Laura Federline – Health and Life Sciences R&D

Analysis of Semi-Structured Text using WHO COVID-19 Situation Reports

Text Origin

Background
The WHO releases daily situation reports on COVID-19. This project focuses on the first 100 reports released January 21st – April 29th.

Report Structure
- Published as a PDF
- 100 reports
- 916 pages total
- 312,831 words
- Semi-structured (deliberate structure & text separable by headers)

screenshot of 100th PDF

Project Goals
- Quantify changes in language and content
- Identify advisory themes
- Evaluate the degree to which these reports reflect current data

Data Preparation

1. Python

```
#format report string
report_text = report_text.replace('\n','')
report_text = ' '.join(report_text.split())

#make list of rawheaders using regex
rawheaders = re.findall(r"\b(?:[A-Z]+\s*:*)+\s+",
                        report_text)
```

snippet of Python script to extract & format text using PYPDF2 package

sitrep	header	text
1	beginning	1 Data as reported by: 20 January 2020
1	SUMMARY	Event highlights from 31 December 2019 to 20 ...
1	SURVEILLANCE	Reported incidence of confirmed 2019-nCoV cas...
1	PREPAREDNESS AND RESPONSE	: WHO: WHO has been in regular and direct cont...
1	COUNTRY RESPONSE	: China: National authorities are conducting a...
...	...	...
100	SURVEILLANCE	Table 1. Countries, territories or areas with...
100	PREPAREDNESS AND RESPONSE	To view all technical guidance documents rega...

resulting text data structure

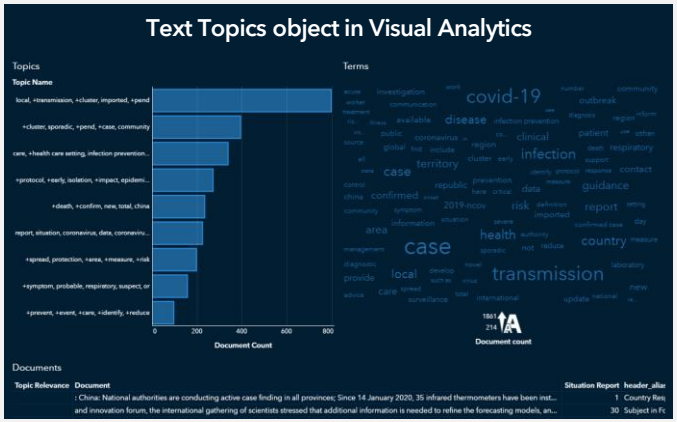
2. SAS Studio
- Separate text into smaller units of observation
- Add descriptor variables and COVID statistics (e.g. death count)
- Clean text further to prepare for analysis

3. Visual Analytics

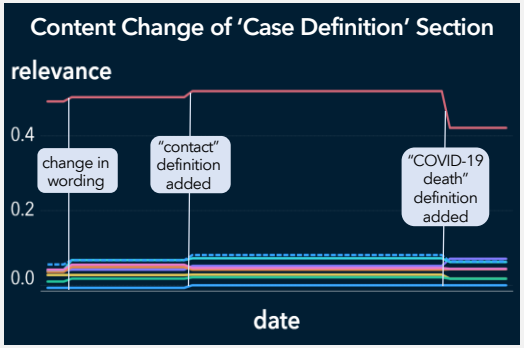
word cloud of most frequent words

Explore text contents with VA objects and adjust data in SAS Studio.

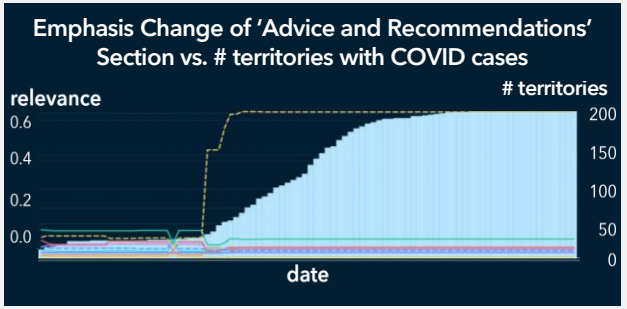
Text Analysis



9 text topics with little overlap were derived. Each text document is assigned a relevance for each text topic.



Each line represents a topic. Stagnant time periods indicate that the Case Definitions were identical between reports.



At the same time territories (bars) began to increase, the 'risk assessment' topic (yellow line) became dominant in 'Advice and Recommendations'.

source: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports