News Coverage of Inequality in NOW Corpus

1. Introduction

By: Xi Cheng

With massive data and research showing that the gap between the wealthy and poor keeps enlarging in recent years, I am interested in how online media frames this phenomenon. Extracting topics and examining inequality coverage trends may offer us a new perspective on the causes and solutions of the inequality. The goal of this project is to analyze online English language news coverage of inequality by employed computational methods including topic modeling methods, word embeddings, and word and phrase frequency analysis to all news containing "inequality" in the News on the Web (NOW) corpus to see how news coverage given to inequality has changed from 2010 to 2019.

2. Research questions

By: Xi Cheng

(1). What are the topics related to inequality news?

(2). How do these topics change from 2010 to 2019?

3. Data and methods

By: Xi Cheng

We used the News on the Web corpus as our data sources, which covers countries and areas in North American, Europe, South and southeastern Asia, Australia, and Africa, including Bangladesh, Canada, the UK, Ghana, Hong Kong, Ireland, India, Jamaica, Kenya, Sri Lanka, Malaysia, Niger, New Zealand, Philippines, Pakistan, Singapore, Tanzania, the US, South Africa between January 2010 to September 2019. The numbers of news on inequality were calculated based on the search term “inequality.” We searched articles that mention the words “inequality.”

3.1 Stratified sampling data

I sampled 70% of each year’s news articles for the following analysis to increase efficiency (sampling sizes have been rounded off). Table 1 below shows the original and sampled data sizes of each year. Then I cleaned and preprocessed the original text data into tokens.

Table 1 A summary of data sizes

Period	# news on inequality	Sample size
2010	905	634
2011	1599	1119
2012	2305	1614
2013	2908	2036
2014	4008	2806
2015	4866	3406
2016	11831	8282
2017	13048	9134

2018	12427	8699
2019 (until September)	11000	7700
Total	64,897	45,429

3.2 Latent Dirichlet allocation (LDA) topic modeling

I first used Latent Dirichlet allocation (LDA) models, a basic probabilistic topic model, to identify inequality themes in the text data. I expected to see that income inequality stands out among all media topics on inequality. The number of topics is assumed to be latent and is determined by myself. I varied the number of topics from 2 to 15 in order to find the optimum number of topics in the whole dataset. Meanwhile, I varied alpha, eta, and corpus size to find the model with the highest topic coherence score and a clear (non-overlapping) and interpretable topic visualization result. Table 2 shows a summary of my hyperparameter tuning process. In total, I looped 840 models.

Table 2 A summary of hyperparameter tuning

Corpus size	Number of topics	alpha	eta
75%	2	0.01	0.01
100%	3	0.31	0.31
	4	0.61	0.61
	5	0.91	0.91
	6	asymmetric	symmetric
	7	symmetric	
	8		
	9		
	10		
	11		
	12		
	13		
	14		
	15		
Total	840		

The LDA topic model generated a probability distribution of all topics. I classified an article as a specific topic of inequality if it contains words whose associated topic weight is greater than the threshold. The threshold is calculated by averaging the score of all words and in all topics for all articles. Table 3 shows a summary of the number of articles on each topic. The theme of each topic was identified by its top-30 most relevant terms (see Appendix A). Note that the topic order shown in this table is slightly different from the order shown in Figure 1, Figure 3, Table 4, and Appendix A since I use different packages to plot topic distribution and generate topic probabilities. However, the topics are the same. Only the corresponding number was changed.

Table 3 A summary of number of articles on each topic

Topic #	Theme	Number of articles
0	Educational inequality	11052
1	American politics	10693
2	Income/Wealth inequality	16061
3	Racial inequality	7291
4	Gender inequality	15894
5	International politics	12864
6	Class inequality	17436
7	Legal system and justice	12168
8	International development inequality	15967

3.3 Dynamic topic modeling (DTM)

I employed a 9-topic Dynamic topic model to see topic changes year by year from 2010 to 2019. Figure 1 shows an intertopic distance map (via multidimensional scaling) of the year 2020 and its top-30 most salient terms. For more results and detailed discussion, see Section 4.2.

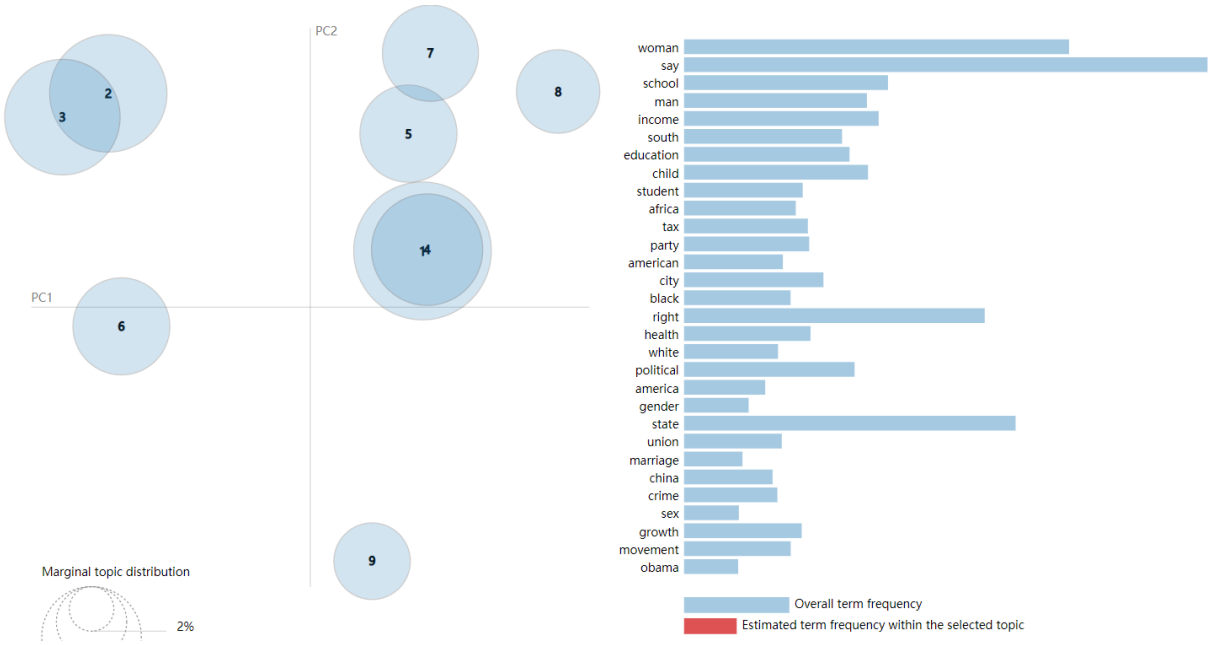


Figure 1 Intertopic distance map (via multidimensional scaling) and top-30 most salient terms of 2010

3.4 Word embeddings

By: Rui Pan

See session 4 below.

3.5 Words and phrases frequency analysis

By: Rui Pan

See session 4 below.

4. Results and Discussion

4.1 LDA topic modeling

By: Xi Cheng

Figure 2 shows the topic distribution of 9 latent topics from the LDA model. Table 4 shows the theme of each corresponding topic ordering by its percentage of tokens. The top 1 topic accounts for 15.9 % of all the tokens, including relevant terms such as class, political, social, human, etc. The top 2 topic accounts for 14.7% of all the tokens, including terms like tax, income, pay, wage, wealth and so forth. The results show that class inequality stands out as the most popular topic among all media topics on inequality. Figure 3 shows the probability distribution of top-25 words among topics.

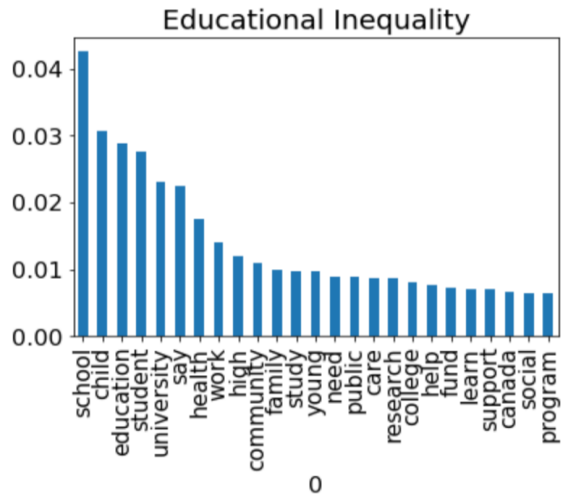
It is different from my hypothesis, which assumes that income inequality will stand out. However, income inequality is positioned on the top 2, which implies that it is still a crucial and well-concerned topic. For more information about percentages of tokens and relevant terms of each topic, see Appendix A.

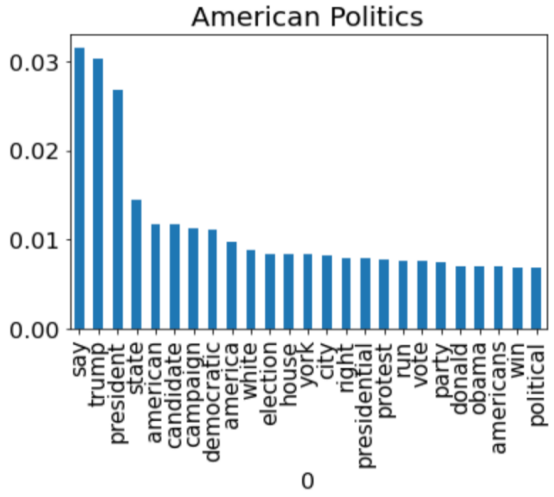
Figure 2 LDA Intertopic Distance Map (via multidimensional scaling)

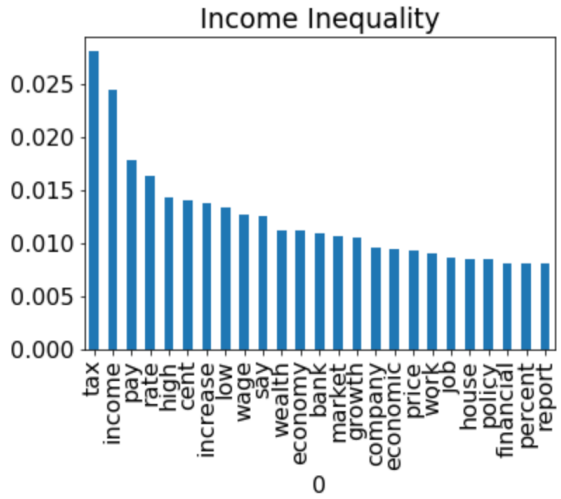
Table 4 Theme of each topic

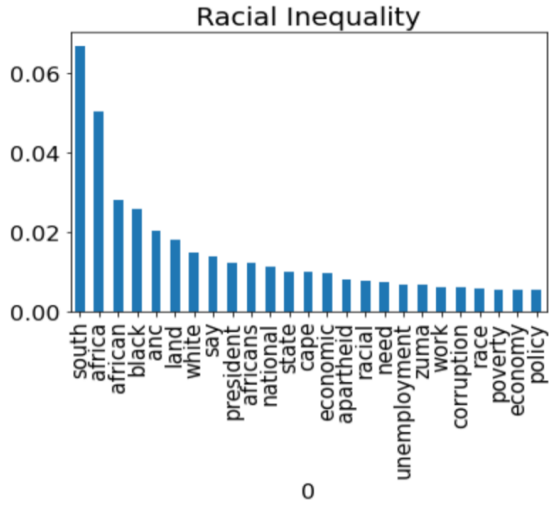
Topic #	Theme

1	Class inequality
2	Income inequality
3	Gender inequality
4	International development inequality
5	International politics
6	Educational inequality
7	American politics
8	Legal system and justice
9	Racial inequality









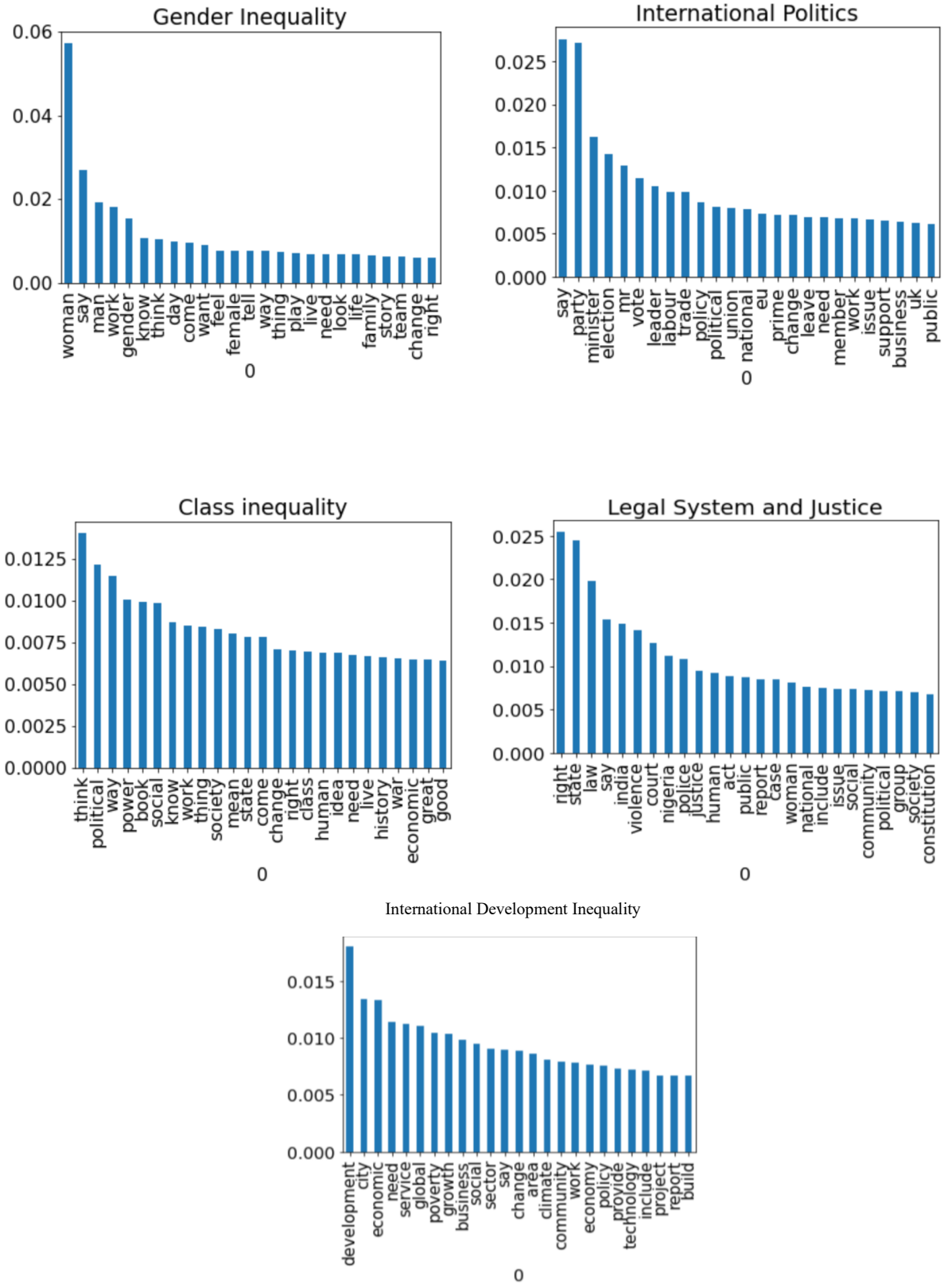


Figure 3 Probability distribution of words among topics

I randomly chose two articles from each year's news and drew the heatmap of topic probability distribution shows in Figure 4. The related topic themes are shown in Table 3 (not Table 4). We can see that topic 2, income inequality, has a high probability for these ten articles. The two articles of 2010 both have a relatively high probability on topic 5, which is about international politics. Indeed, two samples of each year cannot represent the whole characteristics, here I want to address that drawing heatmap is an excellent way to visualize the topic probabilities, and in the future, I want to improve the article topic classifying algorithm by using the maximum and minimum topic probabilities rather than current average topic scores.

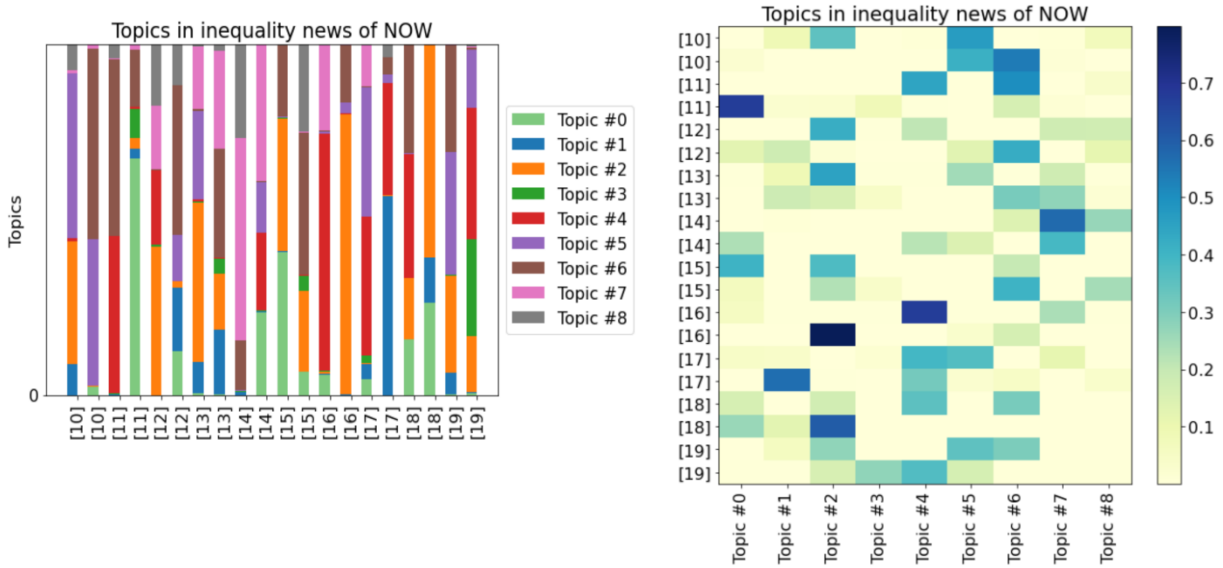


Figure 4 Topic distribution of two articles per year

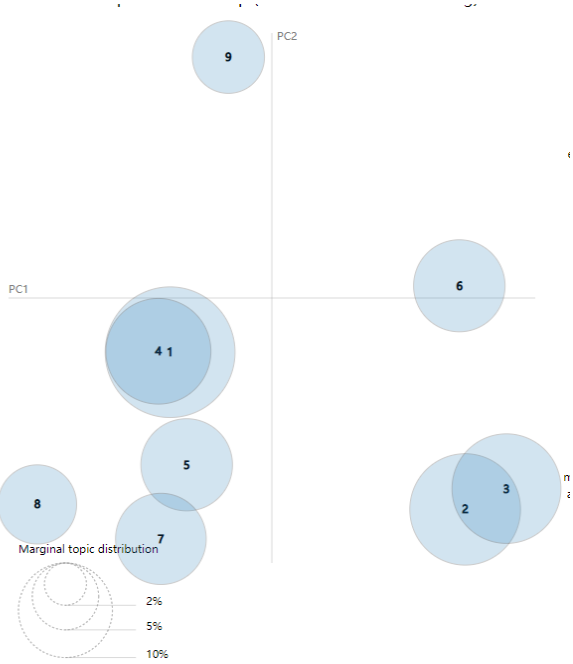
4.2 Dynamic topic modeling (DTM)

By: Xi Cheng

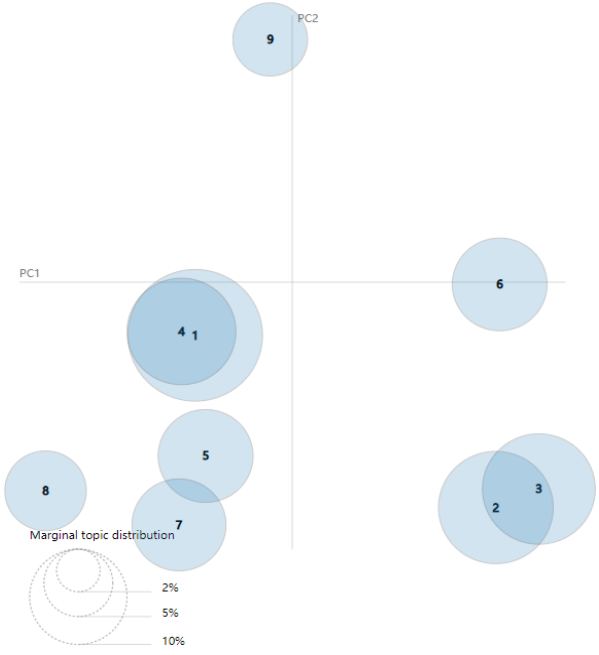
Figure 5 shows the changing trend of topic distribution from 2010 to 2019. Figure 6 shows the 20 most probable words of each year from 2010 to 2019 for each topic. I noticed that the distance between topic 4 and topic 5, 7 keeps increasing, and topics 5 and 7 are also becoming farther in Figure 5. According to the most probable words in Figure 6, topic 4 contained words such as city, house, street, community, etc., and topic 5 includes words such as Africa, crime, court, justice, community, etc., and topic 7 contains words such as party, political, union, vote, etc.

