Collecting and Organizing Far-Left Extremist Data from Unstructured Internet Sources

Eric Perez, Dr. Gina Ligon
The Center for Collaboration Science Research Associate

University of Nebraska Omaha
College of Information Science & Technology

Abstract

Far-left extremism refers to a network of groups who adhere to and take direct action in accordance with one or more of the following ideas: Support for bio-centric diversity, the belief that the earth and animals are in immediate danger, and the view that the government and other parts of society are responsible for this danger and incapable/unwilling to fix the crisis and preserve the American wilderness (Chermak, Freilich, Duran, & Parkin). Far-left extremism groups self-report activities using publicly accessible, online communiqués. These activities include arson, property damage, harassment, sabotage, and theft (Loadenthal). The communiqués are structured like blog posts with no sorting or search feature, making it difficult analyze the data. A Python web-crawler was built to collect data from the communiqués and store in a database, identifying title, date, country, URL, and contents. Using the structure provided by the database, I found frequency of words, countries, and posts over time. The database provided a way to do a keyword search, allowing the quick identification of posts related to arson. There were 3,010 communiqués collected. Of those, 397 were identified as arson. The most frequent words used in communiqués were animal, will, and ALF. The countries with the most posted were the UK, Sweden, and Mexico. The number of postings peaked in 2008 and has declined since. In conclusion, the web crawler was an effective way to store communiqué data to make future research into the topic easier, as shown by the analysis already able to be completed.

Research Questions

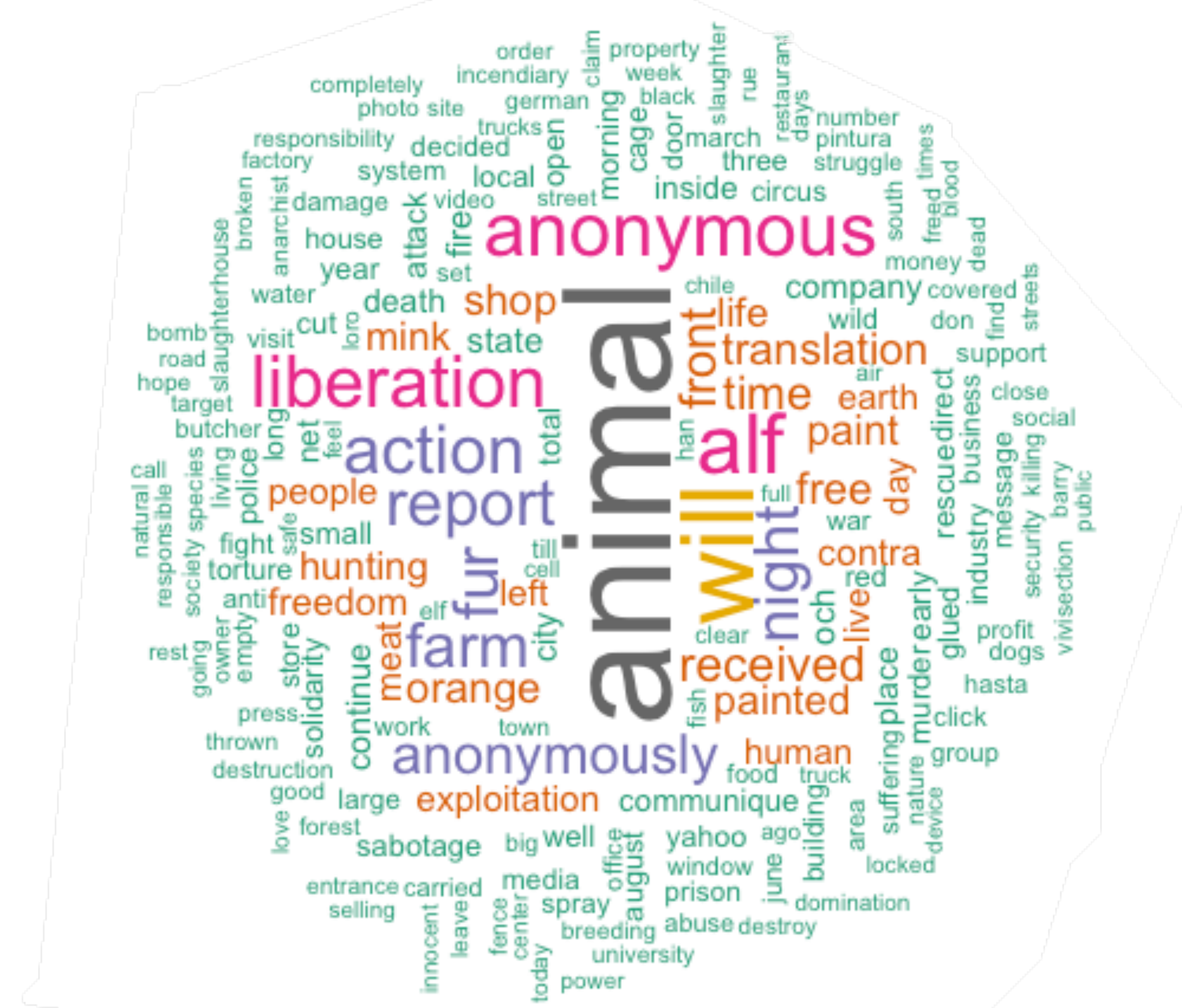
1. To what extent can far-left unstructured communique data be collected and stored?
2. What words are most frequently used in communiqué data?
3. How frequently do different countries appear within the different communiqués?
4. How frequently are communiqués posted over time?

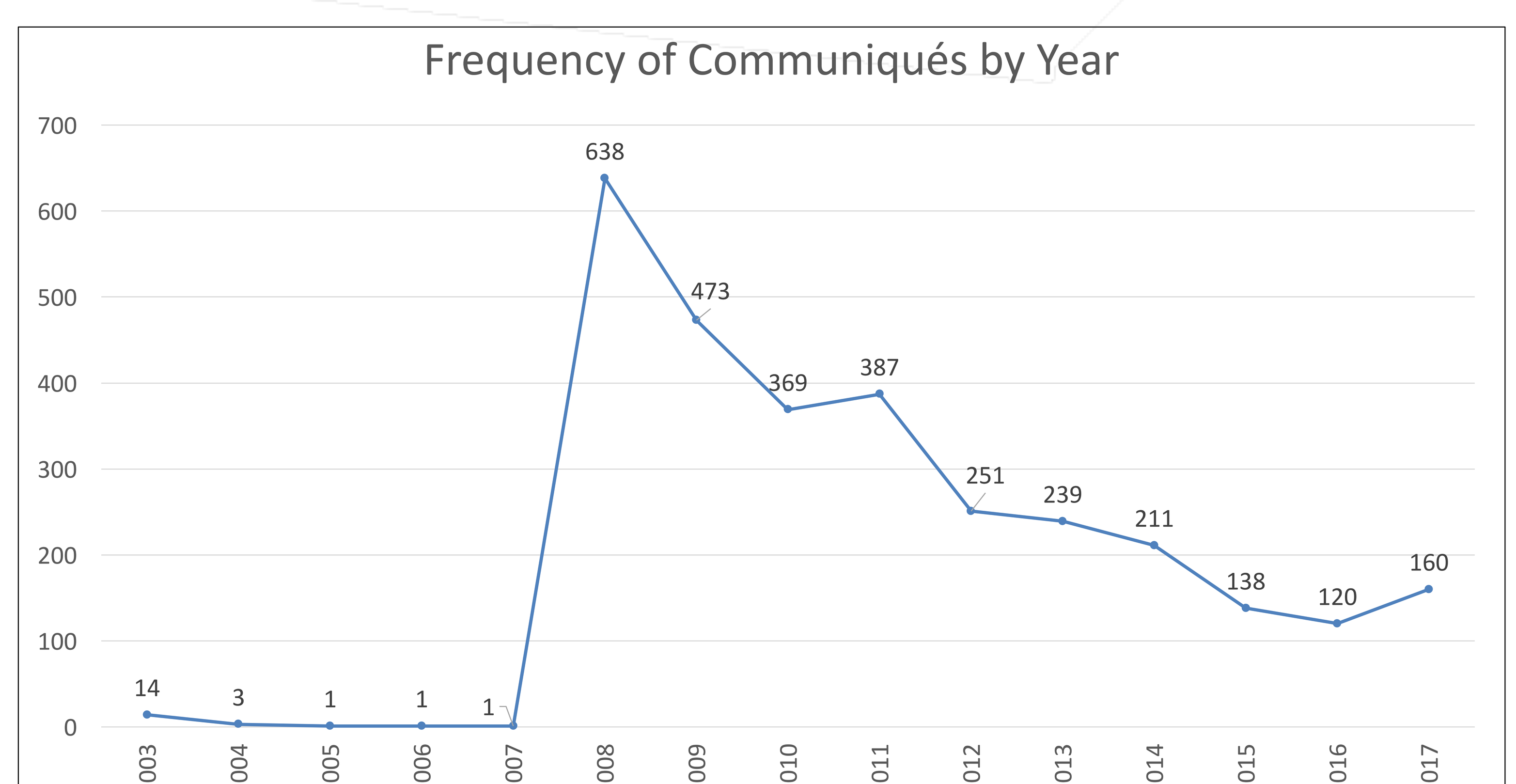
Methodology

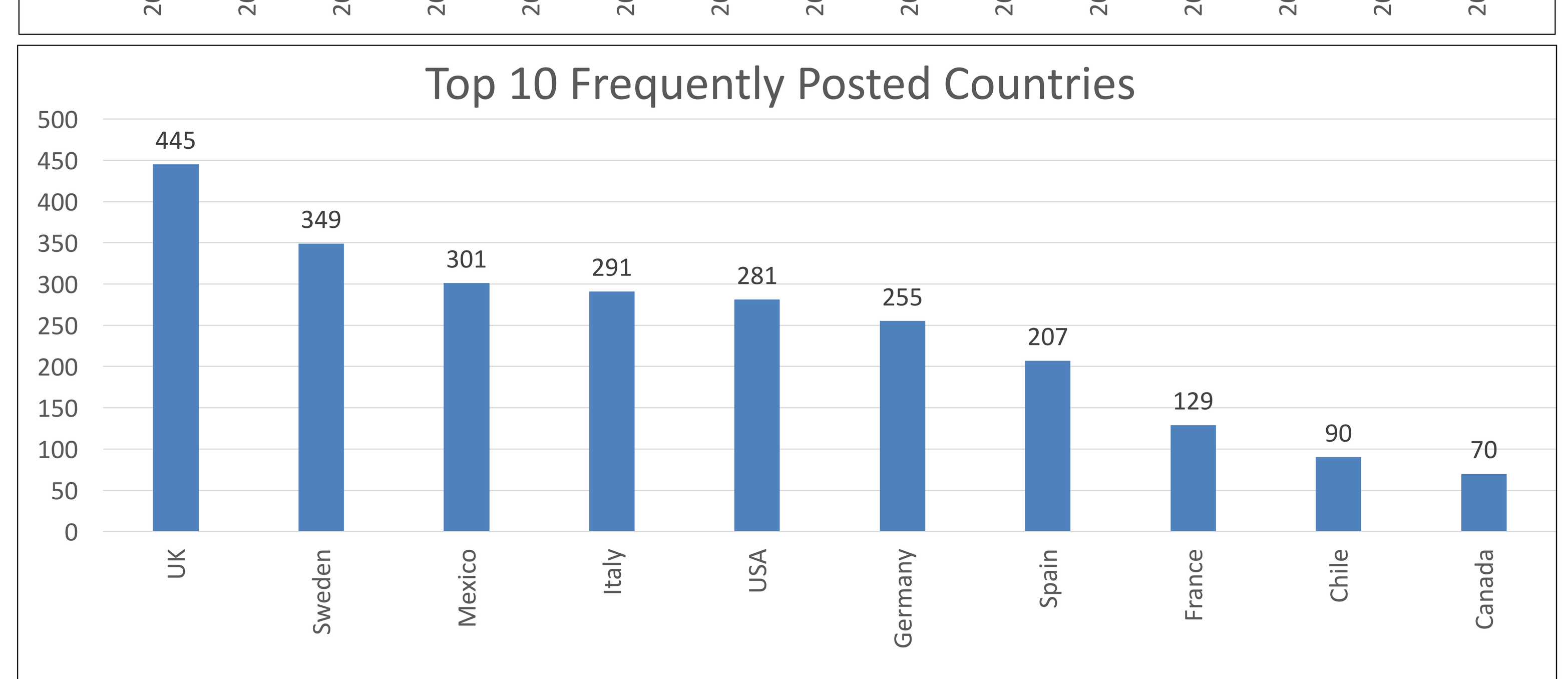
1. Collaborated with doctoral student in criminology to identify websites for communiqués.
2. Wrote a Python web crawler to save communiqués from the website into a MySQL database. The web crawler identified communiqué titles, dates, countries, URLs, and contents.
3. Wrote a Python script to identify arson posts based on the following key words: arson, fire, incendiary, gasoline, petro, and petroleum.
4. Used R to find word frequencies and create a word cloud for the communiqué contents.
5. Used Excel pivot tables to identify most frequent countries and graph posts over time.

