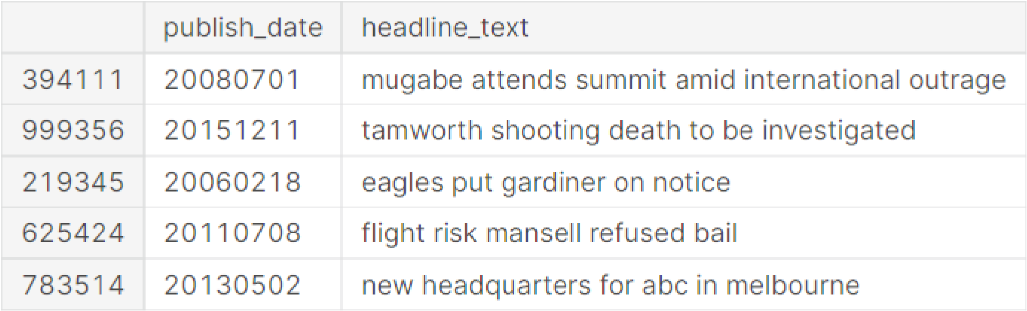
Dataset

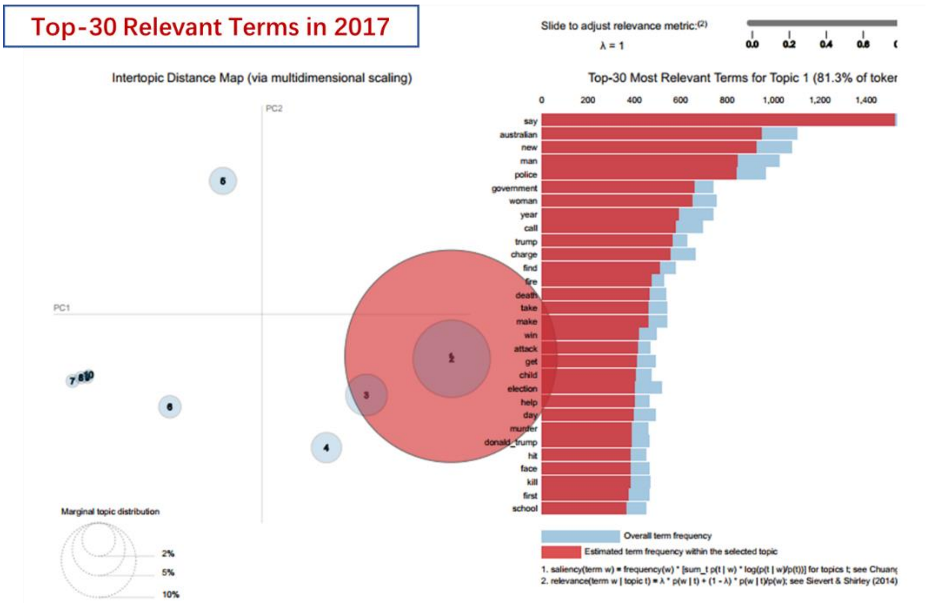
The dataset in this project is from the Kaggle open source, A Million News Headlines.

With a volume of two hundred articles per day and a good focus on international news, we can be fairly certain that every event of significance has been captured here. Digging into the keywords, one can see all the important episodes shaping the last decade and how they evolved over time.

The dataset contains data of news headlines published over a period of nineteen years.
Sourced from the reputable Australian news source ABC (Australian Broadcasting Corporation)
Start Date: 2003-02-19 ; End Date: 2021-12-31
The dataset is too large to train the model (1213004 news headlines), which need more than 10h, so I take the news from 2019.1.1 to 2021.12.31.

