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.

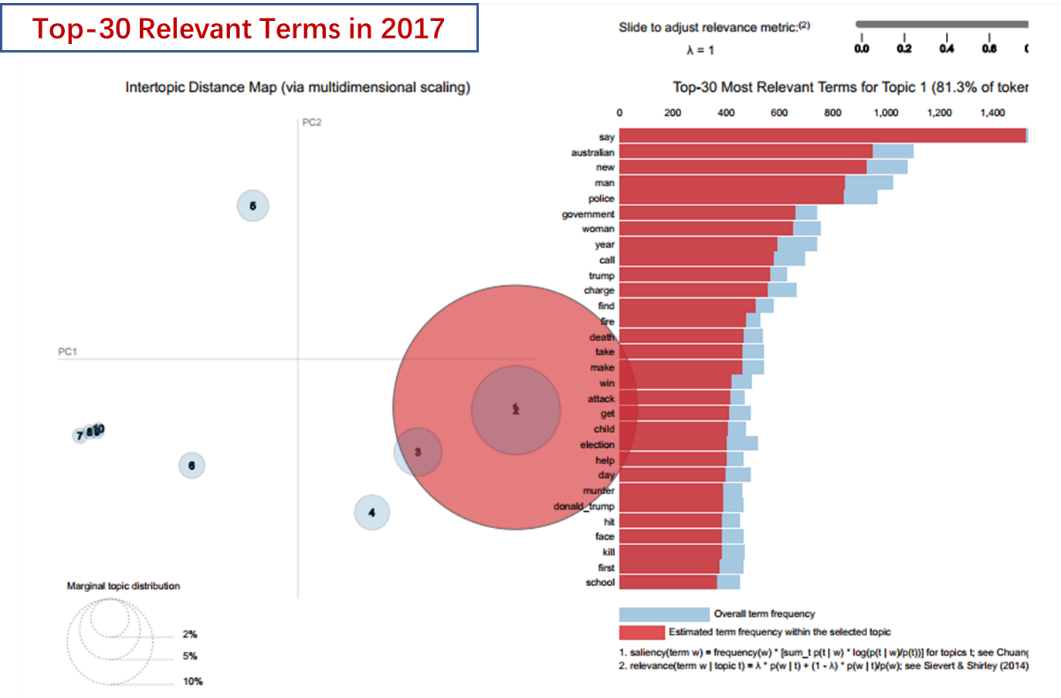
Sourced from the reputable Australian news source ABC (Australian Broadcasting Corporation)

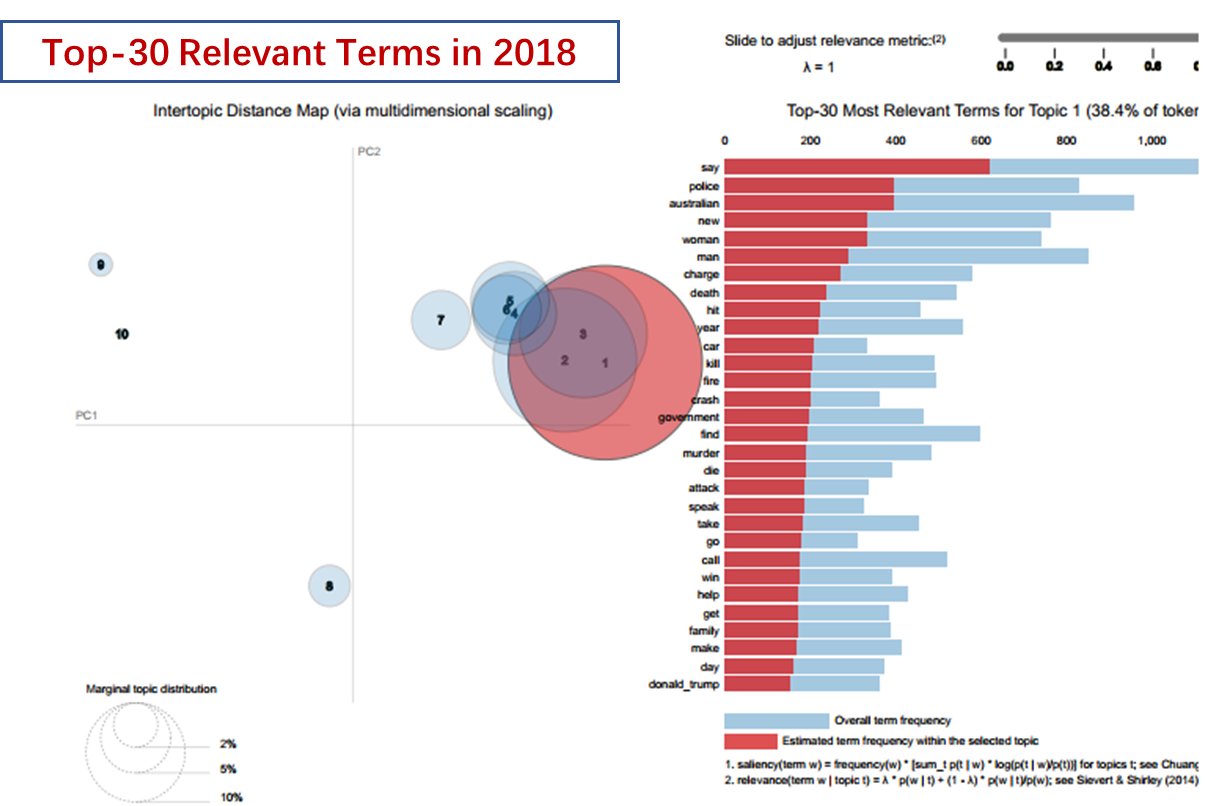
