From Bag-of-Words to Large Language Models: Recent Advances in Topic Modelling

Rafael Campos-Gottardo

McGill University

January 23, 2026

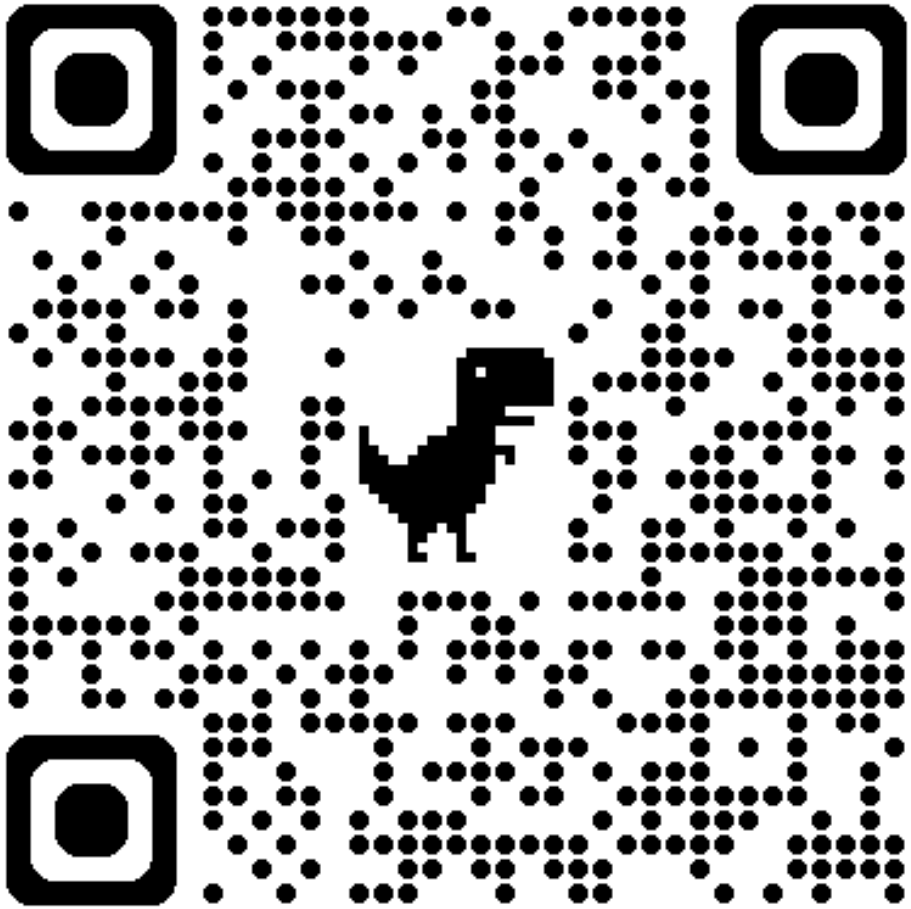
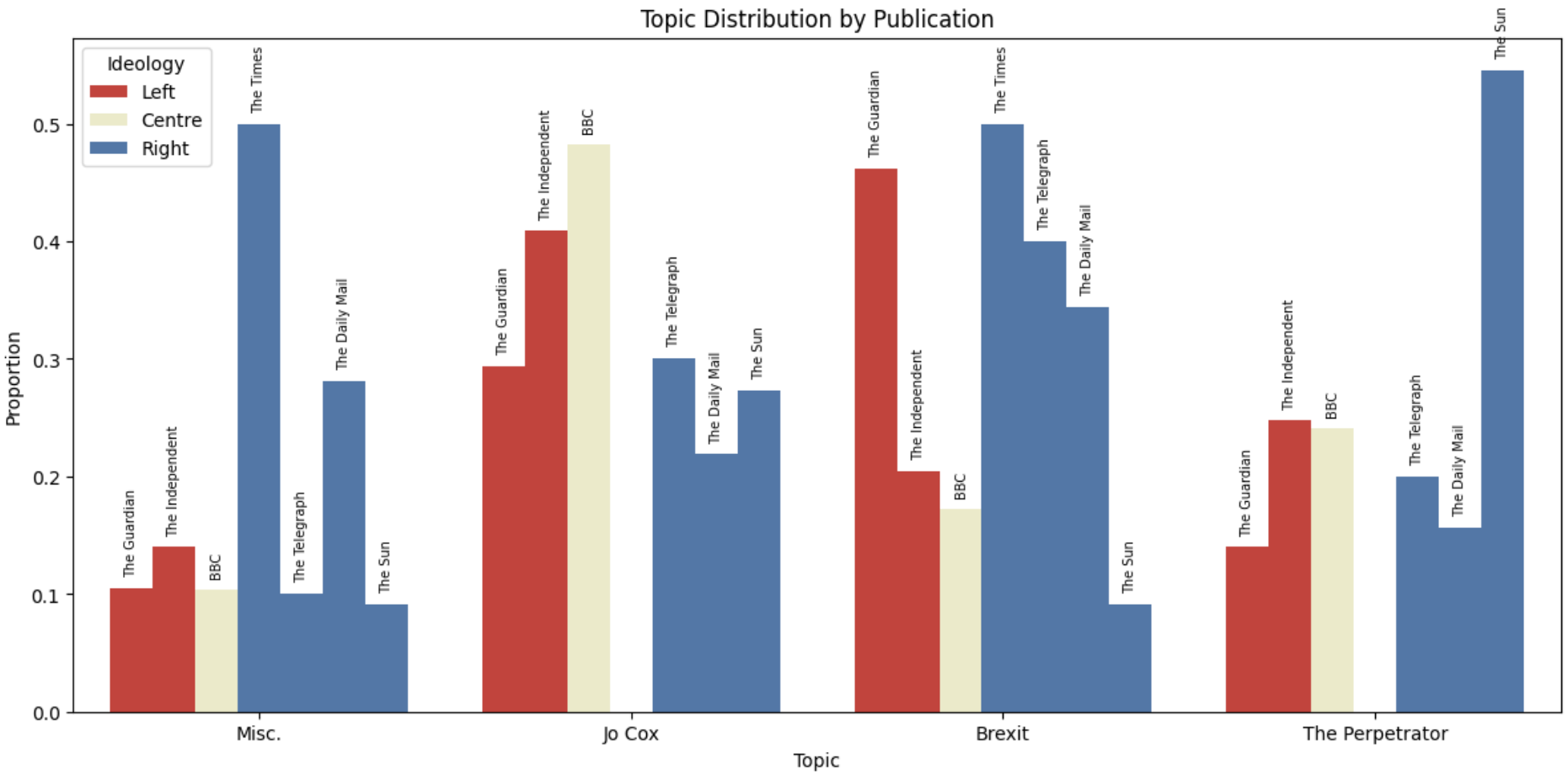