I think the increasing distance between topic 4 and topic 5/7 implies a continuous detachment in online media discussion between local inequality issues (topic 4) and national or global political topics (topic 5 and 7).

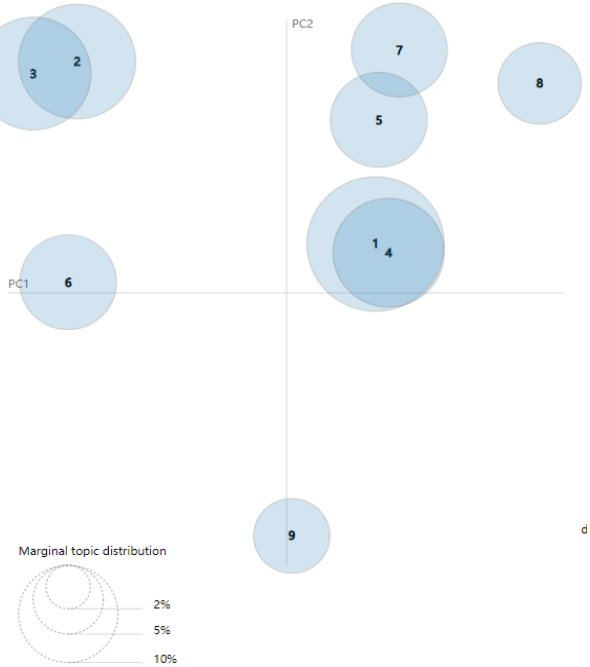
2010



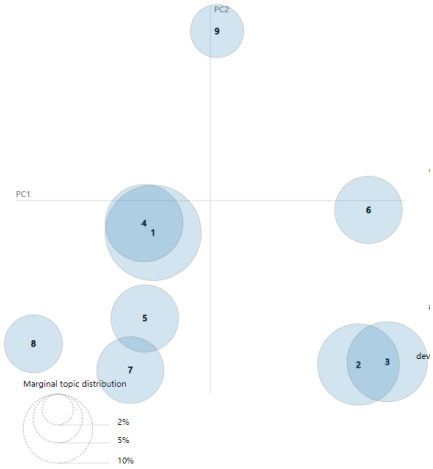
2011



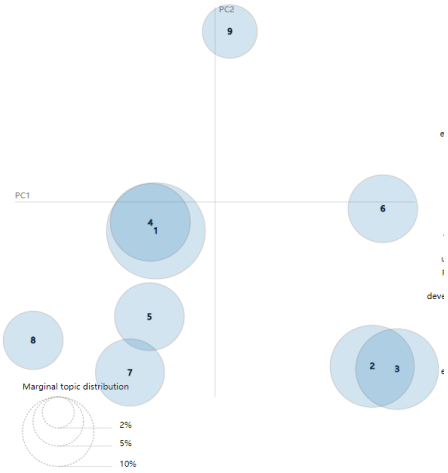
2012



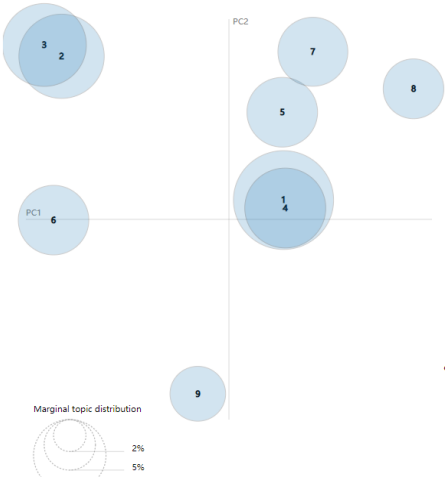
2013



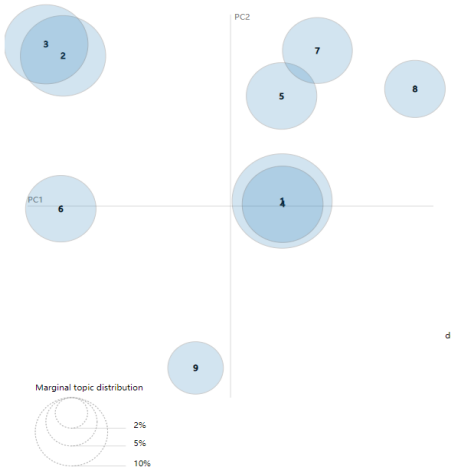
2014



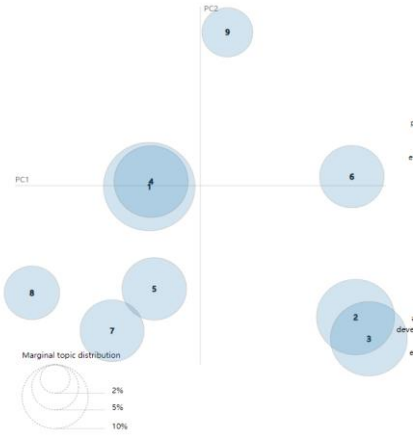
2015



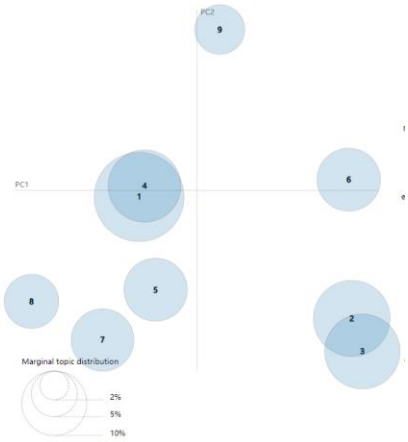
2016



2017



2018



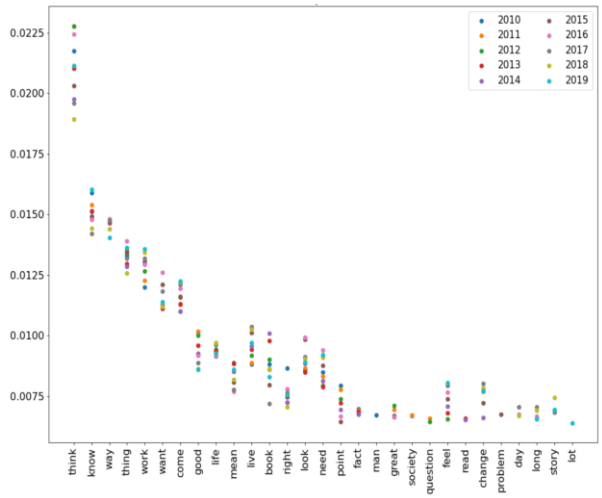
2019

Figure 5 DTM Intertopic Distance Maps from 2010 to 2019 (via multidimensional scaling)

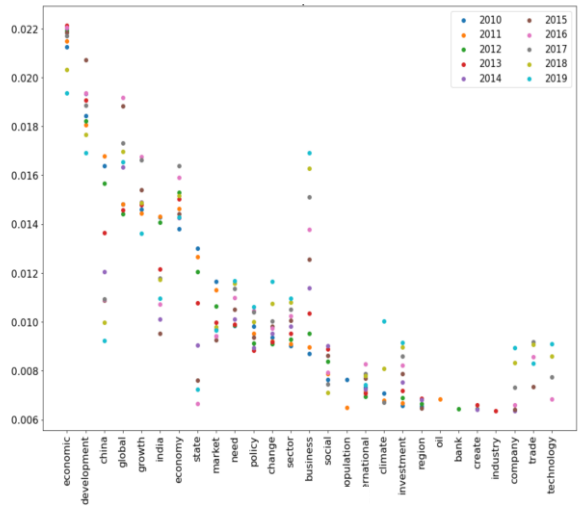
Furthermore, in comparing the total word numbers and their probabilities, I found that some topics experienced more enormous changes from 2010 to 2019 than others. For example, we can see that words on the x-axis of topic 2 (“economic,” “development,” “China,” “global”), topic 5 (“south,” “Africa,” “right,” “crime”), topic 7 (“party,” “political,” “union,” “vote”), topic 8 (“America,” “black,” “white,” “Obama”) are more intensive while those on the on-axis of topic 4 (“city,” “house,” “street,” “community”), topic 6 (“school,” “education,” “child”), topic 9 (“women,” “men”). Topic 2, 5, 7, 8's word probabilities also varied more, while topic 4, 6, 9's word probabilities are relatively stable.

These facts imply that gender inequality, local housing inequality, and educational inequality remain stable over the ten years. However, international and national political topics had a dramatic change in the same period.