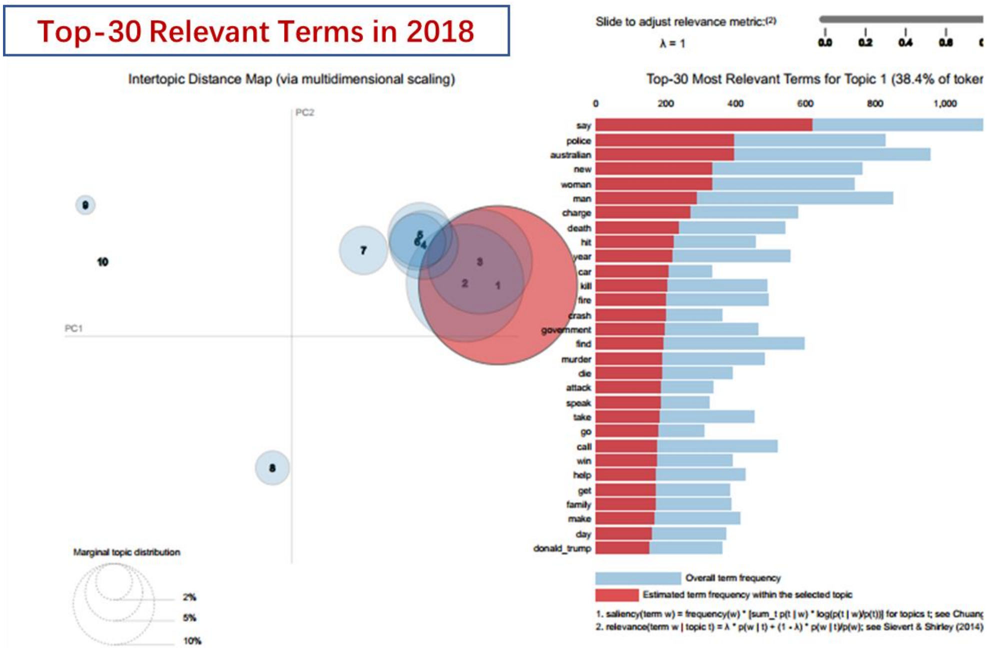


	publish_date	headline_text
394111	20080701	mugabe attends summit amid international outrage
999356	20151211	tamworth shooting death to be investigated
219345	20060218	eagles put gardiner on notice
625424	20110708	flight risk mansell refused bail
783514	20130502	new headquarters for abc in melbourne

