R News Digest

Devis Shehu (ds1902)

Goals & Motivations

- World is a complex and confusing place and we're inundated with (often conflicting) information
- Most workplaces are drowning in documentation and user created content; hard to read through, index, summarize, analyze critically and keep track
- Even highly-organized people using proven note taking techniques such as Zettelkasten struggle with the volume & speed; an agent would be best but first the basics
- Build a fun little project analyzing news articles to reinforce learnings from class and think through ideas
- Dataset should be non-trivial: collect and process a moderate number of news articles and deal with real world-complexities of dirty and incomplete data
- Collect stats and do some analysis to characterize the data
    - News is easier since most of us have "context"
- 1st step toward auto-building organized summaries (Wiki page on Ukraine war) without the curation work: typically low value to others unless domain expert
- Ultimate step: feed sanitized corpus and documents into GPT-3 or alternative - i.e. agent reads and writes summaries, human analyzes and verifies

2022 Russian invasion of Ukraine

From Wikipedia, the free encyclopedia

Russia invaded Ukraine on 24 February 2022, marking a steep escalation of the Russo-Ukrainian War, which began in 2014 following the Ukrainian Revolution of Dignity. The invasion has caused Europe's largest refugee crisis since World War II,[17][18] with more than 5.5 million Ukrainians leaving the country[18] and a quarter of the population displaced.[20][21]

At the start of the war in 2014, Russia annexed Crimea, and Russian-backed separatists seized part of the south-eastern Donbas region of Ukraine, sparking a regional war there.[22][23] In 2021, Russia began a large military build-up along its border with Ukraine, amassing up to 190,000 troops along with their equipment. In a televised address shortly before the invasion, Russian president Vladimir Putin espoused irredentist views,[24] questioned Ukraine's right to statehood,[25][26] and falsely[27] accused Ukraine of being governed by neo-Nazis who persecute the ethnic Russian minority.[28] Putin also said the North Atlantic Treaty Organization (NATO) constitutes a threat to Russia's national security by expanding eastward since the early 2000s, which NATO disputed.[29] Russia demanded NATO cease expansion and bar Ukraine from ever joining the alliance permanently.[30] The United States and others accused Russia of planning to attack or invade Ukraine, which Russian officials repeatedly denied as late as 23 February 2022.[34]

On 21 February 2022, Russia recognised the Donetsk People's Republic and the Luhansk People's Republic, two self-proclaimed statelets in Donbas controlled by pro-Russian separatists.[35] The following day, the Federation Council of Russia authorised the use of military force abroad, and Russian troops overtly entered both territories.[36] The invasion began on the morning of 24 February,[37] when Putin announced a "special military operation" to "demilitarise and denazify" Ukraine.[38][39] Minutes later, missiles and airstrikes hit across Ukraine, including the capital Kyiv, shortly followed by a large ground invasion from multiple directions.[40][41] In response, Ukrainian president Volodymyr Zelenskyy enacted martial law and general mobilisation of all male Ukrainian citizens for between the ages of 18 and 60, who were banned from leaving the country.[42][43]

At the start of the invasion on 24 February, the northern front was launched out of Belarus and targeting Kyiv with a northeastern front launched at the city of Kharkiv; the southeastern front was conducted as two separate spearhead fronts including a southern front (originating in Crimea) and a separate probative southeastern front (launched at the cities of Luhansk and Donetsk).[44][45] On 8 April, the Russian ministry had announced that all its troops and divisions deployed in southeastern Ukraine would be united under General Aleksandr Dvornikov, who was placed in charge of combined military operations, including the redeployed probative fronts originally assigned to the northern front and the north-eastern front which were subsequently withdrawn and reassigned to the second phase to the southeastern front.[46] By 17 April, progress on the southeastern front appeared to be impeded by residual troops continuing to hold-out in abandoned factories in Mariupol.[47] On 19 April, Russia launched a renewed invasion front referred to as an "eastern assault" across a 300-mile front extending from Kharkiv to Donetsk and Luhansk, with simultaneous missile attacks again directed at Kyiv in the north and Lviv in western Ukraine.[48]

The invasion has been widely condemned internationally as an act of aggression.[49][50] The United Nations General

2022 Russian invasion of Ukraine
Part of the Russo-Ukrainian War

Military situation as of 2 May 2022
Ukraine
Occupied Ukrainian land
For a more detailed map, see the Russo-Ukrainian War detailed map

Date: 24 February 2022 – present (2 months, 1 week and 2 days)
Location: Ukraine[a]
Status: Ongoing (list of engagements · control of cities · timeline of events)

Belligerents

Russia Donetsk PR[a] Luhansk PR[a] Supported by: Belarus[b]	Ukraine

Commanders and leaders

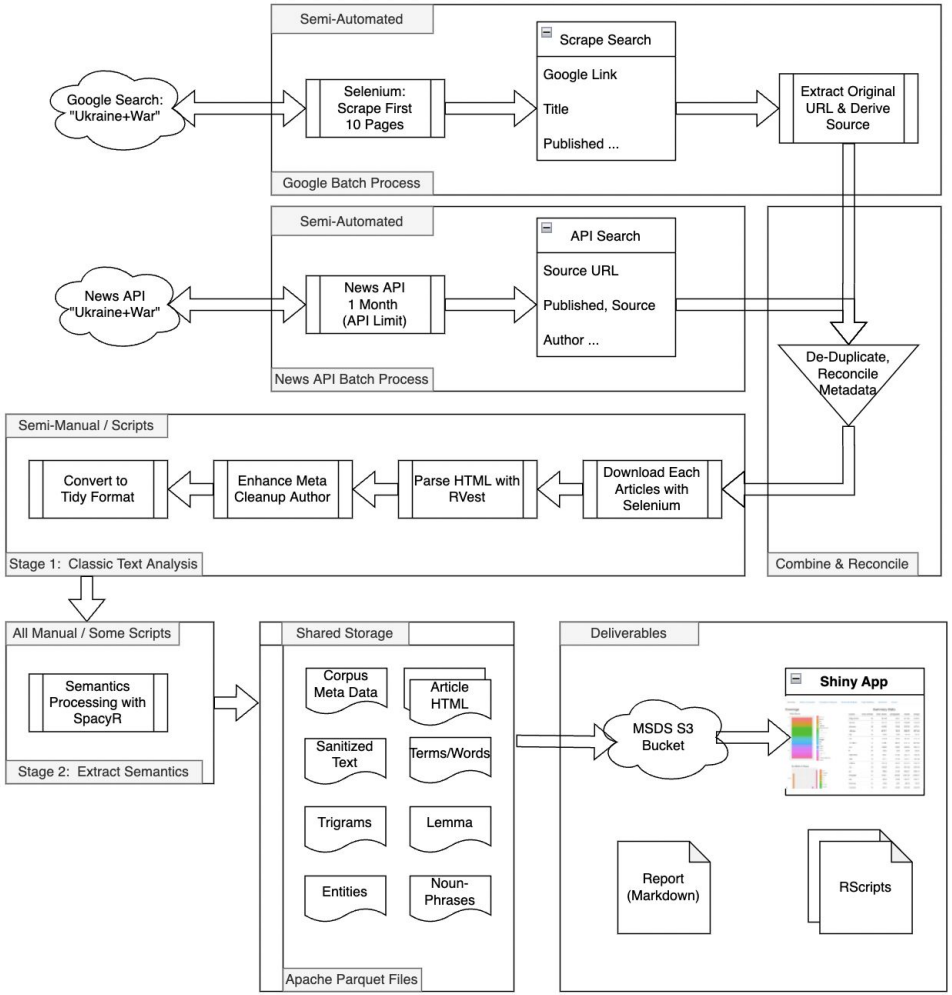
Vladimir Putin	Volodymyr Zelenskyy

Strength

Russia: ~175,000–190,000[14][15] Donetsk PR: 20,000[16] Luhansk PR:	Ukraine: 209,000 (armed forces) 102,000 (paramilitary) 900,000 (reserves)[16]

General Flow

- Identify news providers: Google search and News API were most straightforward (discarded Guardian, NY Times API due to complexity - value tradeoff)
- Run Google searches via Selenium in batches
  - Deal with Google and other sources' throttling
- Call News API iteratively due to limits on free account (1 month of free data) to accumulate data
- Collect or parse all available metadata (dates published, authors, title, url, etc.) over 2 months
- Store metadata in a separate data frame (provider)
- Download files to avoid issues with broken links, re-running process,etc.
- Parse HTML and cleanup auth, etc.
- Sanitize (stop words, etc.) pre-calculate tokens, trigrams and calculate stats for performance
- Convert to Tidy Format and extract semantics (mostly manual)
- Shuffle files to S3 so Shiny App is detached from backend data collection and processing
- Lots of try/catch, error handling for poorly formatted data, Selenium erroring out, throttling, etc.
- Robustness at scale is really hard and time consuming and did not meet the goal!

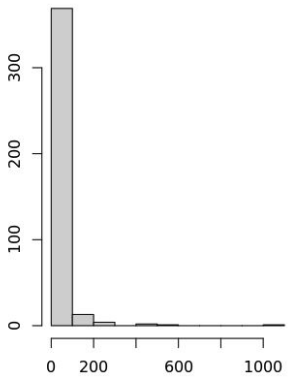


Summary Statistics

- **10,390** Articles
  - 6,996 from Google News including 1,009 attribute to Google but actually made up of random feeds, eg. youtube videos, unknown sources
  - 3,395 from News API
- Timeframe: Jan 31st to April 16th : paused data collection and cleaning
- 679 unique sources but highly skewed with a relatively small # of sources contributing most content (long tail possibly an artifact of data collection)
- News API: 533 secondary sources, attributed to other publications such as AP (262), Reuters(75), VOA (Voice of America) News, ABC, France, New York Times, Bloomberg
- Surprised that NewsAPI, an expensive service, starting at $450 month, has really unclean metadata (especially authors)
- Hundreds of articles were attributed to AP, other news organizations which makes sense but having the author would have enhanced the analysis

Summary Stats

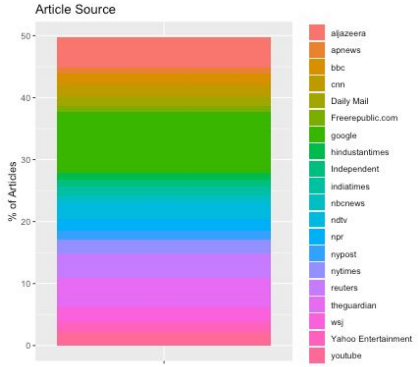
source	total_articles	total_words	paragraphs	words	unique
aljazeera	514	350722	51.80	682.34	372.78
theguardian	500	311412	27.18	622.82	356.92
reuters	408	194237	34.49	476.07	331.19
ndtv	280	71234	20.61	254.41	179.12
wsj	242	23738	10.88	98.09	70.47
cnn	199	95359	38.17	479.19	297.91
npr	196	60100	20.24	306.63	195.72
bbc	171	89608	55.36	524.02	356.10
Yahoo Entertainment	170	161822	57.69	951.89	623.66
Daily Mail	158	125653	73.33	795.27	446.21
indiatimes	151	105693	70.01	699.95	297.17
nypost	143	76473	30.59	534.78	299.19
hindustantimes	126	90834	50.41	720.90	427.82
nbcnews	122	93631	55.02	767.47	467.01
google	111	99474	67.65	896.16	475.89
Freerepublic.com	109	46459	37.14	426.23	296.54
apnews	103	52117	29.03	505.99	337.26
Independent	103	47618	55.07	462.31	276.09
abc	101	89355	89.34	884.70	543.14
newsweek	94	168648	122.80	1794.13	767.54

