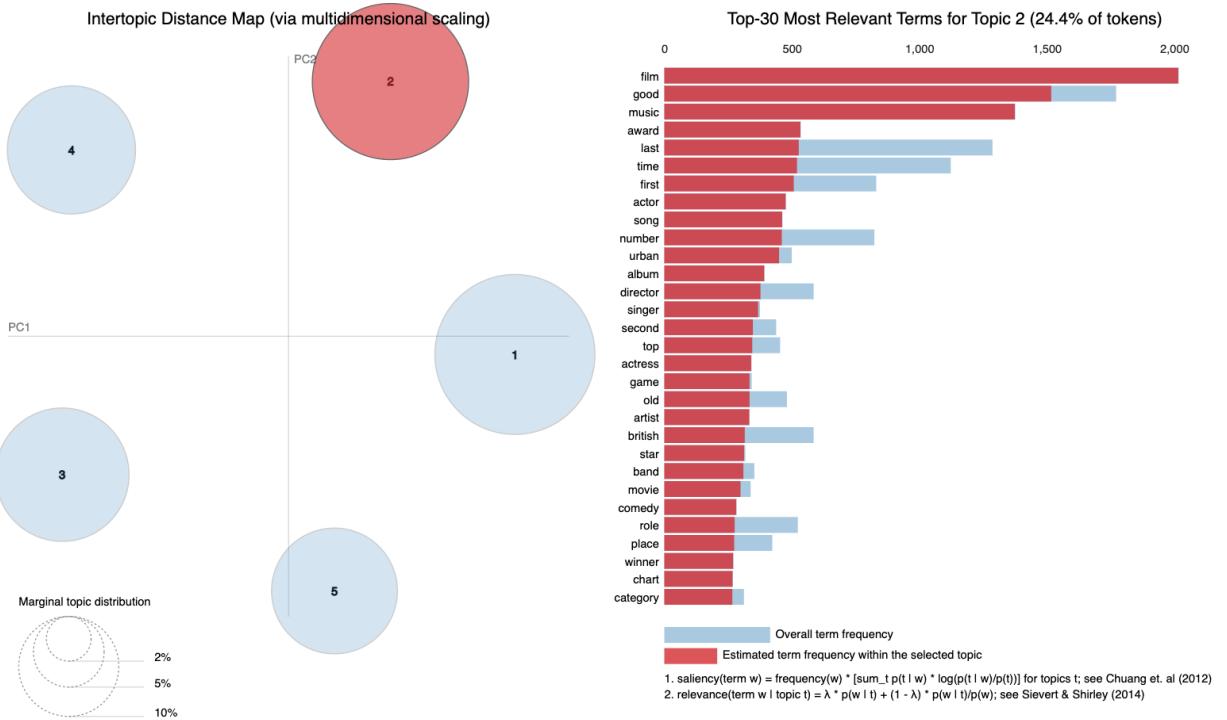
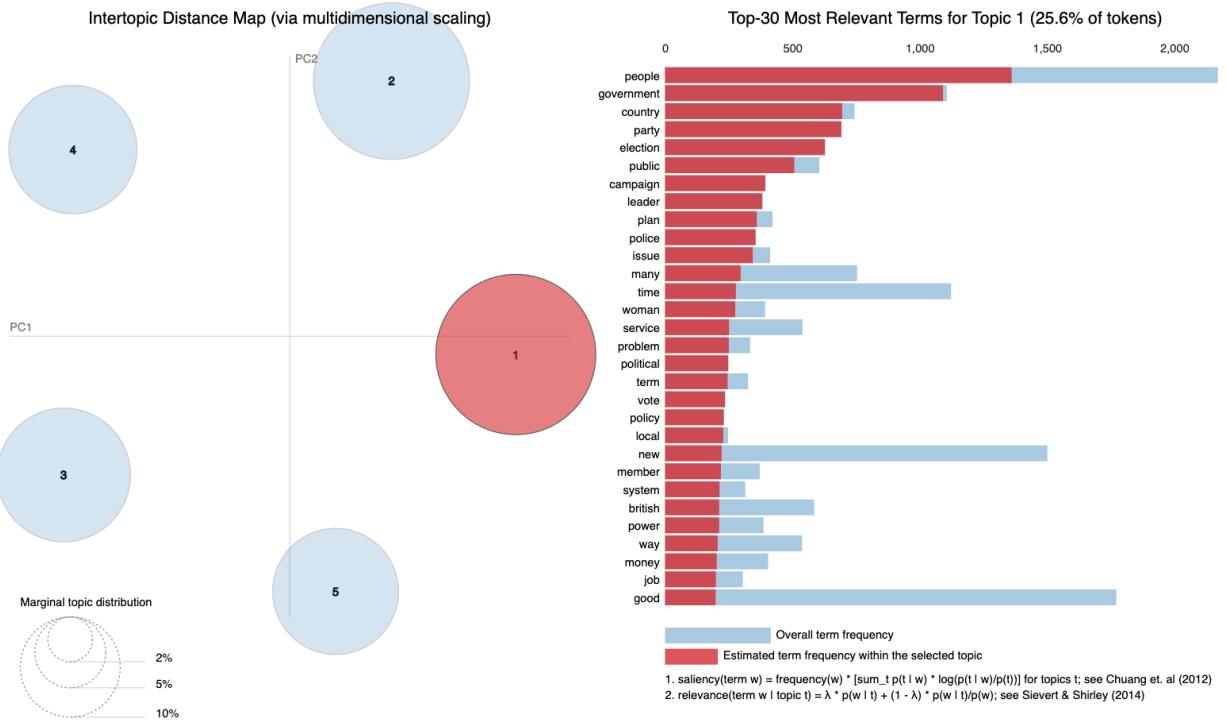
topic		text	polarity	subjectivity	word #		sentences	sent #	title	matches	matches2
0	tech	Game makers get Xbox 2 sneak peek Microsoft h...	0.023355	0.477498	463	[game makers get xbox 2 sneak peek microsoft ...	19	Game makers get Xbox 2 sneak peek	[america]		[US]
1	tech	Man auctions ad space on forehead A 20-year-o...	0.110556	0.414444	267	[man auctions ad space on forehead a 20-year...	13	Man auctions ad space on forehead	[america]		[US, Omaha, Nebraska]
2	tech	Doors open at biggest gadget fair Thousands o...	0.168561	0.438258	702	[doors open at biggest gadget fair thousands ...	31	Doors open at biggest gadget fair	[america]		[Las Vegas, US]
3	tech	Broadband in the UK gathers pace One person i...	0.141597	0.488958	273	[broadband in the uk gathers pace one person ...	17	Broadband in the UK gathers pace	[england]		[UK]
4	tech	US woman sues over cartridges A US woman is s...	0.000831	0.425546	266	[us woman sues over cartridges a us woman is ...	12	US woman sues over cartridges	[america, england]		[US, Georgia, UK]
5	tech	Movie body hits peer-to-peer nets The movie i...	-0.012164	0.372360	337	[movie body hits peer-to-peer nets the movie ...	20	Movie body hits peer-to-peer nets	[america]		[US, Dallas, Hollywood]
6	tech	Online games play with politics After bubblin...	0.066682	0.388439	527	[online games play with politics after bubbli...	26	Online games play with politics	[america]		[US, Uruguay]
7	tech	Europe backs digital TV lifestyle How people ...	0.162908	0.414297	451	[europe backs digital tv lifestyle how people...	20	Europe backs digital TV lifestyle	[france]		[Nice]
8	tech	Consumers 'snub portable video' Consumers wan...	0.132508	0.513921	360	[consumers 'snub portable video' consumers wa...	21	Consumers 'snub portable video'	[france]		□
9	tech	China net cafe culture crackdown Chinese auth...	0.046496	0.223178	387	[china net cafe culture crackdown chinese aut...	20	China net cafe culture crackdown	[china]		[China]