Figure: https://github.com/RafaelGottardo/Topic-Modelling

What is Topic Modelling?

- Method to extract meaning from political documents.
  - Traditional clustering algorithms assume that every document belongs to one cluster (e.g. k-means).
  - Topic modelling assigns each document with proportional membership in multiple clusters (topics).
  - Allows for more interpretable clusters.
- Topics are a collection of words that co-occur together.
  - Can be used to identity themes in text (Erlich et al. 2024)

Uses of Topic Modelling



Uses of Topic Modelling

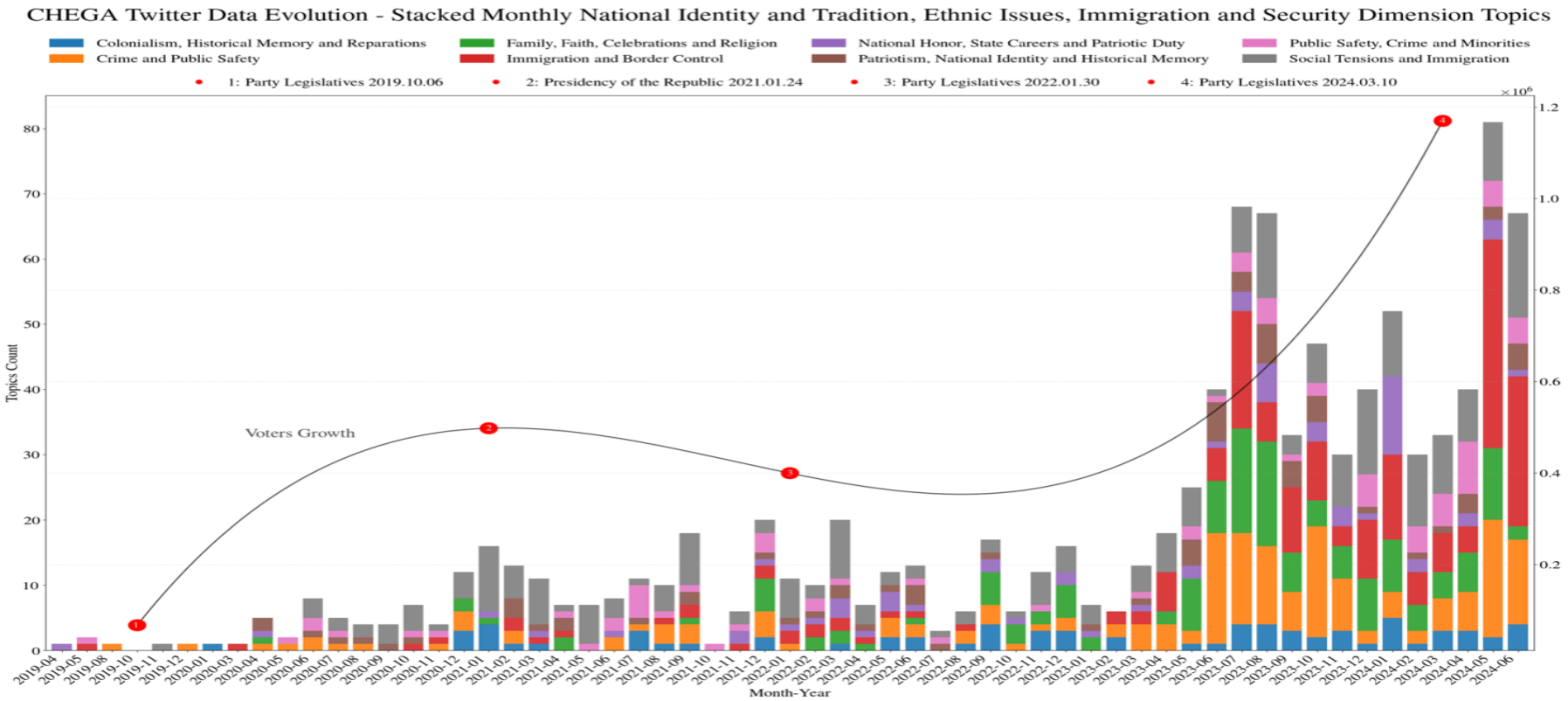


Figure 5 Monthly Evolution of Topics Related to National Identity, Tradition, Ethnic Issues, Immigration, and Security Dimensions (2019-2024).

Figure: Source: Cardoso et al. (2025)