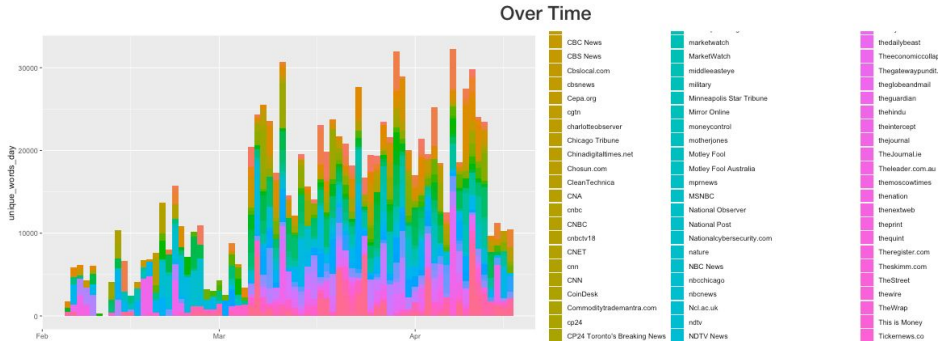


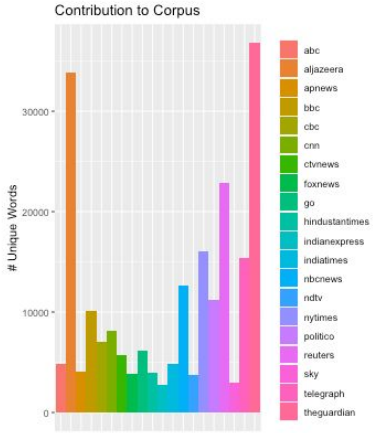
AP	58
Feed / Feeder	57
Facebook (BBC)	53
AP Wire	25
Post Editorial Board	20
ABC Newshour	19
Socialists & Democrats in EU Parliament	18
Alex Millson	14
Staff	13
Angela Dennis	12
Caroline mimbs nyce	11
Press operations	11
SA Transcripts	11
Eustance Huang	10
LW	9
Star Tribune Staff	9
AFP	8
Bob Brigham	8
Deutsche Welle	8

Coverage by Source

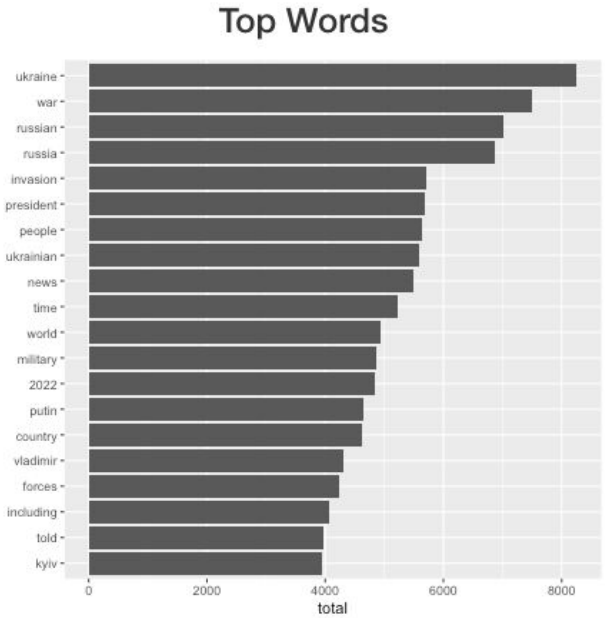
- Given the distribution of sources, wanted to figure which one provides the most information (# of article, unique content)
- Long-term goal to identify statements/facts, cross reference and flag unique content (super valuable)
- Starting out with basic stats as shown in previous slide
- Top 20 sources contributing 50% of the content and are generally the most reliable sources (i.e. **not** social media/alt facts)
- Contribution varies dramatically over time (even NY Times, Guardian and Al Jazeera) - event driven - invasion, major battles, developments such as new sanctions, announcements by Russian, US and British government
- Pleasant surprises: coverage of side stories such as refugees from Africa and India from alternative sources such as the Times of India; less of a western bias
- Conflict is still in the early phase and it would be interesting to see if intensity (and coverage) is sustained in 6 months