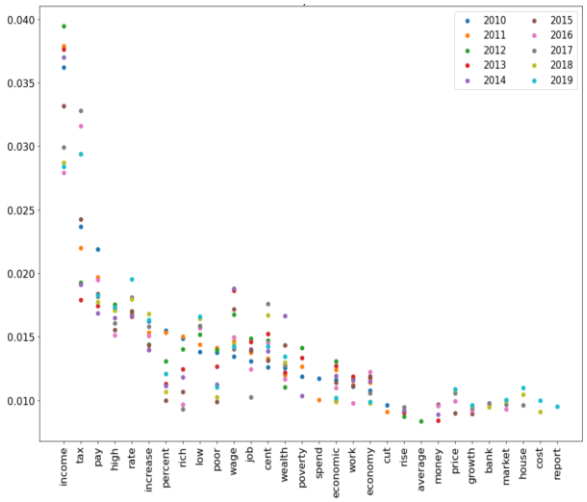
Topic 1



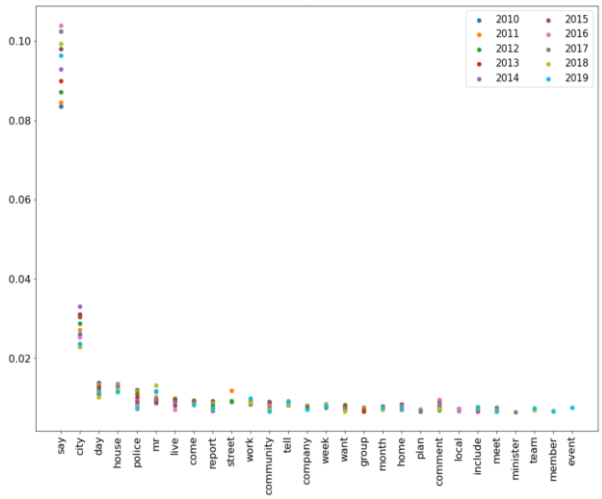
Topic 2



Topic 3



Topic 4



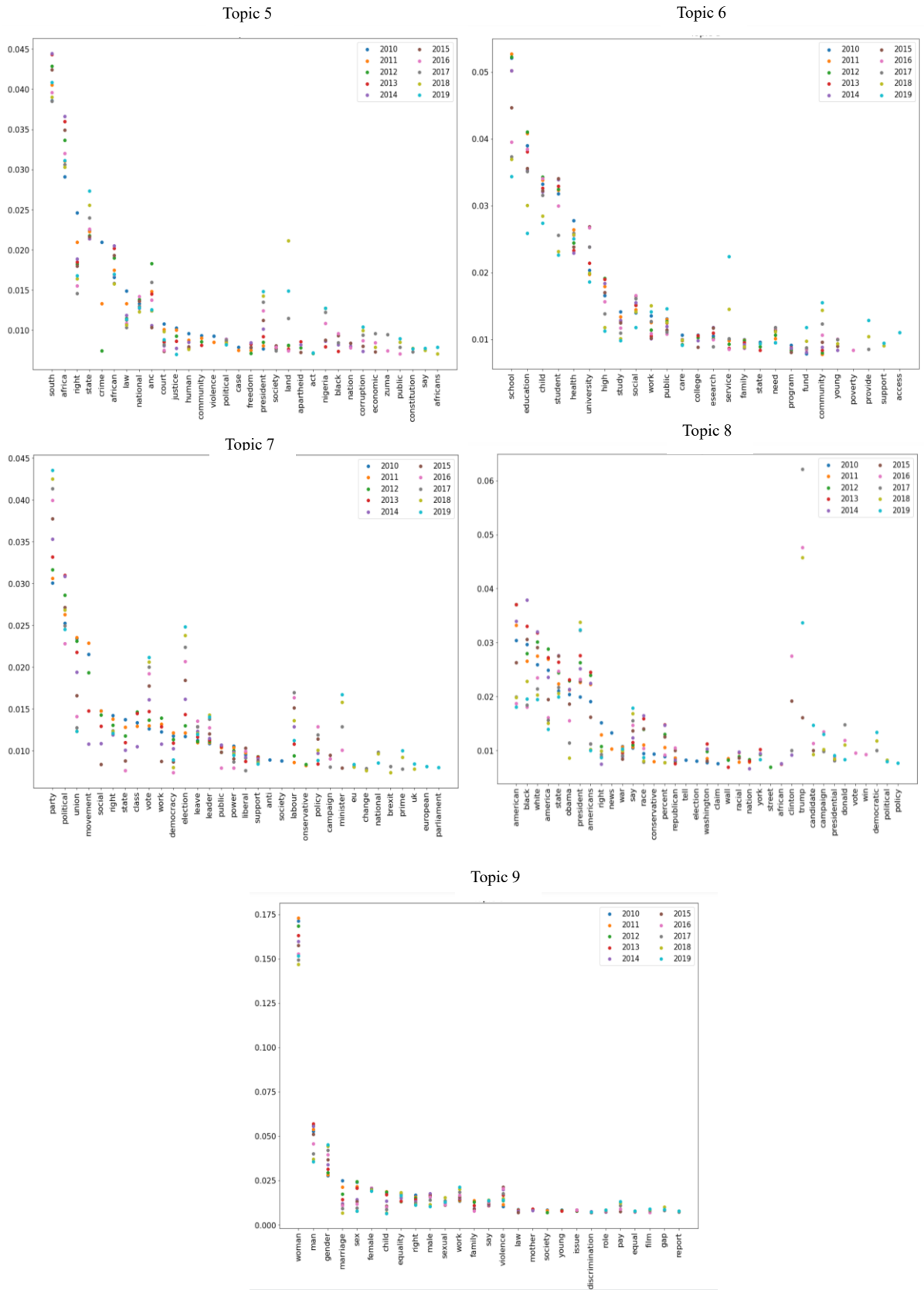


Figure 6 DTM 20 Most Probable words from 2010 to 2019 for each topic

4.3 Word cloud

By: Rui Pan

Word cloud is useful for identifying words appearing more frequently in the text, which will be displayed in larger font sizes. The advantage of the word cloud is to filter the salient words with visualization so that people can easily capture the main idea of the texts. We apply the word cloud to three time periods of the news concerning inequality to capture the change of focus in people's discussion about inequality in news.

To have a glance at the potential word change in the news related to inequality by time, we cut the period by three stages: 2010-2013, 2014-2016, 2017-2019. In the word cloud figures, there are several most common words: government, country, world, etc, which may be the most common context of the discussion about inequality. Especially, "country" and "government" seem to be more salient in 2017-2019 compared to the previous periods ---- more news concerning inequality is discussed related to government and country.

By taking a closer look at the specific words that potentially contain specific content of the topic "inequality", "work" is becoming increasingly salient during the three periods, especially in 2017-2019. We may infer that the inequality related to career and work becomes a focus in the news. While the keyword of "women", is one of the most salient words in the period of 2010-2013, but seems not as salient as before afterward, especially in 2017-2019. According to a previous analysis on topic modeling, "women" mostly occurs in topics of gender inequality. Then it seems to violate our common sense ---- as we generally expect the discussion about gender inequality to increase in recent years. We need to be cautious about the news, as it may be due to a change compared to the relatively increasing appearance of other salient words, but not the absolute change of specific topics.

The most interesting and salient change in the 2017-2019 period is the appearance of the keyword "south africa", which doesn't appear in the previous periods. From the results of topic modeling, the keyword "south africa" is closely related to news concerning some keywords such as African, black, racial, etc, which may cover news related to racial inequality, or the inequality issue in regions of South Africa.

4.4 Part-of-Speech

By: Rui Pan

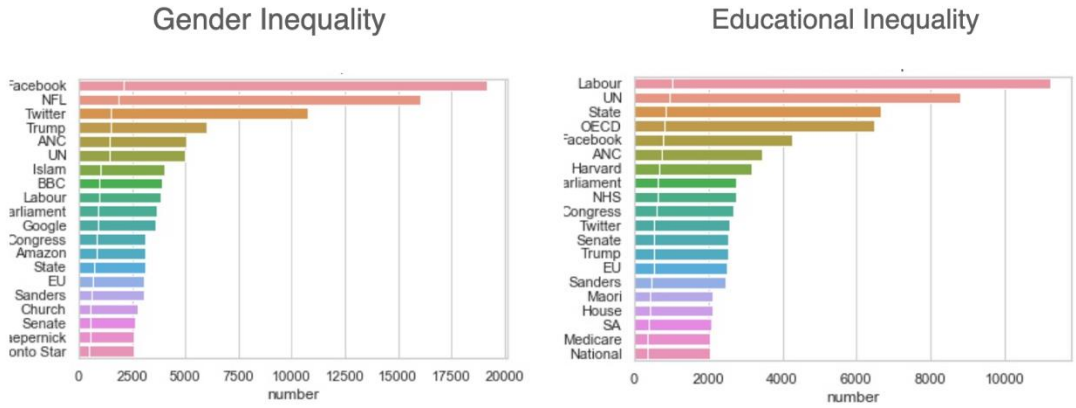
Apart from the general glance at the potential change by time, we are also interested in how the entities in

different topics are different in the news. Which country or region is the public's focus on some specific topic? Where may have a more severe problem in the specific inequality? We take a look at the geopolitical entities to answer that question. Besides, organization entities help us identify the organization that tends to claim or discuss the specific issue, and also the discussion context of the issue.

Named Entity Recognition (NER) can identify named objects with the spaCy package after tokenization. This algorithm supports classifying name entities into Organizations (ORG), Geopolitical Entities (GPE), Nationality or Religious or Political Groups(NORP), law(LAW), etc. We apply the analysis on GPE and ORG to see the different entities related to the discussion of different inequalities.

4.5 ORG: Organization

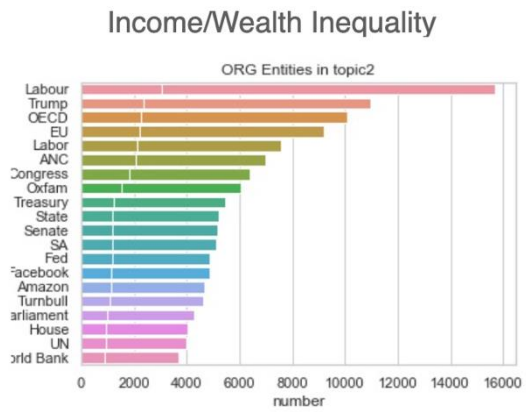
By: Rui Pan

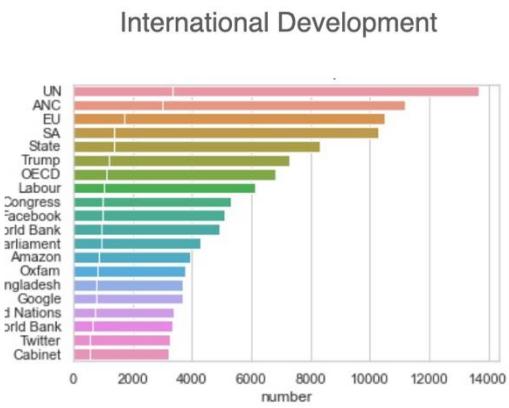


Apart from political organizations, entities of social media, such as "Facebook", "Twitter", are among the top organization words. It seems like gender inequality is also frequently discussed on social media or related to the context of social media. "Islam" is among the top, which reflects the public discussion on the gender inequality problems in the countries and regions with Islam religion.

While for educational inequality, entities of party ("Labour"), political and international organizations (such as "UN", "OECD", "Parliament", "Congress") are more related. The appearance of many international organizations reflects that educational inequality tends to be considered as a global issue with a core value to solve. There are also entities with a strong political impression, such as the UK Labour party, which claims policies on improving education a lot, which may be the reason that it occurs most frequently.

When comparing these two topics, the discussion of gender inequality is adapted to a more personal discussion in the context of social media compared to educational inequality, the latter of which is strongly associated with political and even international organizations. It's kind of corresponding to our common sense, as issues of gender inequality easily attract more public attention and discussion in the private domain, partly because it's close to everyone's position and it easily provokes opposition. The distinction of the private discussion tendency and the formally political context between the two topics is also echoed by the word2vec analysis afterward.