Start Date: 2003-02-19 ; End Date: 2021-12-31

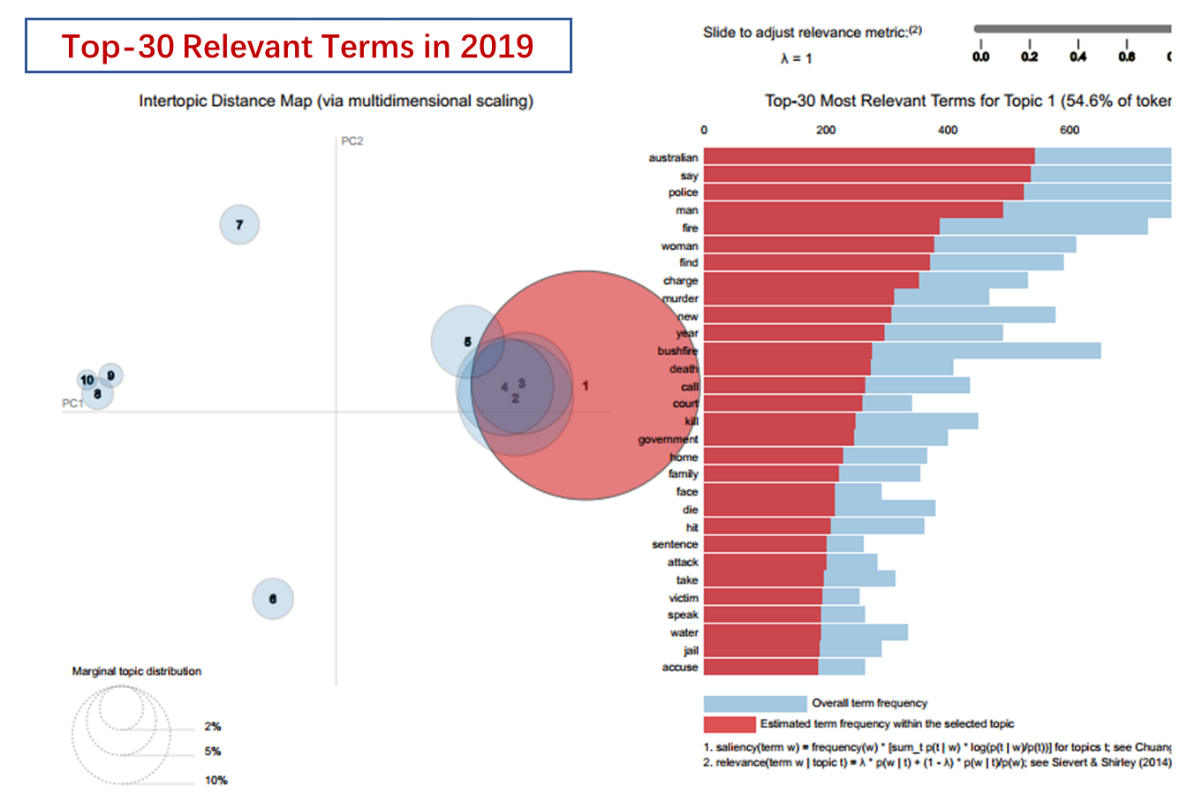
The dataset is too large to train the model (1213004 news headlines), which need more than 10h, so I take the news from 2019.1.1 to 2021.12.31.