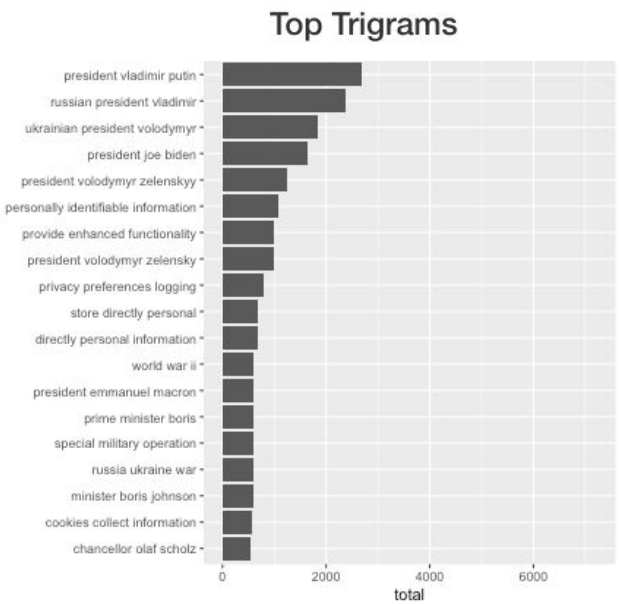




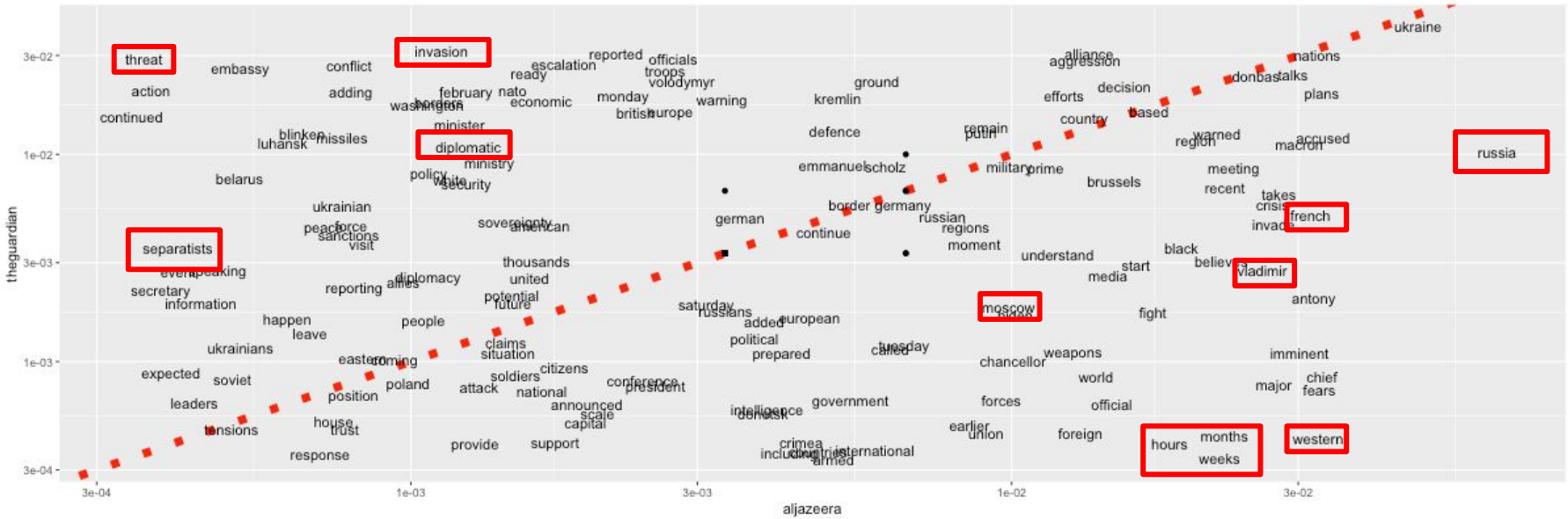
Terms & Concepts



- Overall most articles have a high % of unique words which means the proportion of top words is a low % of the total set of words in the corpus
- Sparsity of words is an issues when computing frequency, etc.
- Terms and trigrams are as expected, mostly bubbling up the key players (Putin, Boris Johnson, Biden) and concept
- “World War II”I, which seem to contribute to fear sentiments pops-up relatively high on the list
- Some data oddities such as “personally identifiable information” and “cookies collect information” due to “boilerplate” web content. not filtered in parsing (improvement)



Comparative Analysis



Comparative Analysis

- From the word comparison and unique terms, some indication that the content of the **The Guardian** differs from **Al Jazeera**
- Both are highly credible and well regarded as sources though The Guardian has a reputation as being somewhat left-leaning and biased toward Britain, and its role and leadership in the conflict
- Some evidences of emphasis on high level diplomacy and threat posed by Ukraine conflict to European, Britain's collective security and action
- Al Jazeera seems to have a more balanced view of the conflict that emphasis broader context; i.e. West's ambiguous relationship with Russia, potentially culpability in the runup to the conflict
- Al Jazeera seems more empirical, objective and focused on the dates and facts but also at times seems more tolerant of Russia
- Unrelated commercial terms (ads) and browser/cookie (boilerplate notice) - small glitch with "belarus"
- Vladimir / Putin is conspicuously prevalent on Al Jazeera's along with French diplomatic efforts and Germany's role (especially early on hesitation to break with Russia and send weapons)

word
guardian
ability
abruptly
access
adviser
analysis
appeared
artillery
attempt
believed
benefit
biggest
billionaire
boris
britain
casualties
chancellor
changing
claimed
claims
clock
collapse
commercial
communities