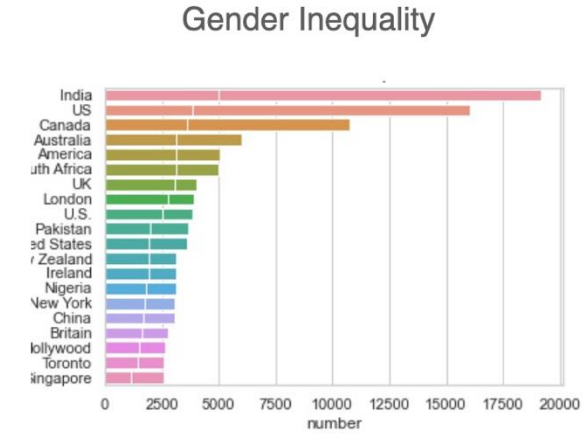


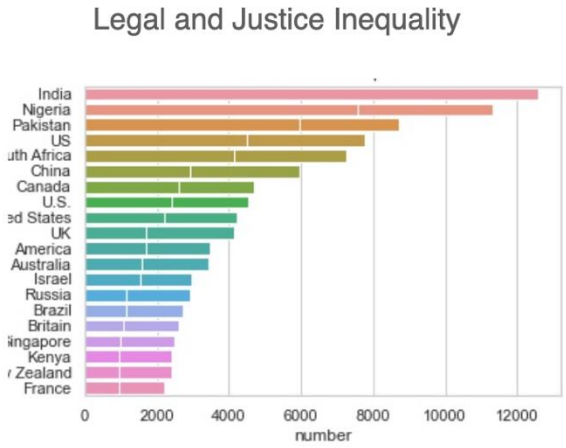


For income and wealth inequality, it's interesting to see Trump should be the second-ranking on most organization entities. But solely from this piece of information, we cannot judge whether the discussion is about Trump's policy on improving income inequality, or news discussing the inferred facts of Trump's accelerating or decreasing income inequality. Apart from that, the Labour party in the UK, which emphasized income inequality appears the most on the organization entities list. For international development, many international organizations occur among the top entities, such as the UN, OECD, World Bank, United Nations. International organizations, such as African National Congress (ANC), European Union (EU), are related to the topic.

4.6 GPE: Geopolitical Entity

By: Rui Pan





For gender inequality, the most salient GPE world is India, which is not surprising to reflect the gender problems in India. When it comes to legal and justice inequality, India, Nigeria, and Pakistan are among the top.

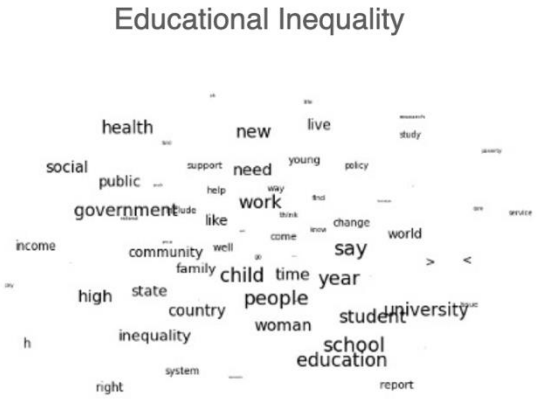
4.7 Word2Vec

By: Rui Pan

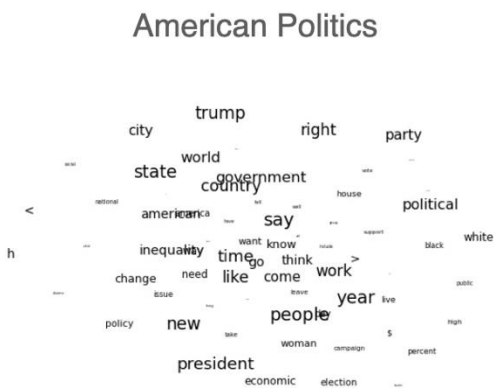
Word2Vec method is one of the ways of word embedding. It can be used for calculating word distance by

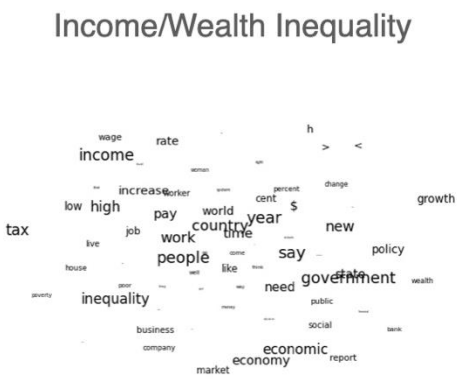
transforming words in the corpus into structured vector space. Then the distance between words is assigned with a vector value.

To see how words are close to or far from one another in the text of each topic, we visualize the word2vec with the top 70 words in each topic.



In educational inequality, the core words are “child”, “family” and “community”, etc. “Country”, “state”, “community”, “family”, “education”, “woman”, “school”, “university” are closely related to one another. When we talk about gender inequality, “family”, “work” and “right” are most related, while “country” and “government” are not salient as educational inequality. It seems like educational inequality tends to be discussed in the context of a public discussion associated with government and the country’s role in it. While for gender inequality, not only those political words are fading, but there are more subjective words, such as “think”, “know” and “like”, which may indicate a more private discussion context with more personal opinion expression and concerns compared to educational inequality.

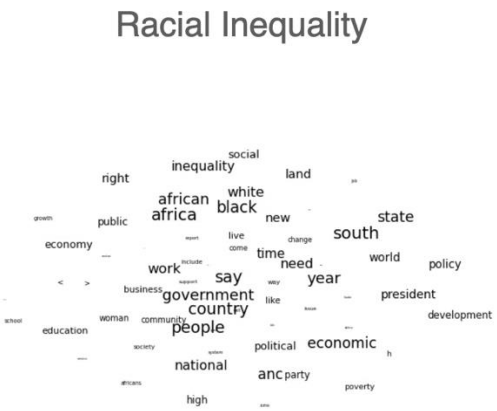




In corpus related to American Politics, several words concerning political issues are mostly correlated, such as government, country, world, state, and trump. The nearest word of “president” is “economic”. This may show

the public concern about electing the president or show the speeches or policy claims issued by presidents. Around the word “political”, which can be taken as one of the core ideas of this topic – American politics, there are “white” and “black” by side. We infer that there are much news discussing American political issue or American political inequality arising from the racial opposition and inequality of opportunities.

In the income/wealth inequality, there are several cores with related words. The word “government” is surrounded by “policy”, “public”, “social” and “wealth” which is more related to public welfare and the policy intervention to affect the income/wealth inequality. Another core, around the word “country”, there are more substantial words, such as “income”, “wage”, “pay”, “work”, “job” and some words describing the status of income such as “high”, “low”. Further than that, concerns related to income are the living cost issue, as we observe “live” and “house”. Another core in the word2vec are words concerning economic issues, such as business and market. So from the word2vec of income/wealth inequality, we could identify several main ideas related to the news, and the potential key points around those main ideas.



What kind of problems are associated with racial inequality? Closely related to words concerning Africa, African are words like “social”, “right”, “economy” and “work”, which are the most core concerned problems compared to the farther words such as “education” and “woman”. In texts of international development inequality, we can identify several issue groups, such as the “income”, “poverty” group, which is the nearest to “inequality”. On the right side, more macro words such as “development”, “growth” and “economy” are nearest to “policy” and “change”. These may indicate the potential logic underlying the news that covers the international development inequality. The former words are at a more micro-level or local issues, while the latter ones are more at a perspective similar to international organizations to identify the macro or global issues and implement policies to promote change.

4.8 Doc2Vec

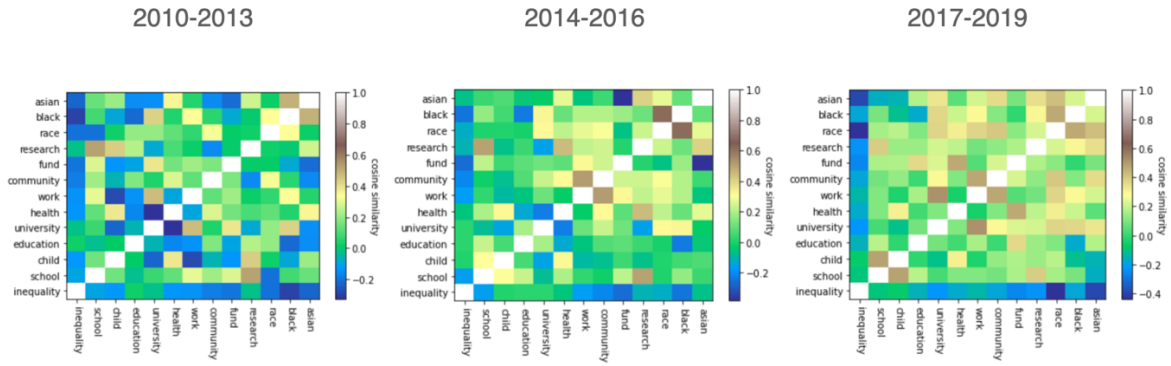
By: Rui Pan

Apart from word2vec, another text analysis tool is Doc2Vec, which represents documents as vectors. Still, we apply Doc2Vec analysis to check the change of closeness between different words in the different topics, from which we may observe the change of people’s focus on different word pairs.

The words for constructing the dimension of doc2vec are mainly from the topic modeling outcome, where different topics are defined with separate keywords. So we choose the top words in the topic modeling, to check

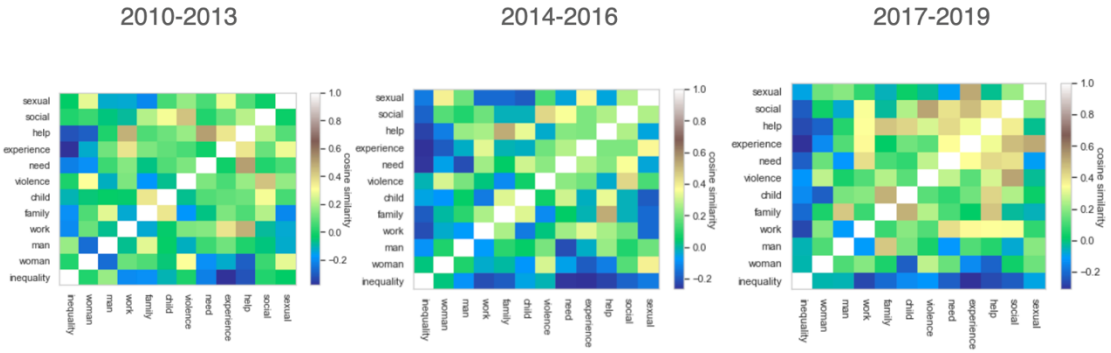
the relation change among these relatively more important words in each topic.

Educational Inequality



For the pair of "work" and "university", the cosine similarity increases a lot through the three periods. This indicates that the discussion combining university education and work seems to increase, especially in 2017-2019. That makes sense as the public's focus on the inequality problems of university admission has been increasing in recent years, which is closely related to personal career development and social mobility.