Uses of Topic Modelling

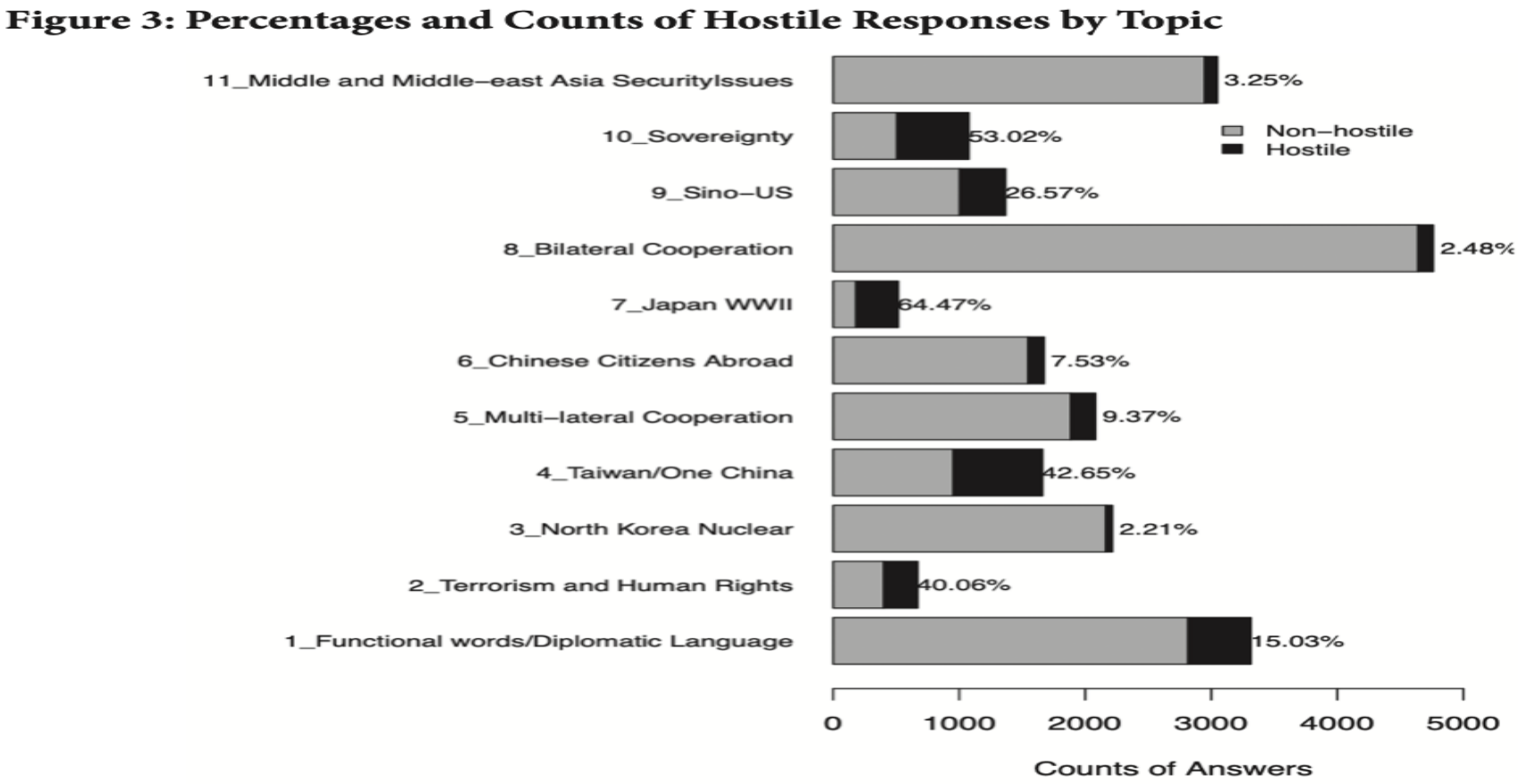


Figure: Source: Dai and Luqiu (2022)

Uses of Topic Modelling

Table 2: Topic Labels and Most Likely Terms within Each Topic

Number	Label	Most Relevant Term	Most Relevant Term Translated
1	Functional Words/ Diplomatic Language	中國，問題，是否，不，沒有，有關，會，進行，一，這個，情況，說，一些，方面,中方	China, question, whether, no, not, relevant, will, ongoing, one, this, situation, say, some, aspect
2	Terrorism and Human Rights	中國，組織，國際，人權，恐怖主義，打擊，反恐，恐怖，社會，合作，政府，反對，國家，人民，宗教	China, organization, international, human rights, terrorism, strike, anti-terrorism, terror, society, cooperation, government, against, nation, people, religion
3	North Korea Nuclear	朝鮮，會談，問題，六，方，中方，半島，各方，解決，對話有關，希望，朝，和平，努力	Korea (N.), talks, problem, six-party, China, peninsula, each party, solve, dialogue, relevant, hope, Korea (N.), peace, strive
4	Taiwan/One China	中國，美，美國 ，臺灣，關係，問題，中，美方，不，中方，一個，和平，發展，對此，國家	China, America, U.S., Taiwan, relationship, issue, China, America, no, Chinese side, one, peace, development, regarding, country
5	Multilateral Cooperation	中國，國家，合作，國際，發展，經濟，世界，中方，組織，會議，將，非洲，積極，歐盟，社會	China, country, cooperation, international, development, economic, world, China side, organization, conference, will, Africa, active, E.U., society

Figure: Source: Dai and Luqiu (2022)

Workshop Outline

- Data Pre-processing
  - Bag-of-words
  - Word embeddings
- Topic Modelling
  - Latent Dirichlet Allocation (LDA)
  - BERTopic
  - LLMs for Topic Modelling
- Lab Component
  - BERTopic in R

Data Pre-processing

Data pre-processing Bag-of-words

- Each document is represented by the number of times each word appear in the document.
- Documents are represented in a document feature matrix.