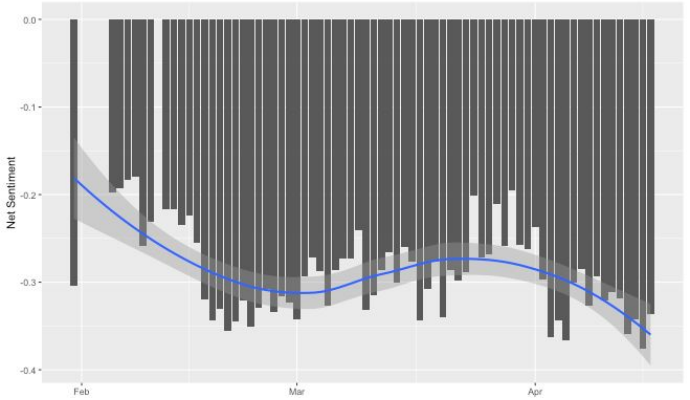
Guardian Only

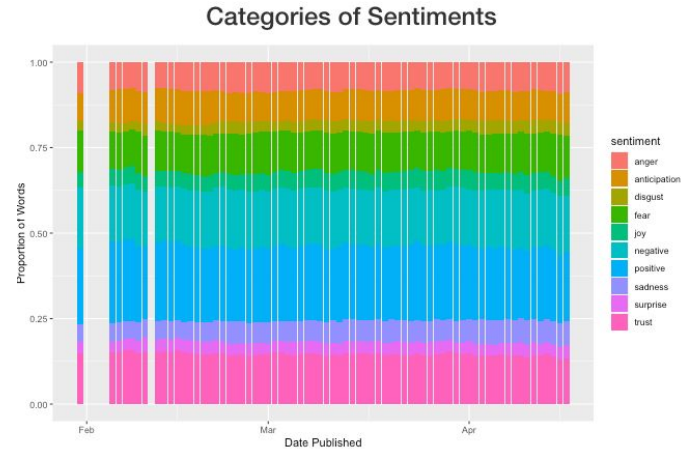
word
jazeera
administration
advertisers
advertising
aggregated
alliance
anonymized
assault
assistance
belarus
black
block
borders
browser
build
business
ceasefire
central
change
china
citizens
collection
concerns
confirmed
consent

Al Jazeera Only

Sentiment Analysis

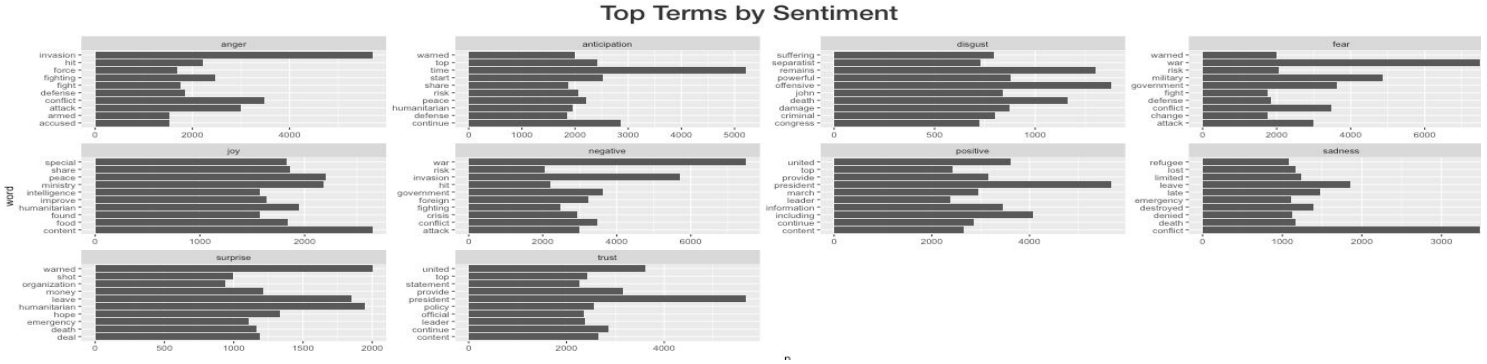
- Sentiment analysis using **bing** and **nrc** to get both a positive/negative and some categorization
- As expected sentiment started out strongly negative and declines precipitously over time as the reality of the war settles in as well as realization that it's not likely to end
- Anticipation was highest at the beginning of the conflict both in terms of if invasion was happening and later on the (misplaced) hope for a peaceful (short term upswing)
    - Less pronounced when looking at 2 ½ months vs 2 week sample
- Expected anger and disgust to increase relative to other emotions as evidence of war crimes surfaced, possibly the sentiment is diluted by sheer number of sources?
- Sadness seems to be increasing but not as much as expected
- Some evidence in the data of fear increasing marginally as terms such as WW3 and nuclear conflict started occurring more frequently
- Sentiment analysis might have been more insightful if run on top 20 news sources and with pruning (separate positive and negative)
- Can we measure fatigues (apathy) somehow?





Sentiment Analysis - Emotions & Terms

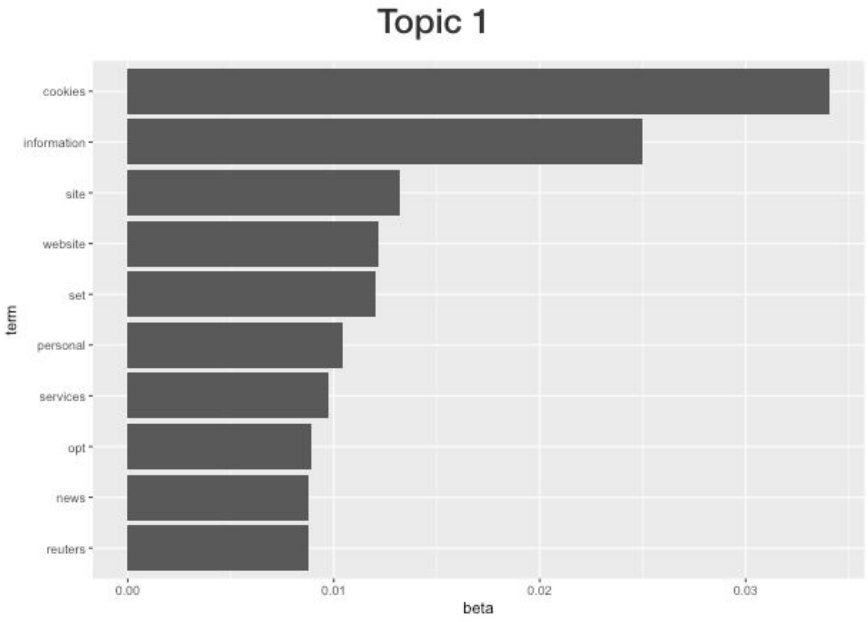
- Given the content, the associated terms aren't too surprising: i.e. invasion, force, fighting associates with anger/fear
- Likewise for refugee, death, emergency with sadness
- Also as expected, temporal (start, continue, time) associate with anticipation: humanitarian (corridors) and peace (conference) do bubble up which tracks with trends
- Some associations are obviously irrelevant ("intelligence" with joy?) - limits of Sentiment Analysis?
- News coverage emphasis the "unexpected" (surprise) of the breadth and depth of humanitarian assistance - is this unexpected or news accentuating the positives and negatives (extremes)?
- Hard to gauge <u>trust</u>: is there significant trust in policy, leaders or even president?
- Sentiment analysis is somewhat limited expect when it bubbles up an unexpected result

