Laura Federline – Health and Life Sciences R&D

About Me

- Major: Statistics, minoring in Biological Sciences
- Graduating November 2020
- Fun Facts: Pack Clogging member, studied abroad in the Czech Republic
- 3rd summer at SAS

NC STATE UNIVERSITY

Ask me about my Previous Projects

- Intern Hack-For-Good event
- Served as an intern virtual lunch leader
- Published data stories on GatherIQ
- Obtained SAS Certified Specialist Certification
- Statistical analysis of top diseases in the US
- Data Visualization Techniques for Communicating Long-Term Trends of Diabetes

This Summer's Project

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

Logistics for Presentation

- I will start my presentation at **10 past the starting time (8:40am; 10:25am; 11:25am)** to allow employees to enter the breakout rooms.
- All questions will be held until the end of the presentation. Please use the chat if you'd like to send questions in advance.

Laura Federline – Health and Life Sciences R&D

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

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)

Text Origin

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

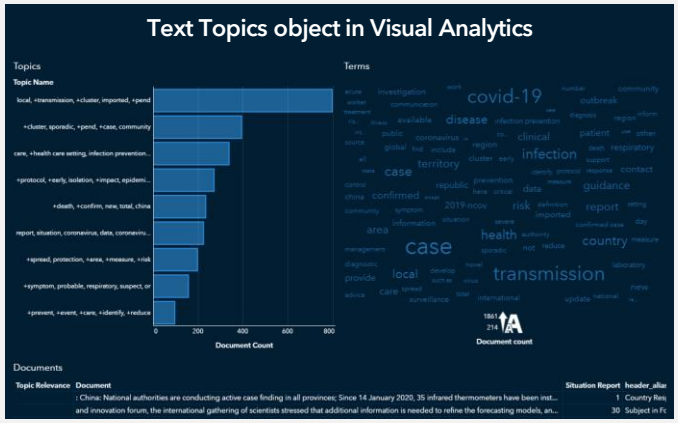
3. Visual Analytics

word cloud of most frequent words

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

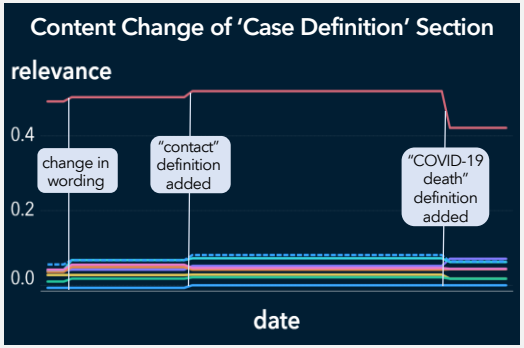
Text Analysis

Text Topics object in Visual Analytics



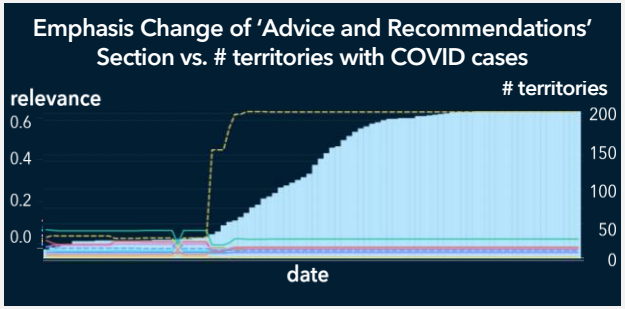
9 text topics with little overlap were derived. Each text document is assigned a relevance for each text topic.

Content Change of 'Case Definition' Section



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

Emphasis Change of 'Advice and Recommendations' Section vs. # territories with COVID cases



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

DEMO

Takeaways

Topics found in Text

- case geography
- counting cases
- prevention
- epidemiological
- risk assessment
- medical aspects
- healthcare
- case origin
- report info

Use of this Project

- Know content without having to read
  - Ex. 'Advice and Recommendations' do not change after report 42
- Identify advisory themes
- Identify patterns in content change and COVID trends

words of topic	topic
+cluster, sporadic, +pend, +case, community	case geography
local, +transmission, +cluster, imported, +pend	case origin
+death, +confirm, new, total, china	counting cases
+protocol, +early, isolation, +impact, epidemiological	epidemiological
care, +health care setting, infection prevention, +setting, prevention	healthcare
+symptom, probable, respiratory, suspect, or	medical aspects
+prevent, +event, +care, +identify, +reduce	prevention
report, situation, coronavirus, data, coronavirus disease	report info
+spread, protection, +area, +measure, +risk	risk assessment

Contact Information

- SAS email: laura.federline@sas.com
- College email: lefederl@ncsu.edu
- https://www.linkedin.com/in/laura-federline/