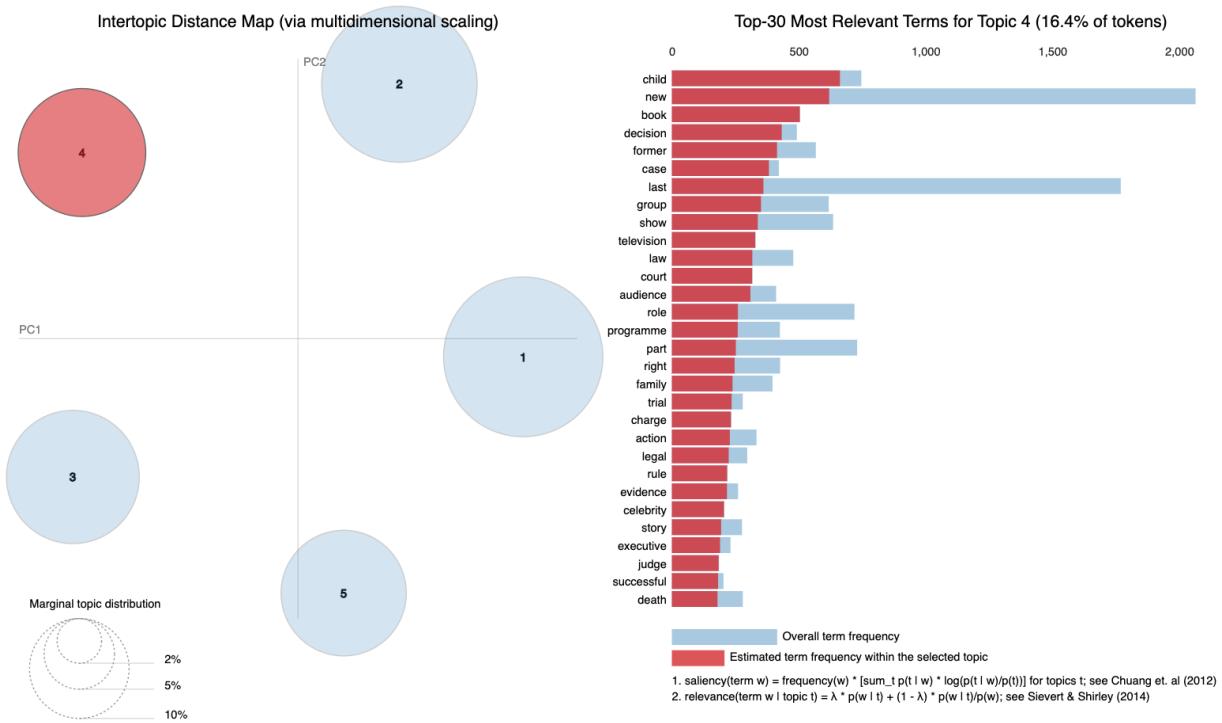
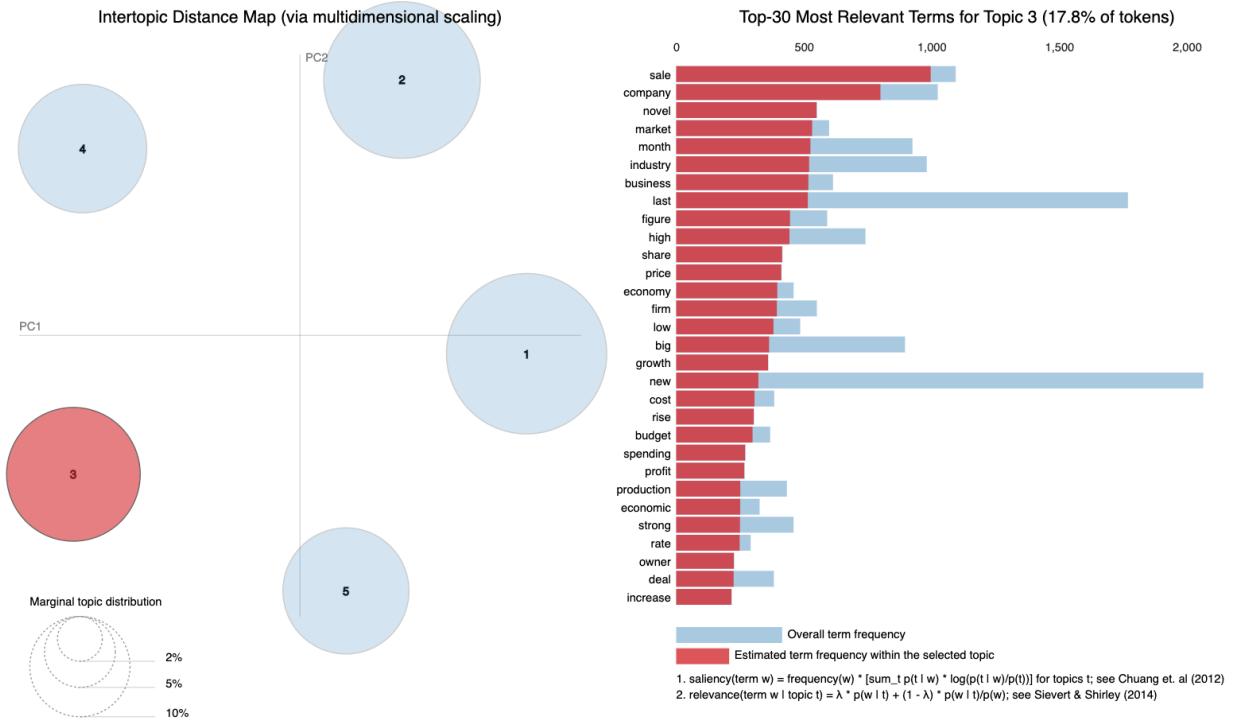
## 6 Treemap Distribution of Most Mentioned G20 Countries by Topic

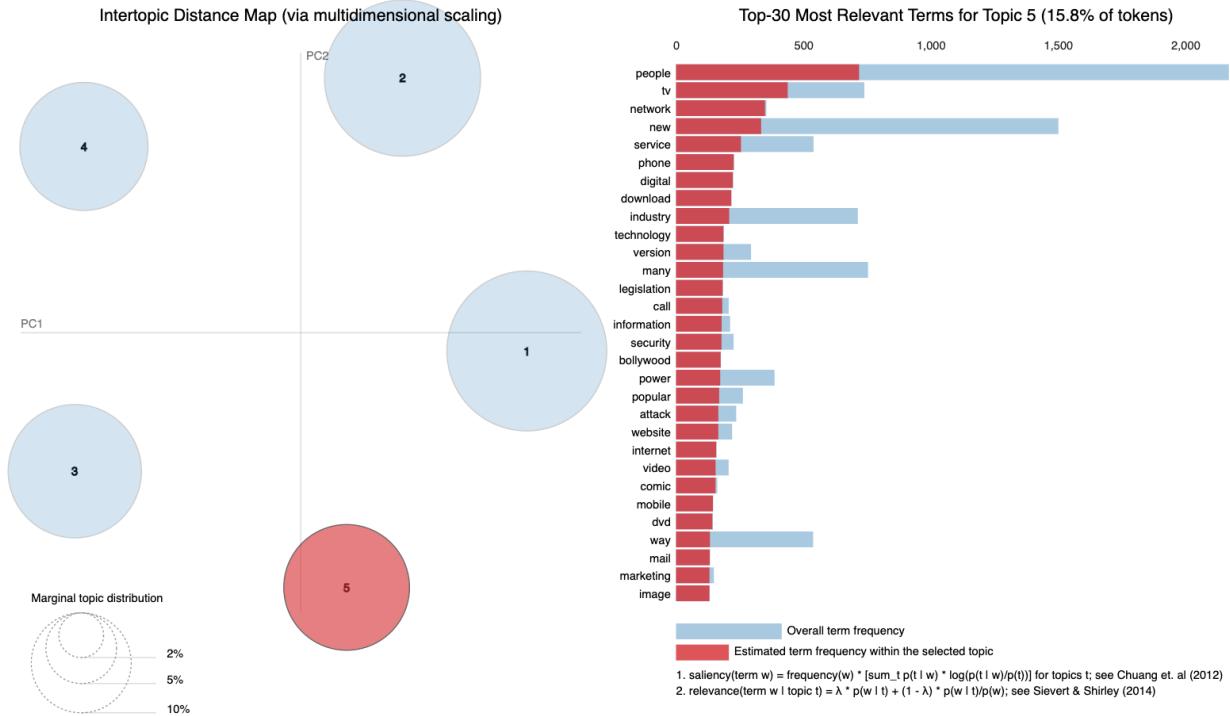


## 7 Topic Modeling (through LDA Model)



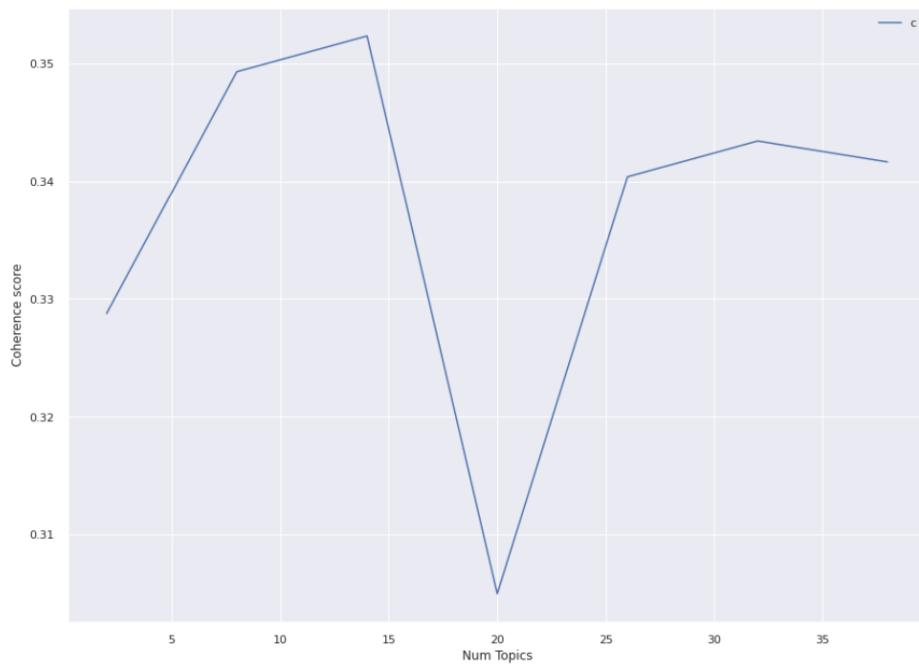






Model Perplexity: -7.6051738956872805

Model Coherence Score: 0.4001820147857291



```
Num Topics = 2 has Coherence Value of 0.3288
Num Topics = 8 has Coherence Value of 0.3493
Num Topics = 14 has Coherence Value of 0.3523
Num Topics = 20 has Coherence Value of 0.305
Num Topics = 26 has Coherence Value of 0.3404
Num Topics = 32 has Coherence Value of 0.3434
Num Topics = 38 has Coherence Value of 0.3417
- - -
```

Document_No	Dominant_Topic	Topic_Perc_Contrib		Keywords	Text
0	0	3.0	0.5426	people, tv, network, new, service, phone, digi...	[game, maker, peek, game, maker, new, console,...
1	1	1.0	0.2710	child, new, book, decision, former, case, last...	[ad, space, old, man, advertising, space, high...
2	2	3.0	0.6148	people, tv, network, new, service, phone, digi...	[door, big, gadget, fair, technology, lover, i...
3	3	3.0	0.5960	people, tv, network, new, service, phone, digi...	[pace, person, internet, second, giant, number...
4	4	3.0	0.4911	people, tv, network, new, service, phone, digi...	[woman, woman, printer, ink, cartridge, certai...