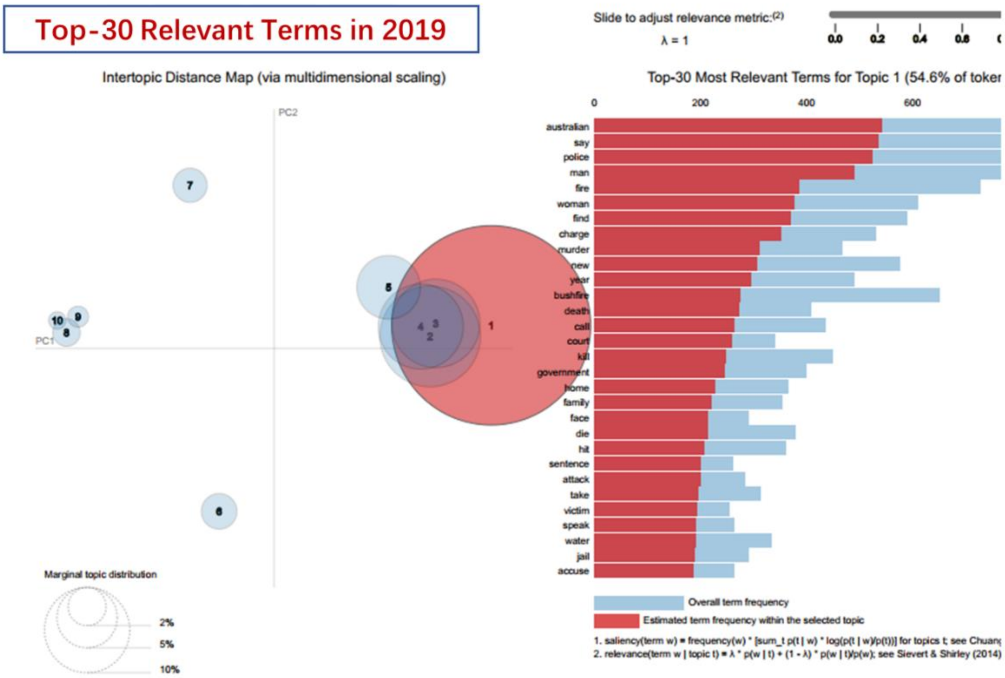


Top-30 Relevant Terms in 2017

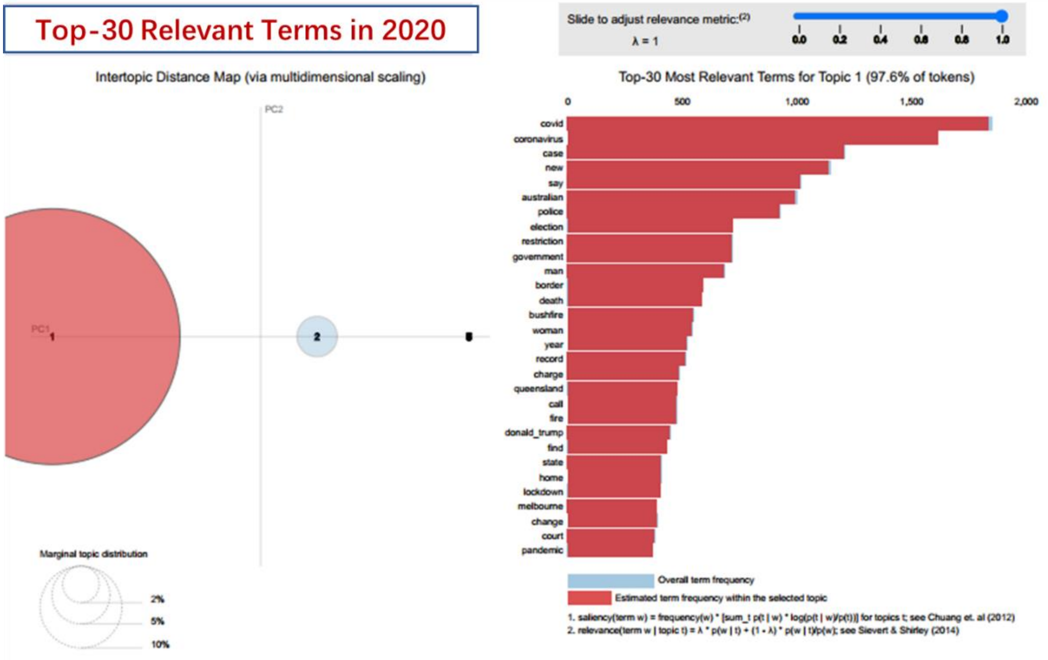
Top-30 Relevant Terms in 2018



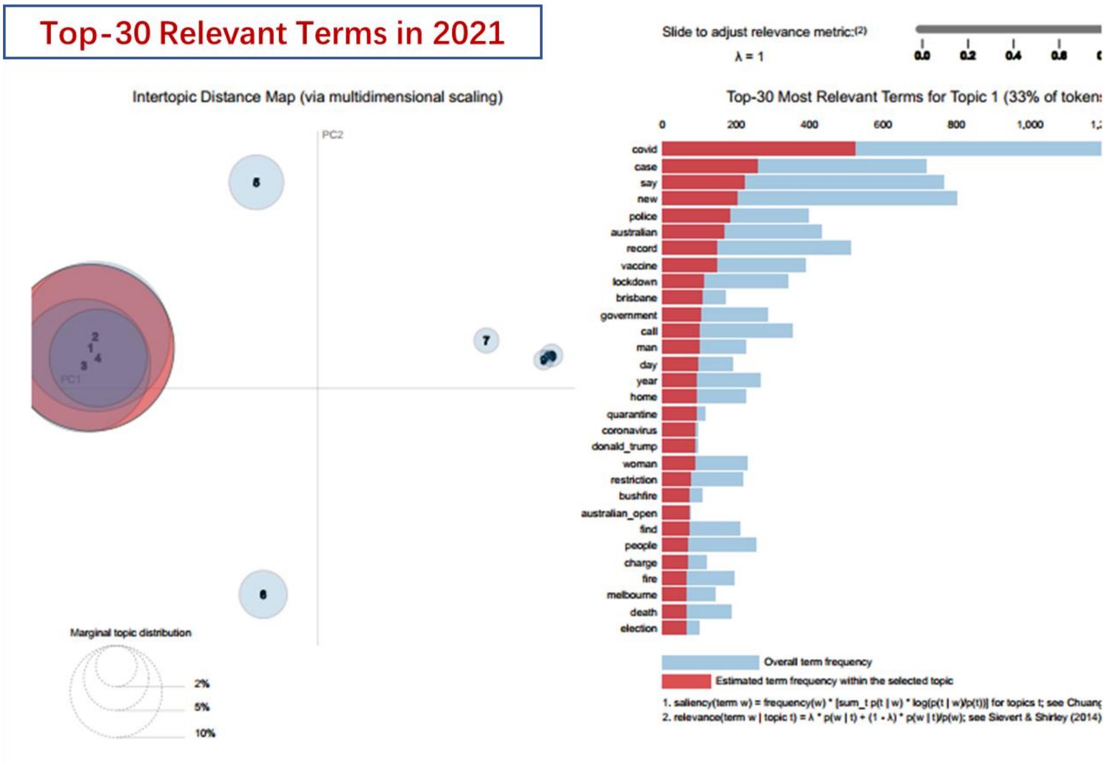
Top-30 Relevant Terms in 2019



Top-30 Relevant Terms in 2020



Top-30 Relevant Terms in 2021



As we can see from the result, the words "Australian, say, policy, men, women" appear a lot of times, which is reasonable, since our dataset comes from Australian Broadcasting Corporation and other words are the common words used in the news. Ignore these words, the key words in the news headline in 2017 is "trump" "government" "fire"; in 2018 are "crash" "fire" "death", in 2019 are "fire" "murder" "bushfire"; in 2020 are "covid" "election" “restriction”; in 2021 are “covid” “vaccine” “lockdown”. From these key words we can recall some important event during that year, for instance, in 2017, the second year of Trump's presidency of the United States, he adjusted and promulgated many bills and bans, which attracted the attention of the whole world and in 2019, the coronavirus was just discovered in China and didn't get the world's attention, while in the next 2 years, people are forced to lock down because of the coronavirus epidemic.