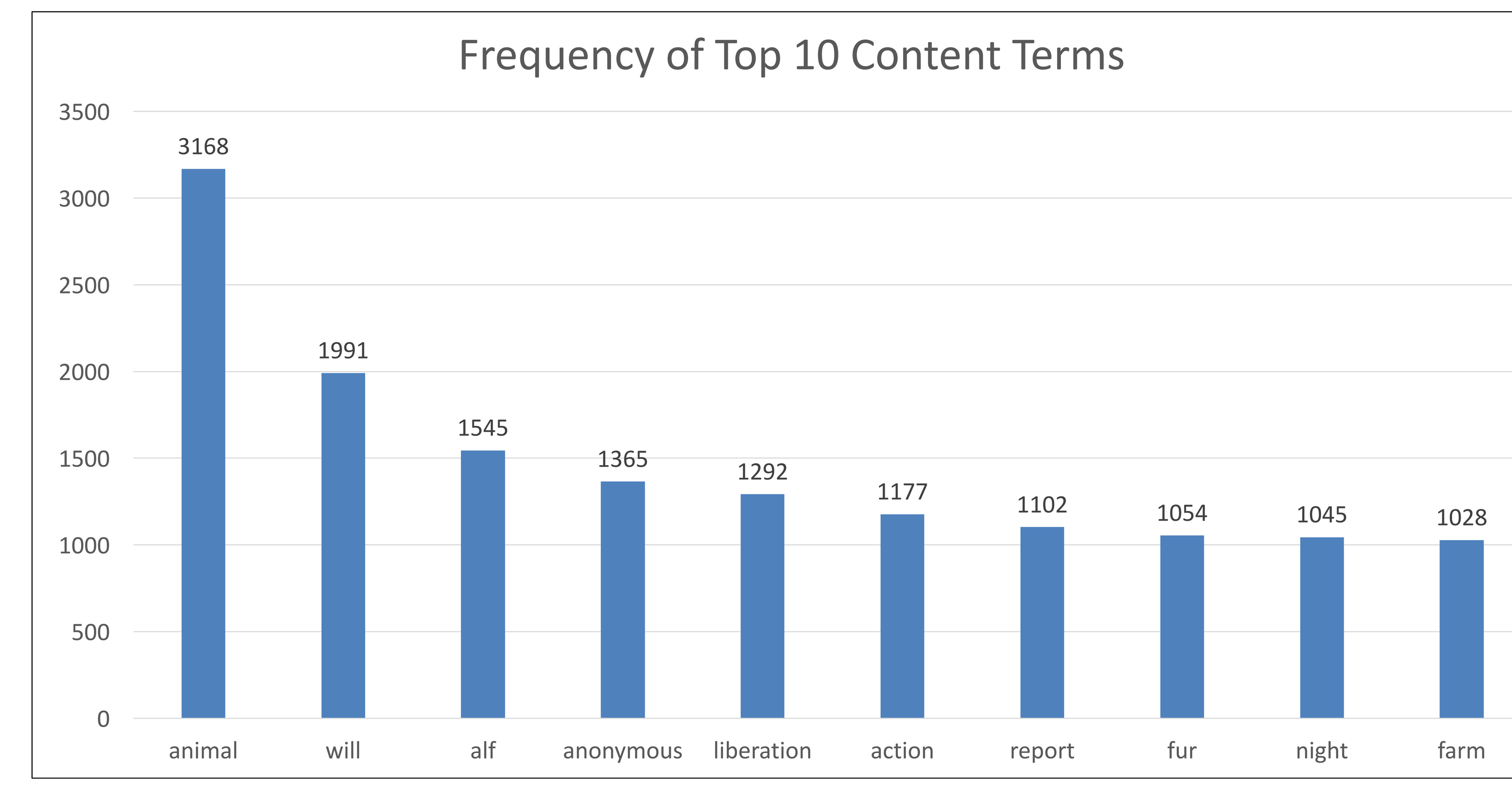
Results

Word Cloud of Communiqué Contents



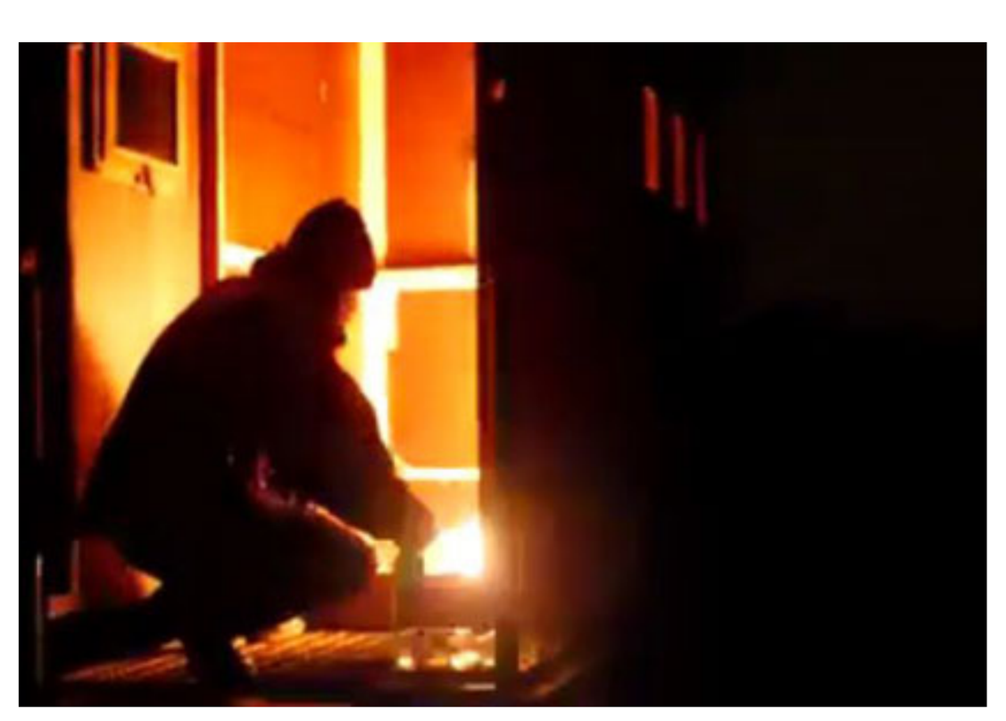






Communiqué Example

Title: *Arson Against Hunting Towers*



Date: *January 29, 2018*

Country: *Germany*

Contents: *We discovered that the sand and stone exploitation by the forest is allowing hunters to create meadows for shooting stands, attracting wildlife to the area to then murder animals without compassion. We destroyed dozens before deciding to send a clearer message, so overnight and with only one security guard in the area we set multiple shooting huts and towers on fire, making sure that all were standing in the clear ground without risks of creating an uncontrollable fire.*

Future Research

1. Build additional web crawlers to collect data from different sites related to far left extremist communiqués.
2. Use Machine Learning to automatically identify activity types as it is collected by the web crawler.

Acknowledgments

This work was supported by the Fund for Undergraduate Scholarly Experience (FUSE). I would also like to thank the Center for Collaboration Science and the MISL lab, specifically Gina Ligon, Michael Logan, and Laramie Sproles for their help with various aspects of this research project.