5	5	3.0	0.7232	people, tv, network, new, service, phone, digi...	[movie, body, peer_peer, net, movie, industry,...
6	6	2.0	0.3971	people, government, country, party, election, ...	[online, game, politic, time, online, game, po...
7	7	3.0	0.4003	people, tv, network, new, service, phone, digi...	[digital, tv, lifestyle, people, digital, ente...
8	8	4.0	0.3977	film, good, music, award, last, time, first, a...	[consumer, snub, portable, video, consumer, mu...
9	9	3.0	0.3831	people, tv, network, new, service, phone, digi...	[net_cafe, culture, crackdown, chinese, author...

Topic_Num	Topic_Perc_Contrib		Keywords	Text
0	0.0	0.9737	sale, company, novel, market, month, industry,...	[house_price, average, recent, evidence, housi...
1	1.0	0.9147	child, new, book, decision, former, case, last...	[greek_sprinter, greek, athletic, body, week, ...
2	2.0	0.9439	people, government, country, party, election, ...	[poll, good, result, local, council, poll, fro...
3	3.0	0.9377	people, tv, network, new, service, phone, digi...	[mail, firm, virus, electronic, card, virus, s...
4	4.0	0.9758	film, good, music, award, last, time, first, a...	[soul, memory, legend, music, world, music, ce...