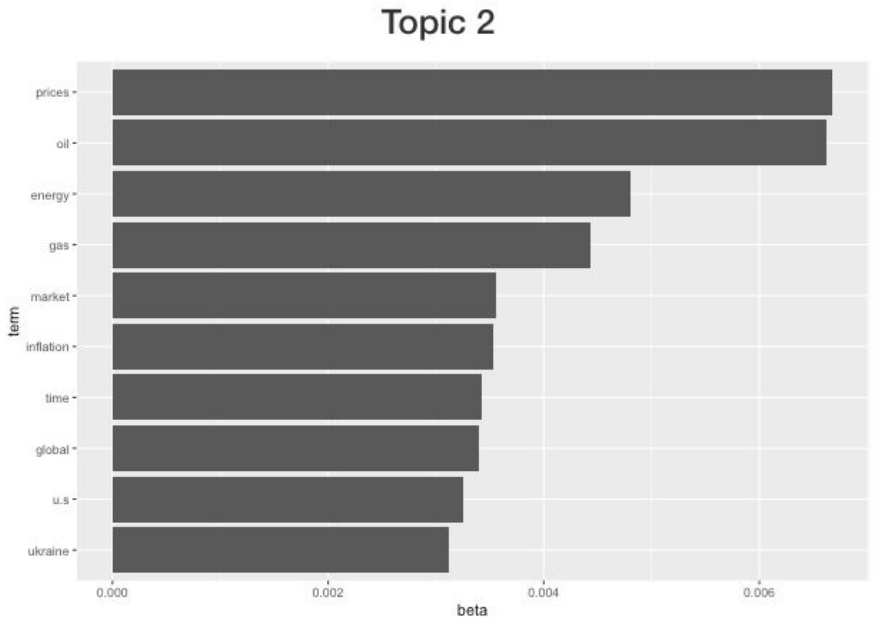


Topic Modeling & Linear Discriminant Analysis

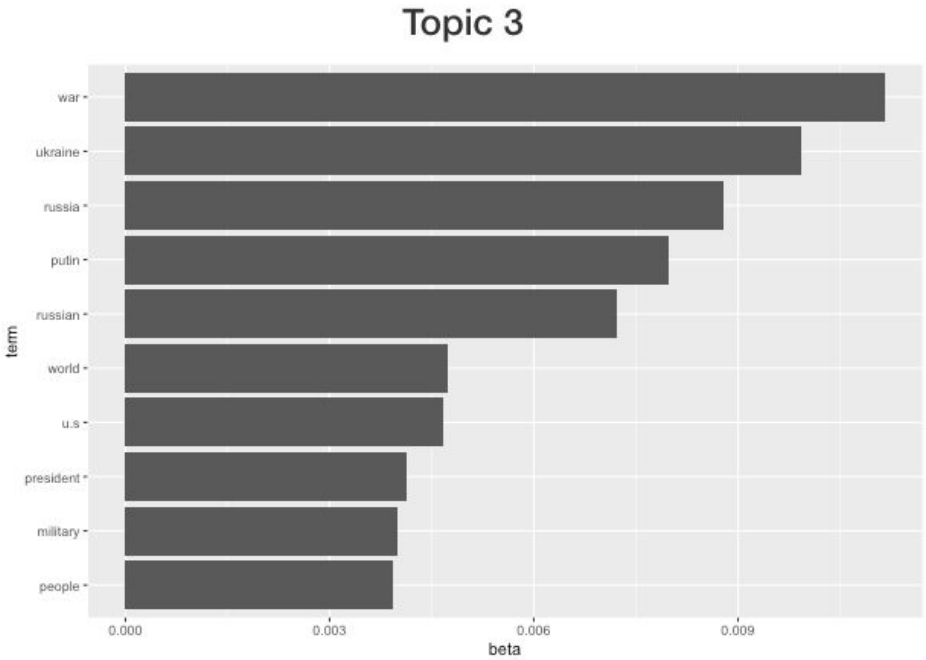
- Ran LDA with default parameters and K=8 and Fisher's method: derived experimentally (trial and error)
  - Needed to convert tidy text with terms sparse matrix
  - *dfMatrix <- dfNewsTidyTerms %>% cast_dtm(document, term, count)*
- Goal was to see if LDA could extract the dimensions of the conflict (political, military, humanitarian, financial (natural resources), etc.
  - Ideally combine this with sources to tag and differentiate sources according to dimenion
- Need to explore coherence score to determine the right number of K and it would help to narrow down the sources to the most trustworthy, highest regarded (see earlier comments)
- Potentially use topic modeling to remove irrelevant search terms, trigrams (advertisement, browser/cookie, etc.)
- Potentially categorize and subdivide the Corpus into subsets and analyze each (drill-down) - automation?
- Some duplication of common words (Ukraine, Russia) is difficult to escape as the words have different context (interesting problem)