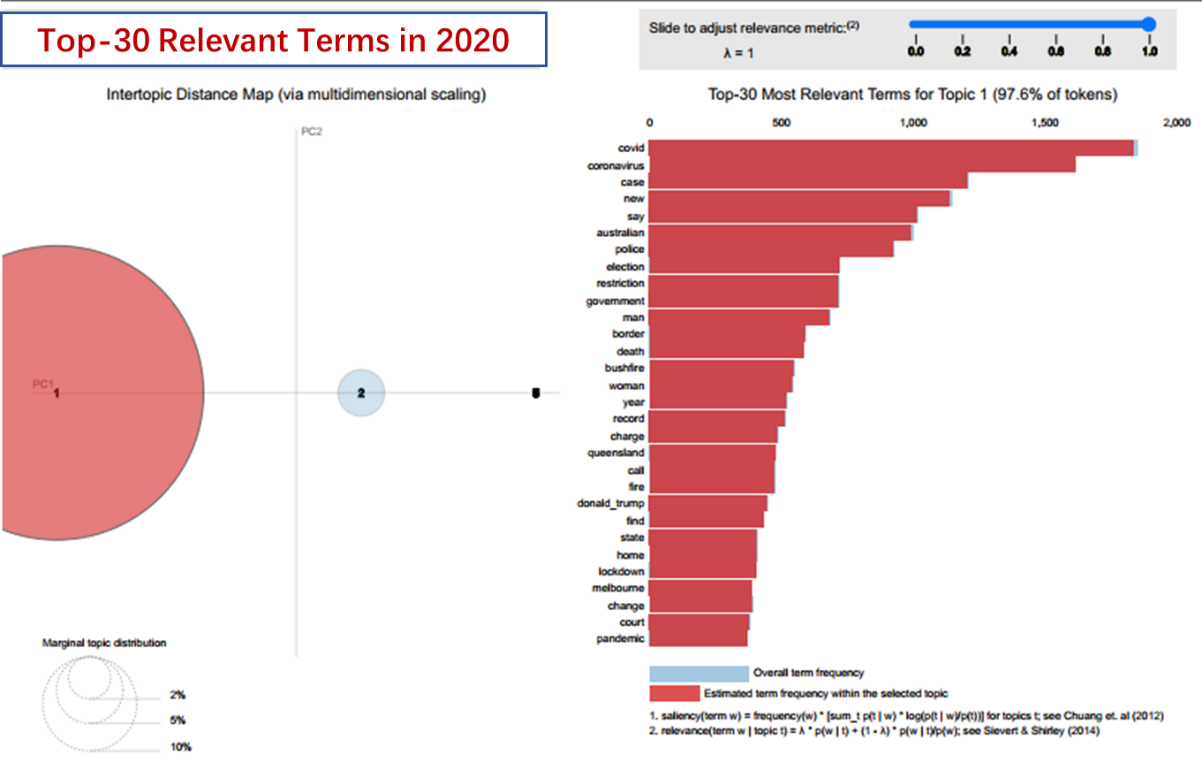
Graphical user interface, text, application, email

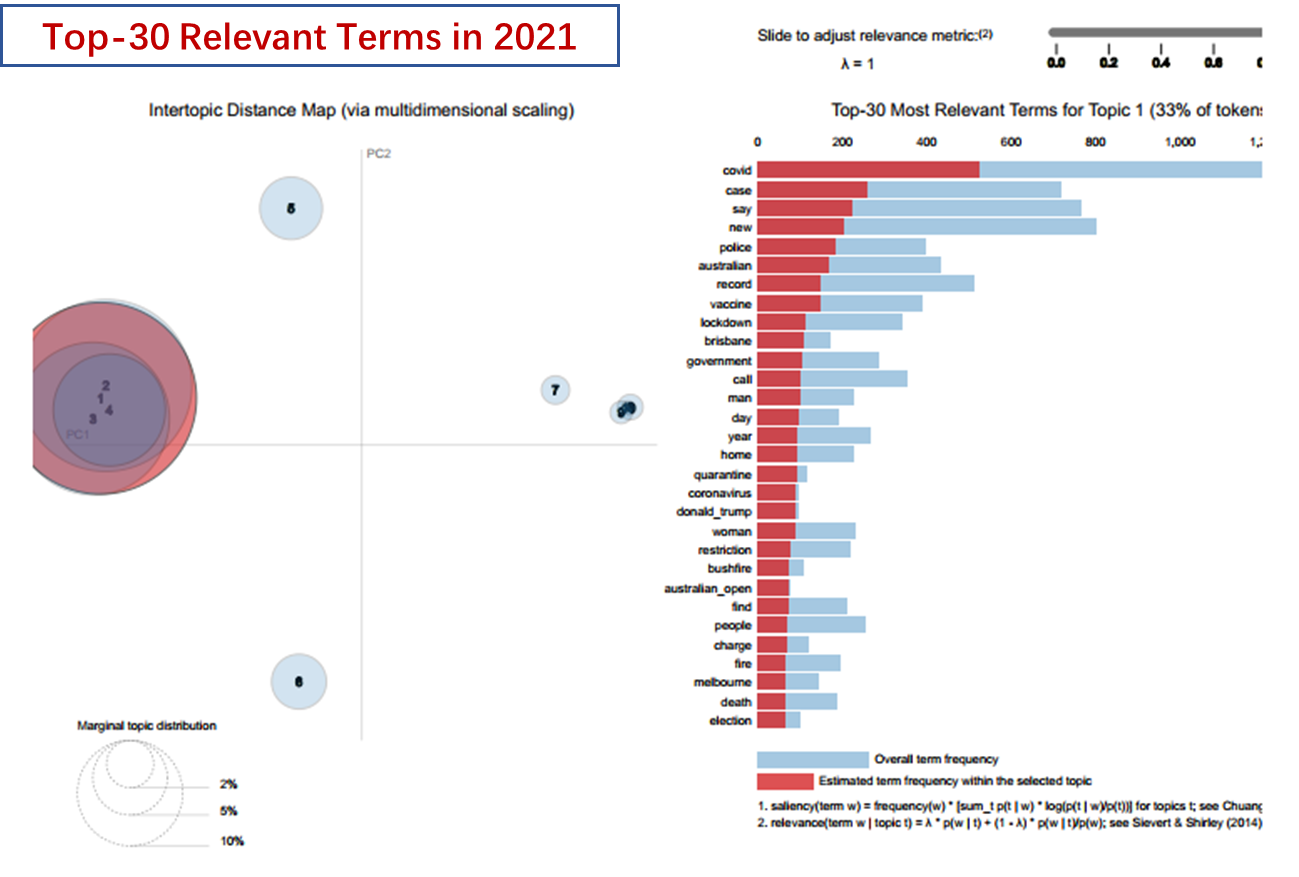
Description automatically generated











As we can see, for the topic 1, the key words in the news headline is covid.