## Report by Rikesh Patel on 2225 BBC Articles from 2005-2006

Questions:

1. For each of the five categories: what are the most common topics among articles?

Tech: service, games, computers, phones, world, websites, networks, work, mobile, software, device, digital, videos, internet, information, research, home, Microsoft, operating system, hard drives, credit cards, consumer electronics, Las Vegas...

Business: economy, growth, government, world, investment, Euro, Dollar, jobs, oil, China, interest rate, stock market, SEC...

Sports: games, England, championships, victories, France, goal, international, Ireland, Six Nations Rugby, World Cup...

Politics: government, Prime ministers, Labour and Tory parties, secretary, tax, laws, campaigns, policies, Leader Michael Howard, Chancellor Gordon Brown...

Entertainment: awards, shows and movies, stars, music, actors, directors, singers, Oscars, festivals, Hollywood, Los Angeles, New York..

2. Across all categories: how many articles talk about each of the G20 countries?

Surprisingly, there are more articles that mention “US” than that for “UK.” According to my count, there are 1700 articles that mention at least one G20 country, and each of the countries are mentioned in at least one article.

3. Describe the methodologies you used in your report.

To answer the question of what is the most mentioned G20 country, I thought the simplest route of retrieving the common topics of the articles was to get the topics directly from the articles’ words. So I used Counter to count the most common word sequences per article category after cleaning the text and then plotted these phrase frequencies. After thinking how this may be flawed because it risks skewing the common phrases towards articles that are repetitive, I then added an interim step to make every word unique per article, so there are no repeats. This new step solves the issue, so I will note and accept the drawback of some two-letter phrases not being counted in some rare circumstances. There is likely another route to count uniquely without taking this loss, so given more time, I would look into that. For a count of repeated word usage, I counted the number of times specific words were found in the text, and for every individual article with a match, I would tally it. I plotted these counts for a polished final result.

4. What tools did you use for this analysis?

I used Python on [Google Colab](#) to test code quickly and look at the data available, and then I used Visual Studio Code to run the Python code locally and create a PDF report.

5. If you had more time how would you strengthen your report?

If I had more time, I would integrate the geopolitical entity detection and geocoding I set up for fun with the simple country name counter to create a more accurate list of most mentioned countries out of all countries globally using the Google Maps API. I would also apply more topic models to check if they may have greater coherence. I would also like to stress test the script’s speed by adding more news articles and articles in different topics to the dataset.

## **6. Analysis**

1. The t-SNE projections were very insightful upon first looking at the dataset. It shows us how similar each group of articles by topics are by taking the similarity of the words used in each article. The disparate usage of words by article topic demonstrates that each topic has some functional classification. There is also sense in how some articles are outliers in an area of a different article topic because overlap is natural
2. The first page of word clouds represents a diagram of the most used *words* by topic, such that more highly used words are larger. At this time, I can see the biggest U.K. players in tech are Apple, Microsoft, and Sony. It is also insightful that government is a common word in both business and politics.
3. Likewise, the second page of word clouds represents a diagram of the most used *phrases* by topic.
4. Each article's words were analyzed for their polarity and subjectivity, or in other words, their positivity and bias. In this case, most BBC articles are more positive than negative and are moderately opinionated. In the last graph, one can see that sports and entertainment articles tend to be more positive on average.
5. For the most mentioned G20 countries, the United Kingdom and the United States are at the top as expected. Below this table, you can see a sample of the table that was created to make the graph above in getting matches for each country's tallies.
6. The treemap is a robust tool for visualization when used correctly. Here, the size of any box corresponds to the times mentioned while the color corresponds to the average polarity (1st graph) and average subjectivity (2nd graph) of those articles. For instance, entertainment articles that mention England are more positive (as shown in the first graph), and overall, articles that mention England are more biased (as shown in the second graph).
7. The next graphs outline the results of an LDA model on the dataset. In the instance of separating out 5 topics, the LDA model assigns a topic group to each article through machine learning until all articles are assigned. In this visualization, you can see the most used words in each artificial topic. The model coherence score is important to maximize, so the topics created in assignment have meaning, so it is equally important to choose the correct number of topics for the LDA model to create to be meaningful.