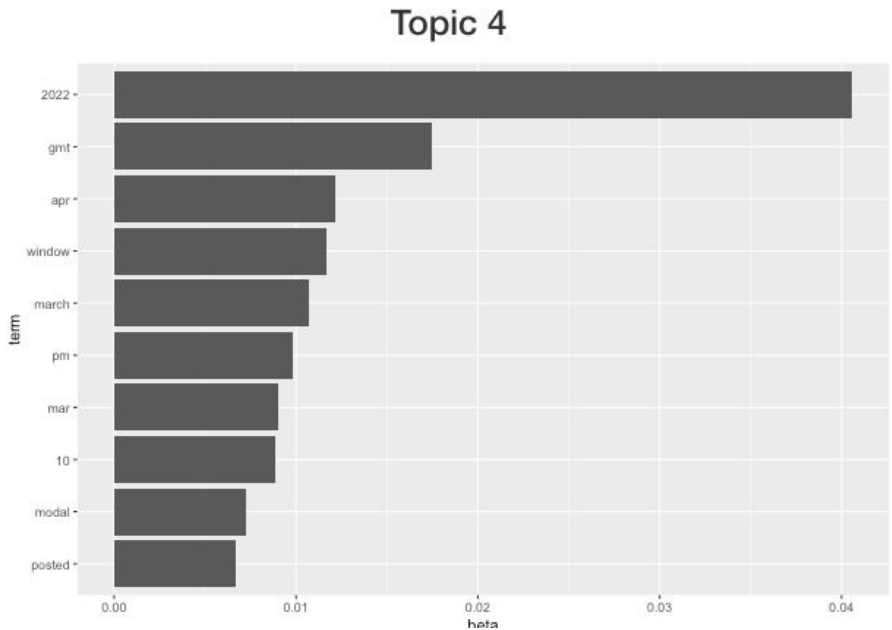
Topic Modeling Results - Part 1



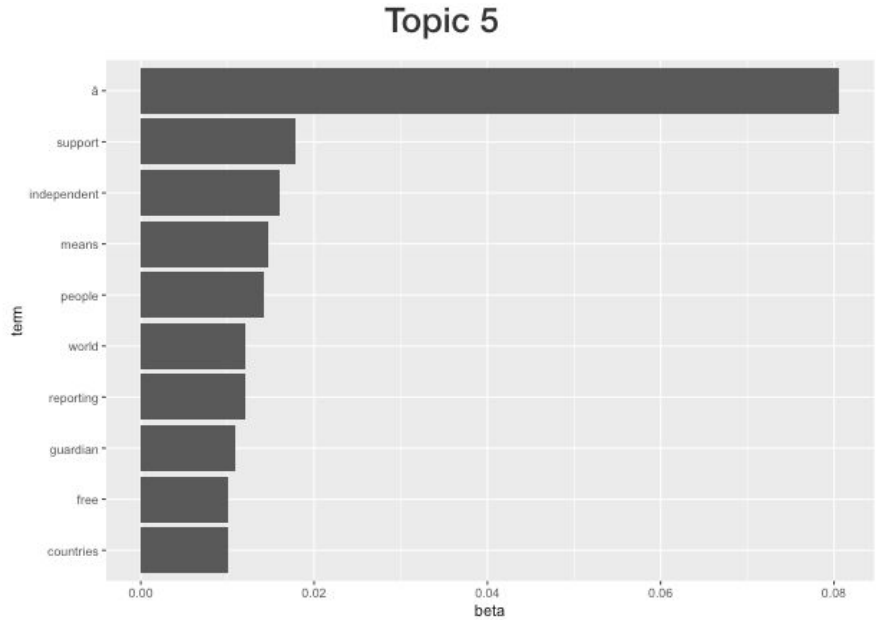


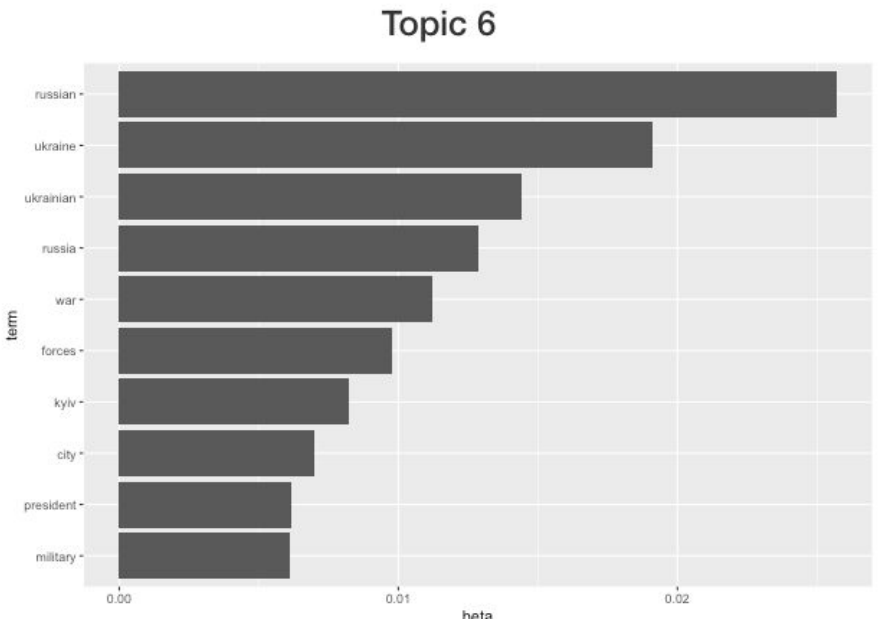
Topic Modeling Results - Part 2



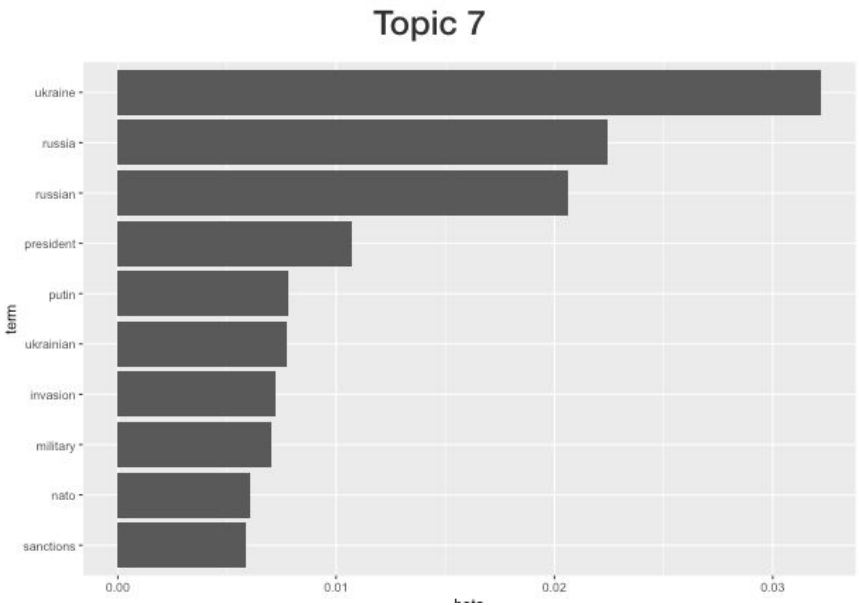


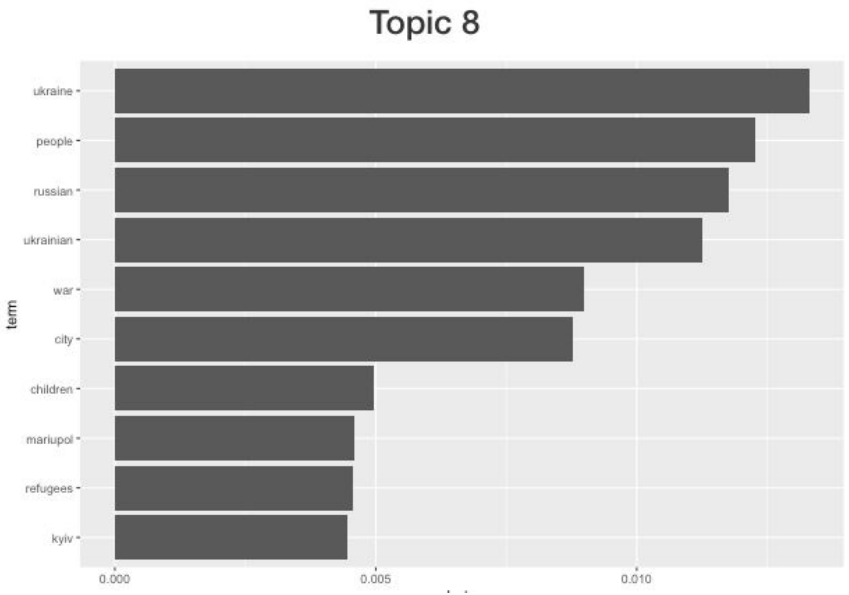
Topic Modeling Results - Part 3





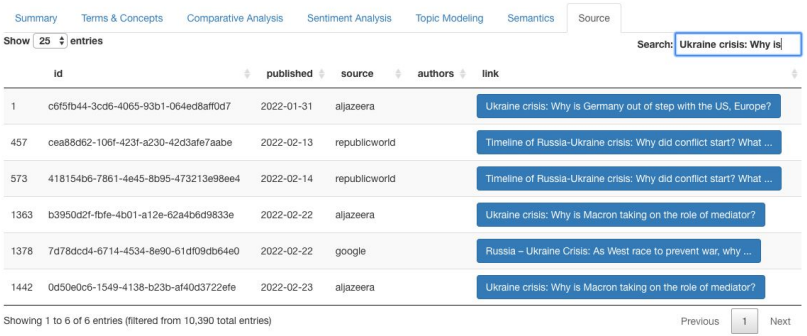
Topic Modeling Results - Part 4





Shiny App, Sources, Misc.

- Build a simple visualization application with some filtering controls (source, timeline)
- Due to issues with Google search throttling & blocking requests and NewsAPI rate limiting, processed offline and uploaded files to S3
- App loads files from S3 and builds visualizations slowly: from <10 seconds for “small” sample set of 1K articles and up to 3-5 minutes for 10K articles and 8 topic LDA (M1 Mac)
- Sources exposed via Shiny UI, can access original articles, if still available; spot checked some at random and link back to stats: another improvement?



Next Steps & Improvements

- SpacyR: semantics will be included in the final report and deliverable, some performance issues running on local machine with > 100 articles; hard to compare with “classic” text mining
  - Possible 1st application of SparklyR and Databricks
- Project report: summarize results and learnings in Markdown file

Future

- Automate data collection, especially scraping, by wrapping R script in Docker container and deploying to AWS Fargate (should help with URL throttling, rate limiting)
  - Look at proxy and other services as an alternative
- Automate the processing and allow for user selectable search argument
- Parse “byline” from Google search articles to extract more authors
- Topic modeling:
  - Explore coherence and algorithms for identifying K automatically
  - Potentially as a post-processing filter - feedback into pipeline
  - Subcategories of topics (multiple corpus) - multi-resolution analysis
- Analyze sentiment across sources and/or filter on trustworthy sources
- Derive a “trust” and “uniqueness” scores for each source to determine if the information is reliable and/or unique (adds to the bodies of data)
- Shiny App performance improvements: Shiny caching of reactive and non-reactive data