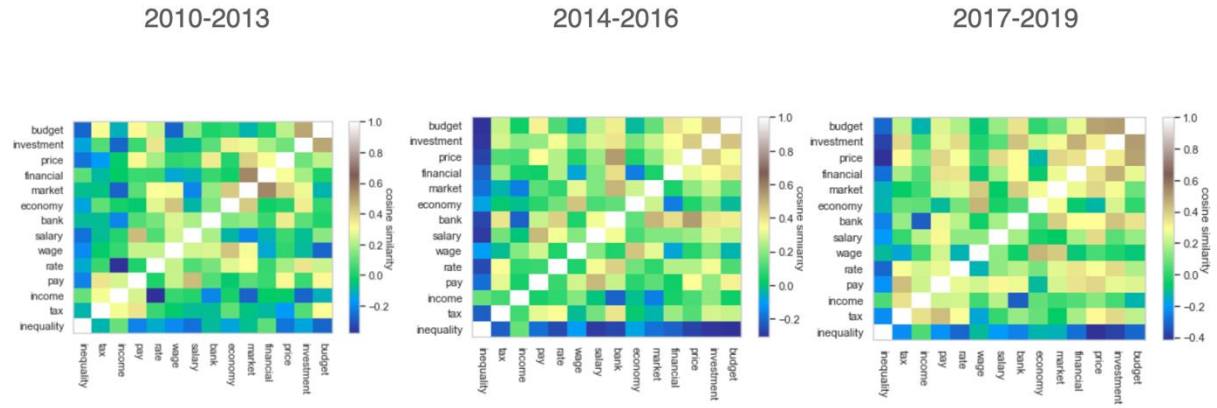
Gender Inequality



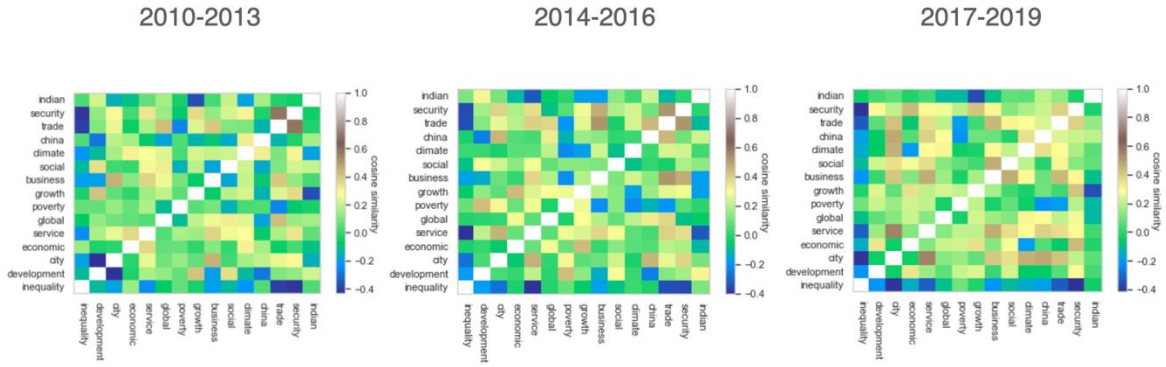
From the texts related to gender inequality, the distance between "woman" and "violence" is becoming more distant through the three periods. Very basically, this reflects the change of discussion that combines these two words are decreasing.

And interestingly, "man" and "family" are closer especially in 2017-2019. There possibly be more discussion on men's role in the family in contrast to the traditional stereotype that women are binding to the family.

Income & Wealth



International Development



In texts related to international development, the cosine similarity of the "social" and "business" pair suddenly increases in the third period of 2017-2019. It possibly reflects the concept of business with social responsibility or corporate social responsibility, etc, so the two words become closely related during that stage.

4.9 Projection

By: Rui Pan

To see how people related to different countries are discussed in the news concerning inequality, we created three dimensions: class, education, employment to get a sense of the context. The word pairs used to define the two dimensions are referenced to Kozlowski et al. (2019).

Dimension	**Keywords: one side**	**Keywords: the other side**
class	Rich, expensive, wealthy, affluence	Poor, cheap, inexpensive, poverty
education	Literate	Illiterate
employment	Employer, owner, manager, boss, capitalist	Employee, worker, staff, proletarian

For the class dimension, “Japanese” and “American” are among the top, which is defined as “rich”, “expensive” etc. “Asian” and “Chinese” are at the middle part, while “African”, “Mexican” and “Indian” are discussed on the opposite side.

For the education dimension, “Indian” and “Chinese” are at the top, which is defined as “literate”. It seems not surprising, as Indian and Chinese people are thought to emphasize education a lot. For the employment dimension, “Japanese”, “Chinese”, “Asian” and “American” tend to be discussed at the side of “employer”.

Due to the limitation of the text corpus, we couldn’t put more words to define the education dimension, because words such as “education” are not included in the corpus. It will be more precise to define the dimensions with more related words to capture.

5. Future improvements

By: Xi Cheng

Methods. This is the first time that I deal with such a big size of text data, so it was a nice trial and practice. However, I realize it is just a start, and I think the following points can be considered when designing and conducting further research:

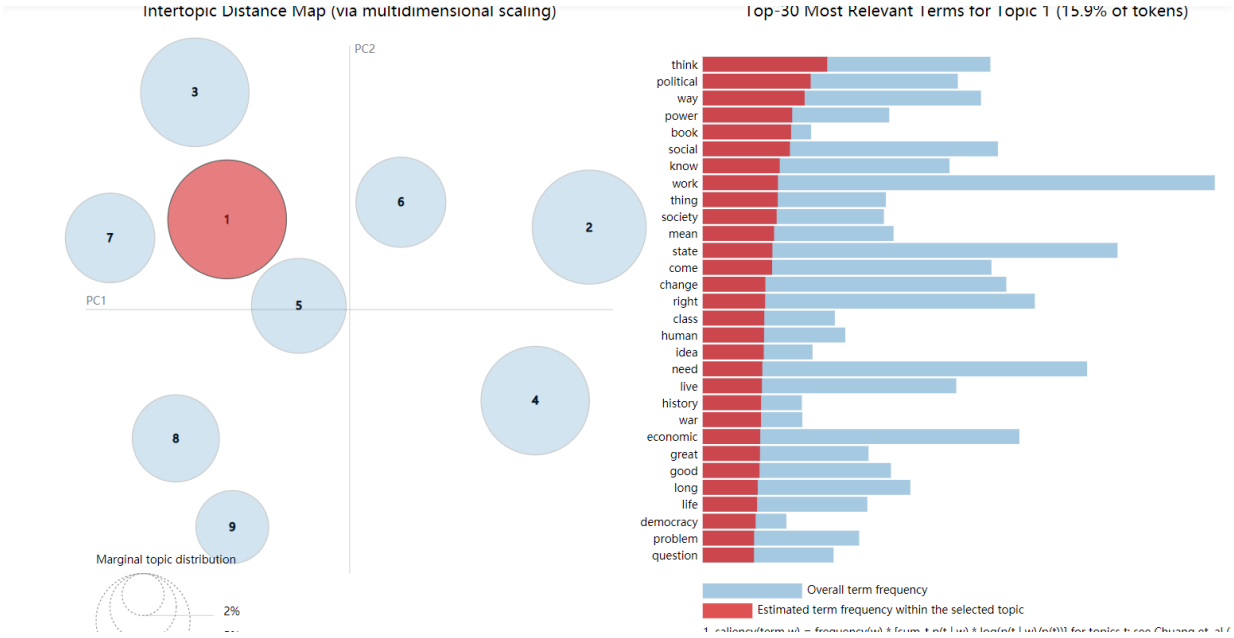
(1). More restricted data cleaning. There is still a word like “say” existing in some topics, which results in some vague topic themes.

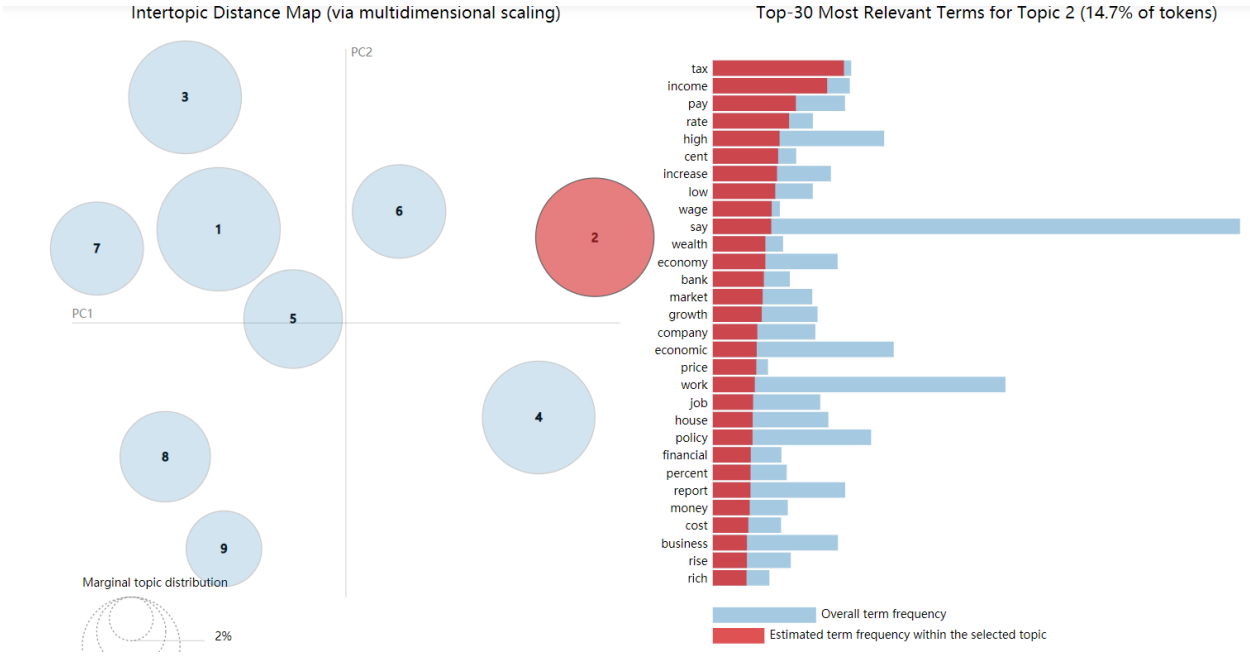
(2). A better method to identifying each article’s topic. Now the number of articles on each topic is relatively big compared to the whole size of the corpus. Although it can be possible that one article covering several inequality topics at the same time, a better threshold is needed.

