PART 2: Text Pre-processing and Keywords Extraction

Naiyu Jiang

The University of Chicago
MACS 30122 Computer Science with Social Science Applications 2
Final Project

March 19, 2021

Special Packages

- **jieba**: this is a word segmentation module for both English and Chinese.
- **tqdm**: this module makes a terminal progress bar.
- **nltk.corpus**: this module contains a convenient list of all English words.
- **matplotlib.pyplot**: this module creates various data visualizations.
- **scattertext**: this module creates beautiful visualizations of what words and phrases are more characteristics of a given category.
- **seaborn**: this is a data visualization module based on matplotlib.
- **wordcloud**: this module is used to generate word cloud.

Five Steps - 1

STEP 1

We construct a class called “data center”, storing methods used to convert the scraped data into the prepared data for text analysis. There are around eight methods for pre-processing purposes: (1) read data from json; (2) pre-process data (split words/ stop words/ lowercase); (3) count frequency; (4) calculate top k frequency; (5) transform the date; (6) collect data by date; (7) transform date to quarter; (8) transform quarter to date.

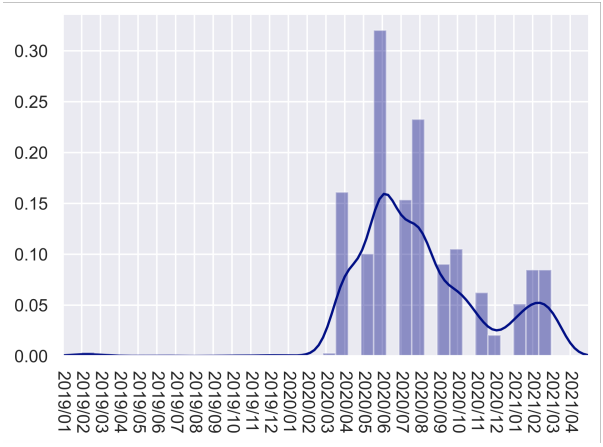
Five Steps - 2

STEP 2

We analyze the whole set of bills from 2019 to 2021 and draw a bar chart to capture the number of bills related to COVID-19 over time. We find that most of the COVID-related bills are proposed in 2020. Therefore, for this project, we want to narrow our scope down and only analyze COVID bills in 2020.

STEP 2 Output

The number of COVID bills changes over time



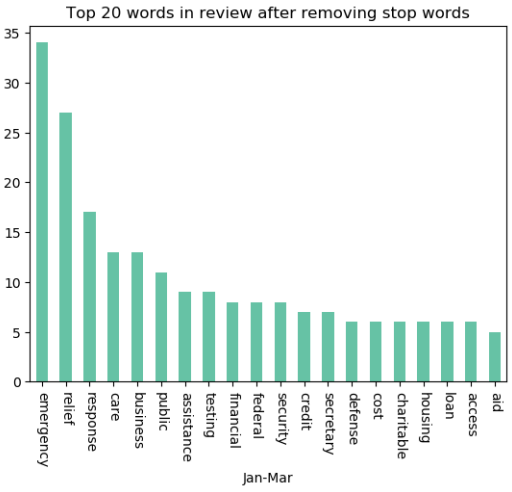
Five Steps - 3

STEP 3

We extract keywords from each bill after removing stop words and identifying as correct English words. Taking three months (one quarter) as a period, we group keywords together and calculate the frequency for each word in each period. Then, we draw plots on the top 20 frequently occurred words.

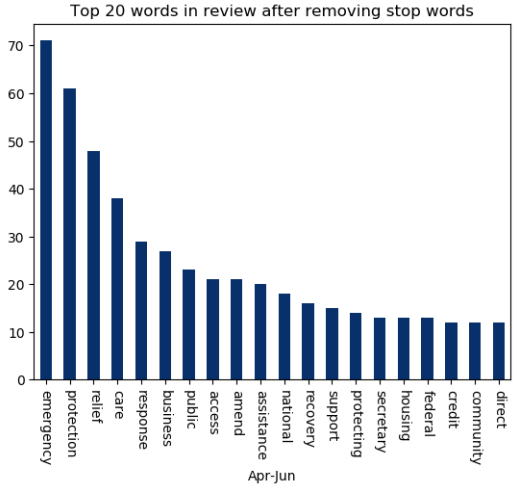
STEP 3 Output

Quarter 1 - Top 20 words



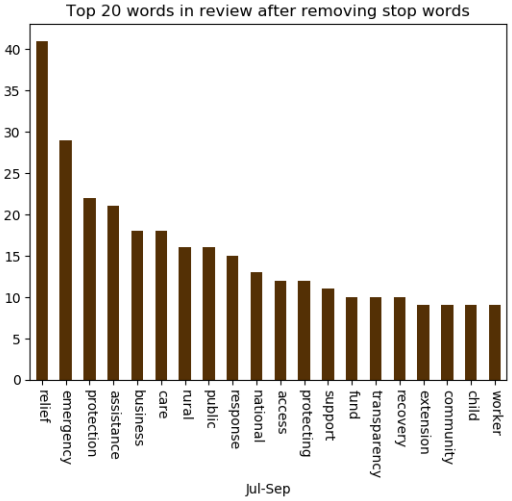
STEP 3 Output

Quarter 2 - Top 20 words



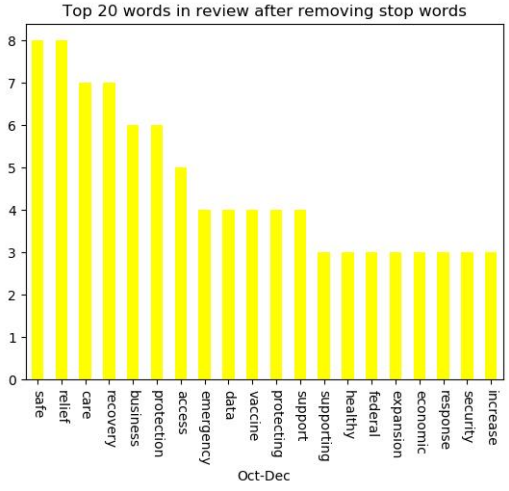
STEP 3 Output

Quarter 3 - Top 20 words



STEP 3 Output

Quarter 4 - Top 20 words



Five Steps - 4

STEP 4

In this step, we draw the word cloud graph for each quarter using the special module. In the word cloud graphs, the importance of each is shown with font size or color.

STEP 4 Output

Figure: Quarter 1

Figure: Quarter 2

STEP 4 Output

Figure: Quarter 3

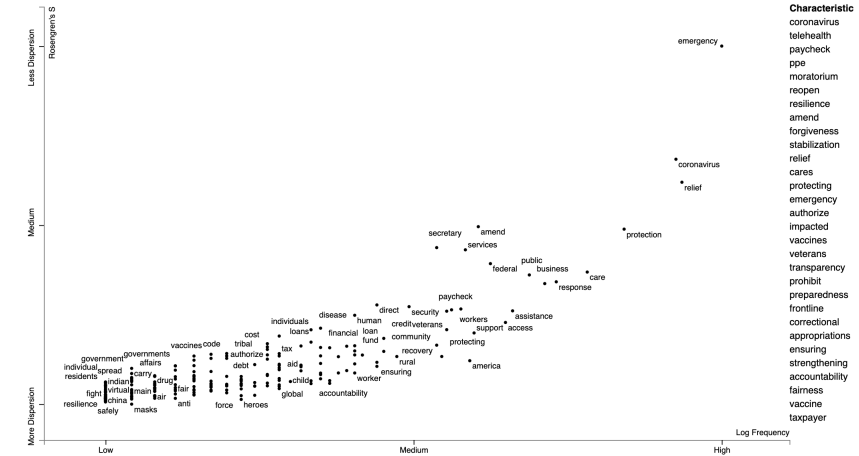
Figure: Quarter 4

Five Steps - 5

STEP 5

In this step, we draw the scatter plot to show overall keywords pattern of all bills. We want to find distinguishing terms in small-to-medium-sized corpora, and presenting them in an interactive scatter plot with non-overlapping term labels. Exploratory data analysis just got more fun.

STEP 5 Output



Search the chart

Document count: 1,590; **word count:** 6,754

STEP 5 Output

government

Document count: 1,590; **word count:** 6,754

Term: government

Frequency: 8
Range: 8
SD: 0.070754
VC: 14.062361
Juilland's D: 0.549543
Rosengren's S: 0.016057
DP: 0.982639
DP norm: 0.982639
KL-divergence: 6.097758
DA: 0.004405

Matched 8 out of 1,590 documents: NaN%

None (BILL)
direct President appoint Medical Supplies Response Coordinator coordinate efforts Federal **Government** supply distribution supplies equipment relating -.

None (BILL)
support county municipal **government** entities reducing spread - standardized testing evaluation measures, .

None (BILL)
Indian Tribal **Government** Coronavirus (-) Disaster Assistance Cost Share Relief

None (BILL)
Chinese **Government** - Accountability

None (BILL)
direct President appoint Medical Supplies Response Coordinator coordinate efforts Federal **Government** supply distribution supplies equipment relating -.

None (BILL)
permit, - emergency, Federal financial regulators temporary waiver requirements , territory, local **government** matching cost-sharing funds receiving grant Federal financial regulator, reprogramming funds support unemployment, childcare, healthcare programs, .

None (BILL)
Coronavirus Relief Fund Flexibility Local **Government**

None (BILL)
Defense, Commerce, Justice, Science, Energy Water Development, Financial Services **Government,** Labor, Human Services, Education, Transportation, Housing, Urban Development Appropriations ,

The End