EU Privacy News Dataset

Akhil Acharya & Robert Tapp

Project Overview

Goal: Build Corpus using an API for EU Region using Guardian API

Focus: Privacy news about government agencies regarding surveillance or regulation

Research Questions

- RQ1: What privacy-relevant keywords are most effective in finding relevant news articles?
- RQ2: Under which circumstances (context) are most privacy related news articles published?
- RQ3: What are common topics of privacy related news articles?
- RQ4: What is the common sentiment of privacy related news articles?

Approach

- Step 1: Corpus Extraction
  - Determine keyword/search queries relating to EU government agencies
  - Perform keyword search on Guardian web API
  - Download and archive text and metadata
- Step 2: Summarization
  - Find frequency of article publication
  - Topics of news articles
  - Sentiment of news articles
  - Find Clusters

Approach: Corpus Extraction

1. Enumerated all intelligence agencies for all EU member states with privacy-relevant terms ("surveillance", "disclosure")
2. Created boolean search queries to remove US related news as best possible
3. Extract articles for each query in parallel

Query Example:

```
[{
   "name": "SURVEILLANCE AND LEAKS",
   "keywords": ["SURVEILLANCE", "LEAKS"],
   "search_term": "SURVEILLANCE AND LEAKS NOT (nsa OR cia OR fbi OR cia OR usa)"
}, {
   "name": "HACKING AND SURVEILLANCE AND DISCLOSURE",
   "keywords": ["HACKING", "SURVEILLANCE", "DISCLOSURE"],
   "search_term": "(HACKING AND SURVEILLANCE AND DISCLOSURE) NOT (nsa OR cia OR fbi OR cia OR usa)"
}, {
   "name": "GCHQ AND SURVEILLANCE",
   "keywords": ["GCHQ", "SURVEILLANCE"],
   "search_term": "GCHQ AND SURVEILLANCE NOT (nsa OR cia OR fbi OR cia OR usa)"
}, ...]
```

Results

- Generated 60 queries
- Roughly 4,400 articles per query (rate limits)
- Total Articles: **257,512**
- Total corpus: **1.35** gigabytes

Approach: Analysis

1. Relevance filtering
2. Topic Modeling (LDA)
3. Word Cloud visualizations
4. Date clustering
5. Text clustering
6. Sentiment analysis

Relevance Filtering (RQ1)

- Performed second pass keyword search for > 1 match
- Reasoning: Search within a wide net

Interesting Results:

- Most relevant: 15% remain, GCHQ and surveillance
- Least relevant: 0.6% remain, Slovakia and surveillance
- Overall: ***2.1%*** remain, **5659 remaining overall**

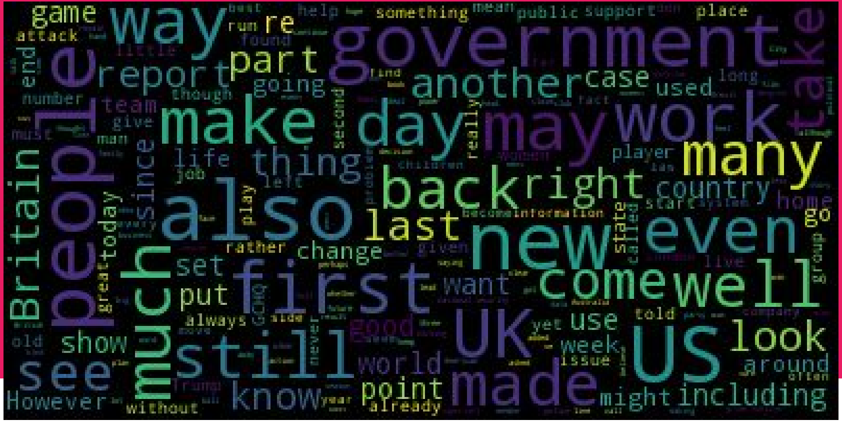
Topic Modeling (RQ2, RQ3)

Used **gensim** to do LDA topic modeling on the filtered articles

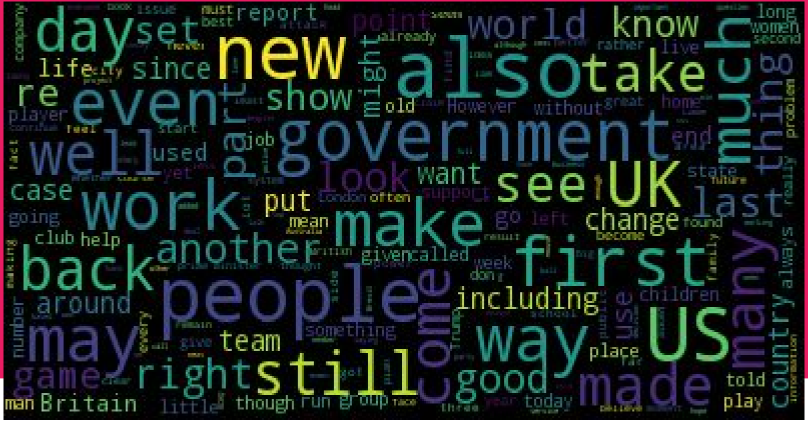
Interesting Results:

UK AND SURVEILLANCE	intelligence, surveillance, security, gchq, government
BELGIUM AND SURVEILLANCE	attack, police, people, like, intelligence

Word Cloud



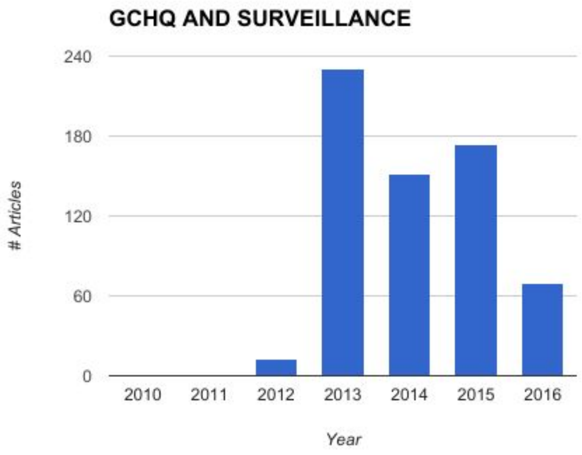
GCHQ AND SURVEILLANCE

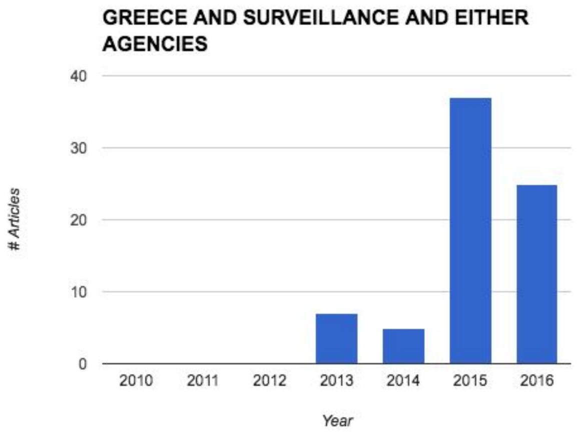


GCHQ AND PRIVACY

Date Clustering

- Found frequencies of article publication throughout the dataset over the last 20 years (1996-2016)





Incidents:

- 2011: Czech Republic AND Privacy
  - Google Street View blocked in Czech Republic due to privacy concerns
- 2011: Hacking AND Disclosure
  - News Corp phone hacking scandal
- 2012: GCHQ AND Surveillance
  - Controversy over expanded GCHQ powers
- 2013: GCHQ AND Surveillance
  - Snowden revelations - "Five Eyes"
  - Largest spike in all categories

Incidents

- 2014: Greece AND Surveillance
  - Surveillance vessel found off coast of Cyprus
- 2015: GCHQ and Surveillance
  - Controversy over "Snooper's charter", also allowed expanded GCHQ powers

Text Clustering

- K-means clustering with 30 clusters
- Size of clusters had a high variance
- Better as a filter, than a grouping measure

Interesting Results:

- GCHQ Cluster
  - Monitoring MP’s (2 articles)
  - Firewall (3 articles)
- Surveillance Cluster
  - EU decision on UK surveillance laws (3 articles)

Sentiment Analysis (RQ4)

- Utilize VADER heuristics to gauge sentiment
- Fast inference without training

Interesting Results:

- Sentiment generally neutral (~88%)
- Most positive: Portugal and Privacy (11% +)
- Most negative: France and Surveillance (10% -)

Limitations

- Datasource coverage biased towards U.K. issues/topics
  - Problems w/ other data sources - language barrier, no API access
- Concerns about rate limits restricted articles per query at ~5,000, though the limit appears to be flexible
- Some "cross contamination" by nature of text search
  - Germany queries have articles about London terror attacks because of German comment
- Filtering is naive - could lead to false negatives

Lessons Learned, and Future Steps

- Better filtering
  - Build a classifier?
- More narrow search queries - false negatives
- Fetch more articles per query
  - Limit set to 4400, can be expanded

Questions?