Natural Language Processing and Geographic Data Visualization: A Two-Part Analysis

Problem 1: NLP for text analysis

Introduction

Natural Language Processing (NLP) allows us to process text data correctly so that we can get useful information. In this issue, I am processing the list of BBC news headlines using NLP processes like word frequency, text pre-processing, and term frequency-inverse document frequency (TF-IDF) scores you are giving. I am also showing keywords using a wordcloud.

Methodology

- Data Pre-loading and Pre-processing
- CSV file was pre-loaded in R.
- Column headers were pre-renamed to understand better and named text and category.
- The text corpus was pre-loaded using the tm package.

Text Cleaning

- I converted all the text to lowercase in order to normalize it.
- All the punctuations and numbers were deleted.
- Whitespaces were deleted to keep the dataset clean.
- Word Frequency Analysis
- Built a Document-Term Matrix (DTM) to look for word frequencies.
- Calculated word frequencies and built the top 10 frequent words.

TF-IDF Calculation

- Performed TF-IDF weighting of the DTM.

- Built the top 10 most significant words by TF-IDF score, finding significant words in individual documents.

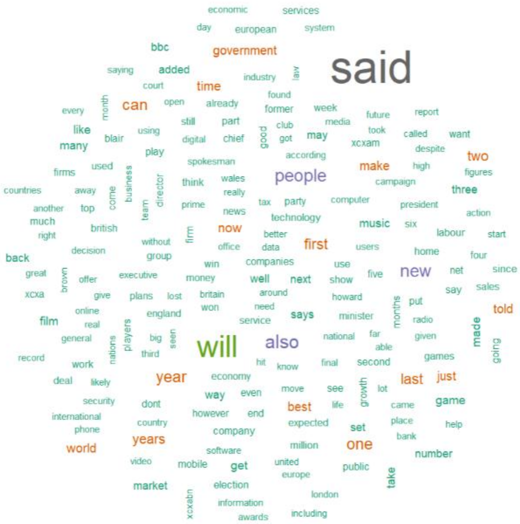
Visualization

- Built a word cloud where words are more frequent words as they are more significant.
- I used the RColorBrewer package for pretty color palettes.

Results

- Top 10 most common words provided detail on the top most common topics in the data set.
- TF-IDF also provided the most common words per article individually.
- Word cloud presented the most common words in visual form. Hence, it was easy to identify the top topics.

Graph



Problem 2

Intro

The COVID-19 pandemic struck the world severely in 2020. To visualize its spread more vividly, US county-level COVID-19 cases were visualized here. A Choropleth map was prepared to identify major areas and see how the epidemic spread.

Methodology

Data Preparation and Cleaning

- Used imported COVID-19 case data from CSV file.
- Normalized dates with lubridate package.
- Filtered out duplicate rows by requesting data to be grouped by county, state, and date.
- Rolled up 2020 cases by county.
- County-Level Case Aggregation
- Concatenated case totals by county.
- Normalized county names to lowercase ASCII for consistency.

Geospatial Data Processing

- Loaded United States county boundary data from a GeoJSON file.
- Cleaned and normalized county names to COVID-19 standards.
- Imputed missing case values with zero and capped cases at 500,000 to eliminate outliers.
- Dropped non-continental states (i.e., Hawaii, Puerto Rico) for concentrated analysis.

Choropleth Map Construction

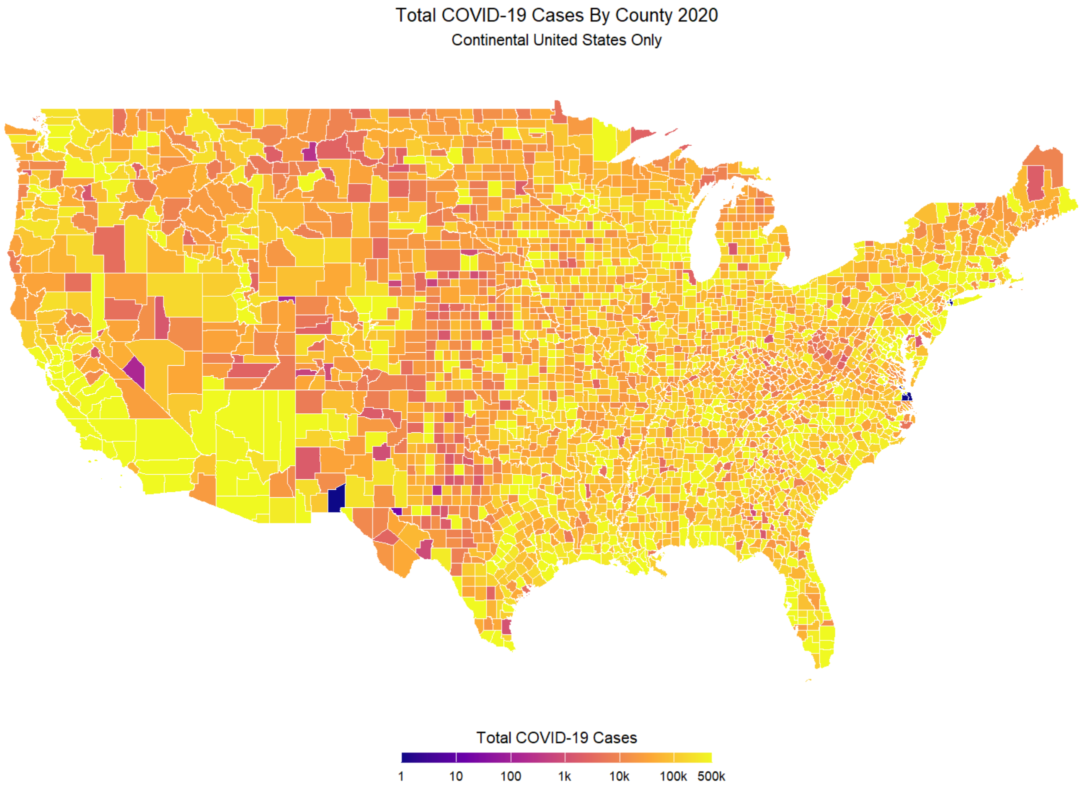
- Used sf and ggplot2 packages to create the map.

- Superimposed onto a log-scaled Viridis color map to observe case density variation.
- Used transparent color legends and titles.

Results

- The map indicated COVID-19 transmission across US counties.
- Regions of high cases were clearly distinguished from those of low cases.
- Log scaling brought the observation into normalization such that even slight changes were observed.
- Virelfare color map was readable, and trends were apparent.

Graph


Total COVID-19 Cases By County 2020
Continental United States Only
Total COVID-19 Cases
1
10
100
1k
10k
100k
500k

Conclusion

This paper presents the use of geospatial visualization in epidemiology. Choropleth Map constructed an open data-driven COVID-19 spatial pattern visualization and exposed it to scientists and policymakers alike. By amalgamating visualization, geospatial, and data, the process could enhance public health domain awareness and decision-making.