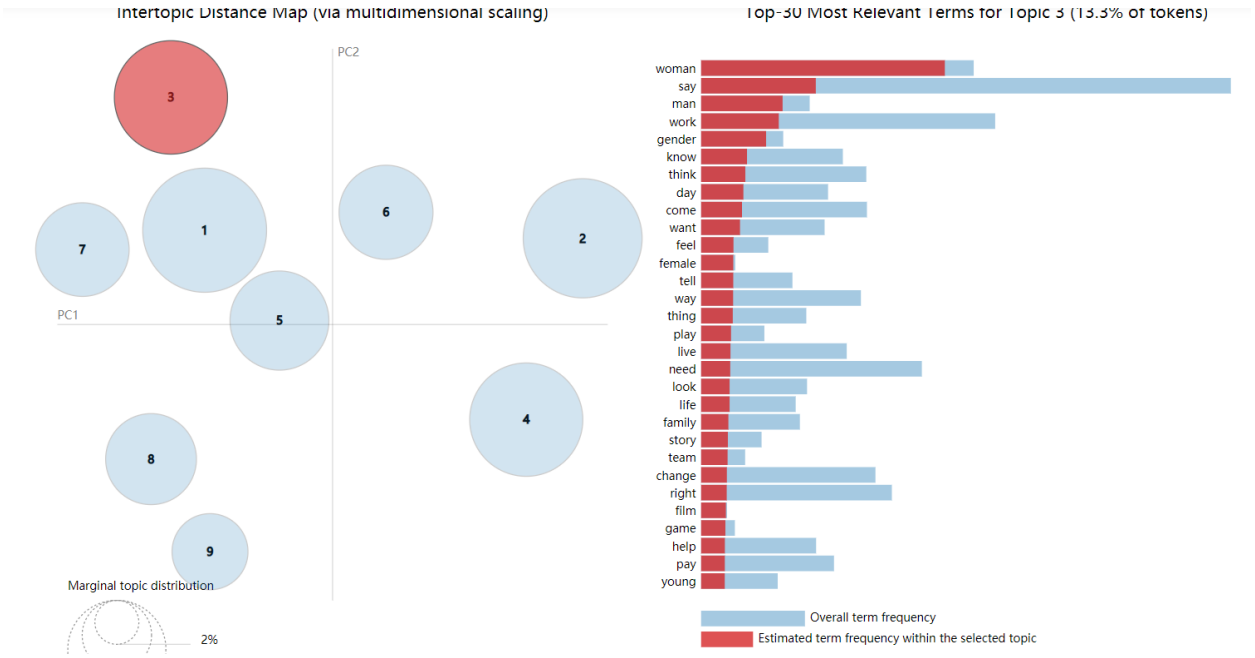
(3). A better model evaluation plan. To make things easy, this time, I just considered the Cv topic coherence score and visualization plots. Other methods like u_mass and perplexity may also be needed.

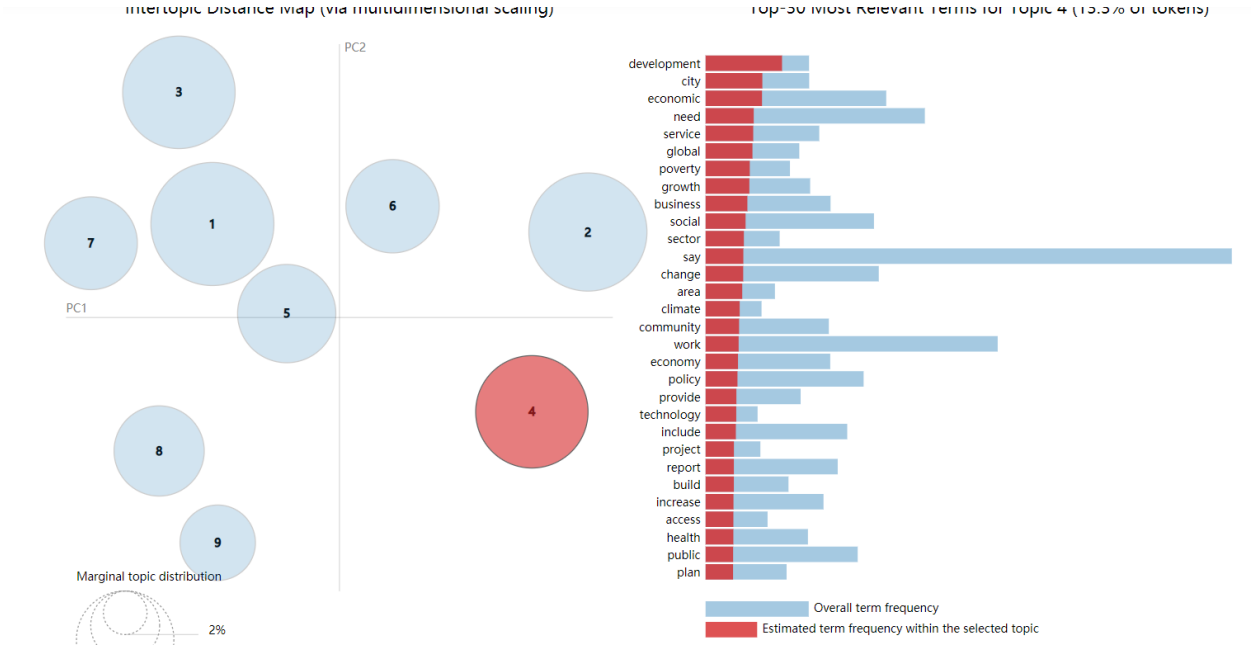
Theories. This project is preliminary research on news coverage of inequality, so the research objective is to see the trend. However, considering news itself is biased and more deep topics behind the inequality news, I need to consider more interesting research questions later after learning more about this topic.

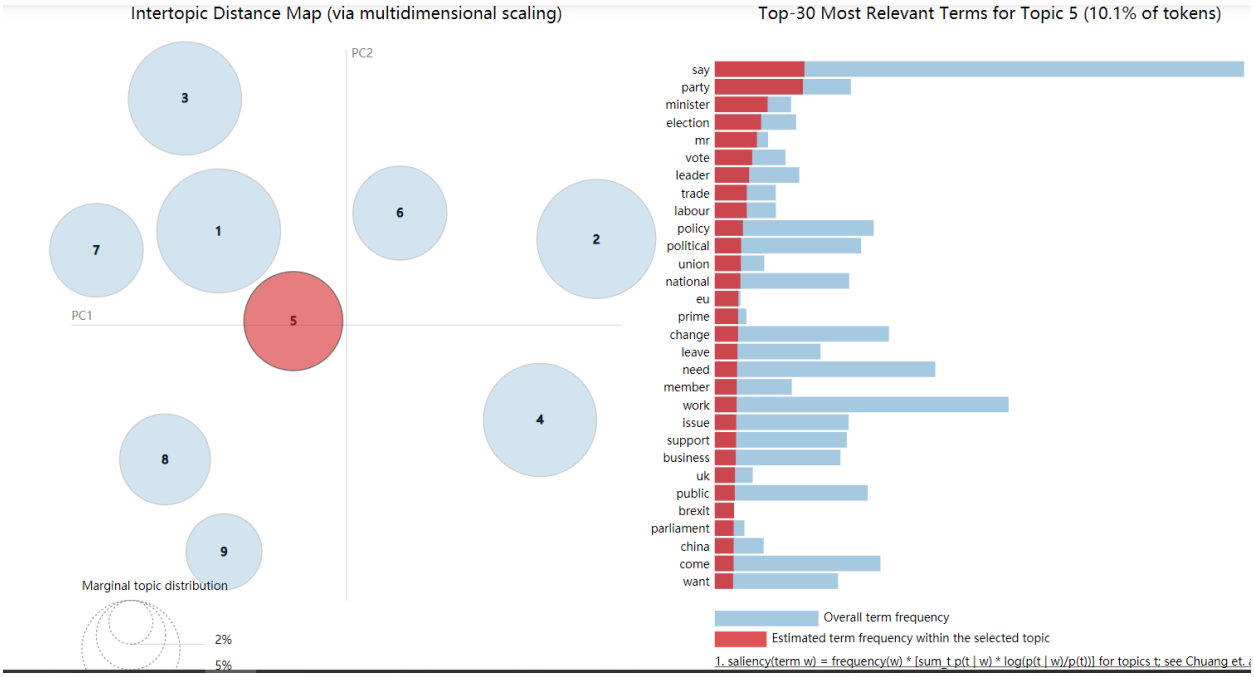
Appendix A Intertopic distance map and top-30 most relevant terms per topic of the 9-topic LDA model

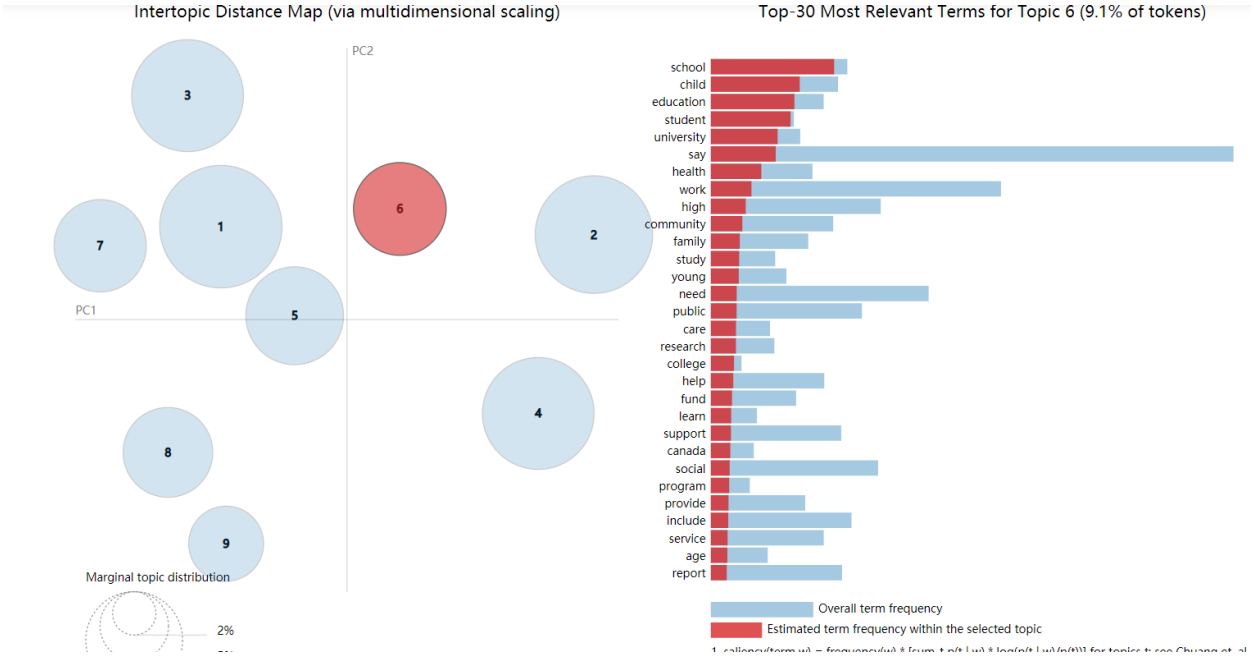


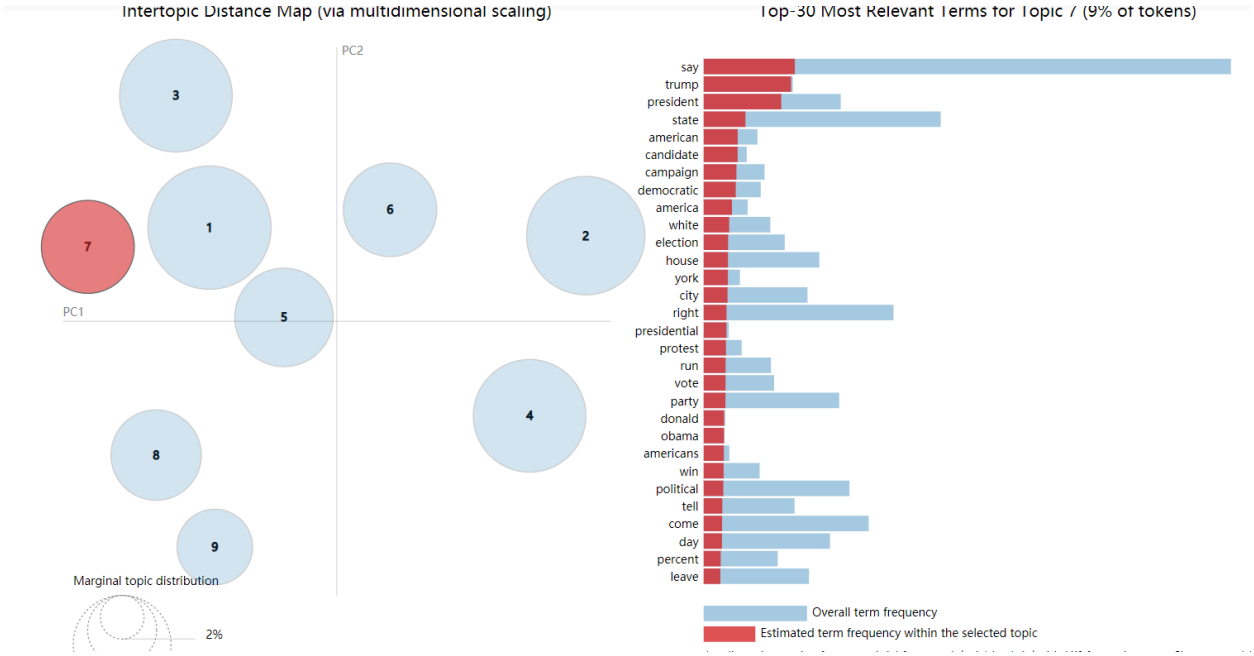


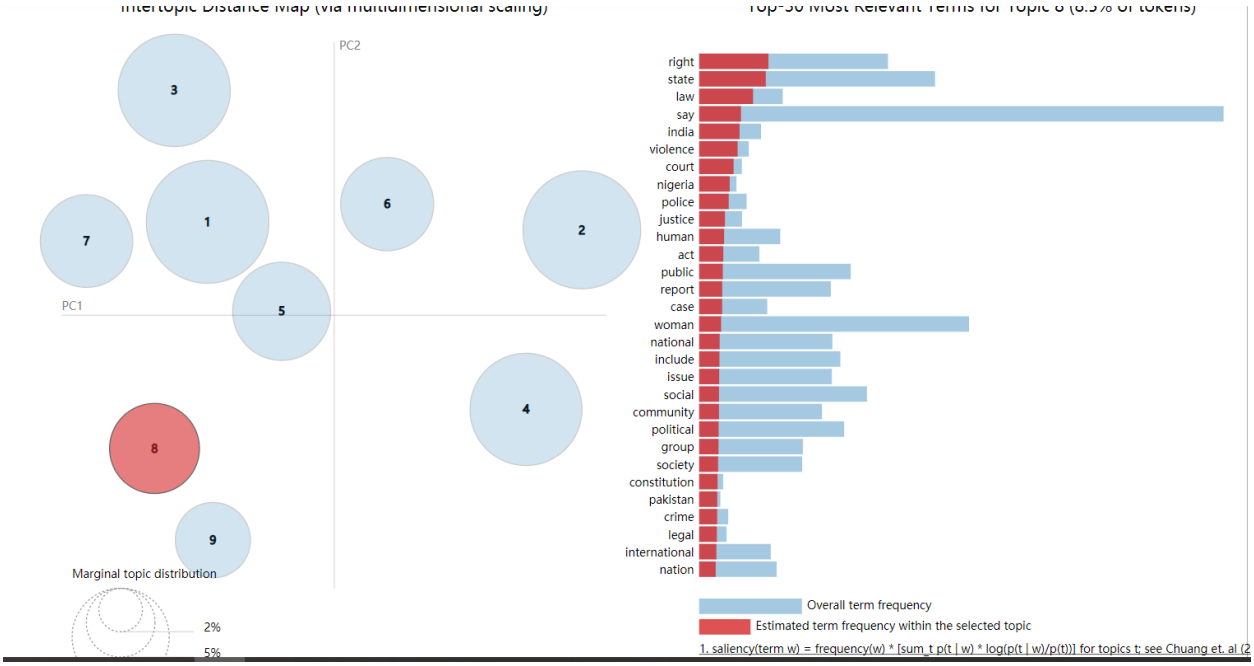

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
1
6
2
7
5
4
8
9
Marginal topic distribution
2%
Top-30 Most Relevant Terms for Topic 3 (13.3% of tokens)
woman
say
man
work
gender
know
think
day
come
want
feel
female
tell
way
thing
play
live
need
look
life
family
story
team
change
right
film
game
help
pay
young
Overall term frequency
Estimated term frequency within the selected topic

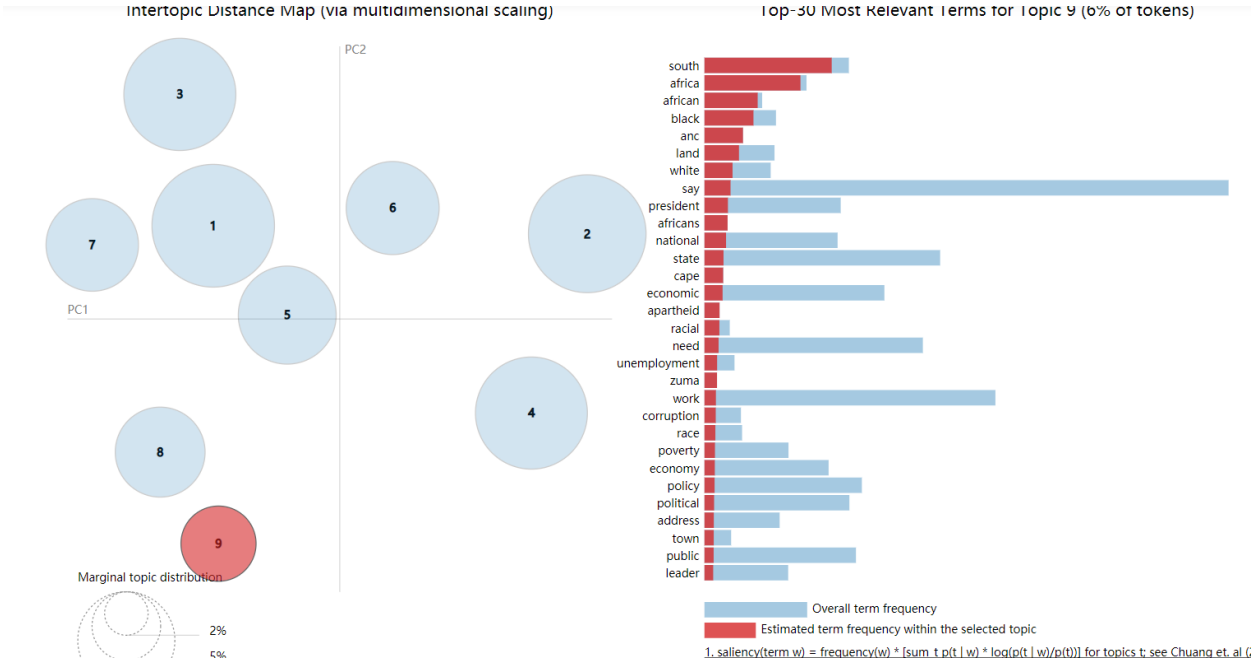

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
1
6
2
7
5
4
8
9
Marginal topic distribution
2%
Top-30 Most Relevant Terms for Topic 4 (13.3% of tokens)
development
city
economic
need
service
global
poverty
growth
business
social
sector
say
change
area
climate
community
work
economy
policy
provide
technology
include
project
report
build
increase
access
health
public
plan
Overall term frequency
Estimated term frequency within the selected topic


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
1
7
6
2
5
4
8
9
Marginal topic distribution
2%
5%
Top-30 Most Relevant Terms for Topic 5 (10.1% of tokens)
say
party
minister
election
mr
vote
leader
trade
labour
policy
political
union
national
eu
prime
change
leave
need
member
work
issue
support
business
uk
public
brexit
parliament
china
come
want
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. a


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
1
7
6
2
5
4
8
9
Marginal topic distribution
2%
Top-30 Most Relevant Terms for Topic 6 (9.1% of tokens)
school
child
education
student
university
say
health
work
high
community
family
study
young
need
public
care
research
college
help
fund
learn
support
canada
social
program
provide
include
service
age
report
Overall term frequency
Estimated term frequency within the selected topic


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
1
7
6
2
5
4
8
9
Marginal topic distribution
2%
Top-30 Most Relevant Terms for Topic 7 (9% of tokens)
say
trump
president
state
american
candidate
campaign
democratic
america
white
election
house
york
city
right
presidential
protest
run
vote
party
donald
obama
americans
win
political
tell
come
day
percent
leave
Overall term frequency
Estimated term frequency within the selected topic


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
3
1
7
6
2
5
4
8
9
Marginal topic distribution
2%
5%
Top-30 Most Relevant Terms for Topic 8 (8.5% of tokens)
right
state
law
say
india
violence
court
nigeria
police
justice
human
act
public
report
case
woman
national
include
issue
social
community
political
group
society
constitution
pakistan
crime
legal
international
nation
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
Marginal topic distribution
2%
5%
Top-30 Most Relevant Terms for Topic 9 (6% of tokens)
south
africa
african
black
anc
land
white
say
president
africans
national
state
cape
economic
apartheid
racial
need
unemployment
zuma
work
corruption
race
poverty
economy
policy
political
address
town
public
leader
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2

Reference

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. American Sociological Review, 84(5), 905-949.

S. Syed and M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation," 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 2017, pp. 165-174, doi: 10.1109/DSAA.2017.61.