Document Feature Matrix

- *Officials in Minneapolis on Friday accused federal authorities of "hiding the facts" over the shooting of a United States citizen by an officer with the Immigration and Customs Enforcement (ICE) agency.*
- *Jacob Frey, the mayor of Minneapolis, criticized the response of the federal authorities to the shooting in Minneapolis.*

Document	Officials	Minneapolis	Friday	accused	federal	authorities	hiding	facts	shooting	United States	citizen	officer	Immigration	Customs	Enforcement	ICE	agency	Jacob	Frey	mayor	criticized	response
Doc 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
Doc 2	0	2	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1

Table: Document-Feature Matrix

Step 1: Tokenize

- Officials in Minneapolis on Friday accused federal authorities of ``hiding the facts'' over the shooting of a United_States citizen by an officer with the Immigration and Customs Enforcement (ICE) agency.
- Jacob Frey, the mayor of Minneapolis, criticized the response of the federal authorities to the shooting in Minneapolis.

Step 2: Reduce Complexity

- make text lowercase
- remove punctuation
- remove stop words (e.g. the, and, it, or; le, la, et, ou).
- `officials minneapolis friday accused federal authorities hiding facts shooting united_states citizen officer immigration and customs enforcement ice agency`
- `jacob frey mayor minneapolis criticized response federal authorities shooting minneapolis`

Step 3: Stem or Lemmatize

- Reduce words to their base:
  - Stem (authority $\rightarrow$ authorit)
  - lemmatization (better $\rightarrow$ good)
- `official minneapolis friday accuse federal authorit hid fact shoot united_states citizen officer immigration custom enforcement ice agency`
- `jacob frey mayor minneapolis criticiz response federal authorit shoot minneapolis`

Weaknesses of Bag-of-words

- Sometimes it is important to analyze stop words (e.g. determining authorship).
- Assumes that words that share the same base have the same meaning (e.g. car and cars; good and better)
- Ignores semantics: similar words in semantic space are treated differently (e.g. barrister and solicitor).
- Ignores order of the sentence.

Alternatives to Bag-of-words: Word Embedding

- Represents each word as a dense vector in a low dimensional space.
- Estimates the semantic similarity of words based on the words that appear with them.
  - The *usage* of a word suggests the *meaning* of a word.
- Stopwords are not removed because they are needed to encode meaning.
- Advantages:
  - Encode similarity
  - Allow for automatic generalization (learning about one word automatically tells us about another word)
  - provide a measure of meaning (similar meaning when used in similar contexts)
  - Fewer pre-processing steps required.

Estimating Word Embedding

- Word embeddings need to be trained.
- The most common training method is self-supervision.
- There exist a number of pre-trained embeddings, that can be drawn on for topic modelling.
  - `word2vec`
  - `doc2vec`
  - `GloVe`
  - `MTEB-French`
  - and more available from huggingface.co

Training Word Embedding

- Skip-gram and continuous bag-of-words.
- “I saw my · for lunch today”
- “I saw my grandmother for lunch today”
- “grandmother” can be used to predict the context words (Skip-gram)
- Context words can predict centre word (“mother,” “sister,” “friend”) (continuous bag-of-words).

Word Embeddings in Practice

- Word embeddings also allows for analogy completion.
- King $-$ Man $+$ Woman $=$ Queen
- Montreal Canadians $-$ Montreal $+$ Toronto $=$ Toronto Maple Leafs

An example from Grimmer et al. (2022)

- Two 300-dimensional vectors for the word `manufacturing` (pre-1960 and post-1960).
- Use cosine difference to compare the closest neighbours in each time period.
- **Pre-1960:**
  - `exportation`, `oil-related`, `concomitant`, `detriment`, `war-relate`.
- **Post-1960:**
  - `retool*`, `jobs`, `downtowns`, `offshoring`.

Word Embedding in R

```r
# load data
data("movie_review", package = "text2vec")

# prepare data for doc2vec
df <- data.frame(#create a unique id variable for each review
                 doc_id = paste0("movie_", movie_review$id),
                 # use a cleaning function from word2vec
                 text = txt_clean_word2vec(movie_review$review),
                 stringsAsFactors = FALSE)

# train the doc2vec model
model <- paragraph2vec(x = df, type = "PV-DBOW",
                       dim = 100, iter = 10, min_count = 5,
                       lr = 0.05, threads = 4)

# extract the document embeddings
embedding <- as.matrix(model, which = "docs")
```

Topic Modelling

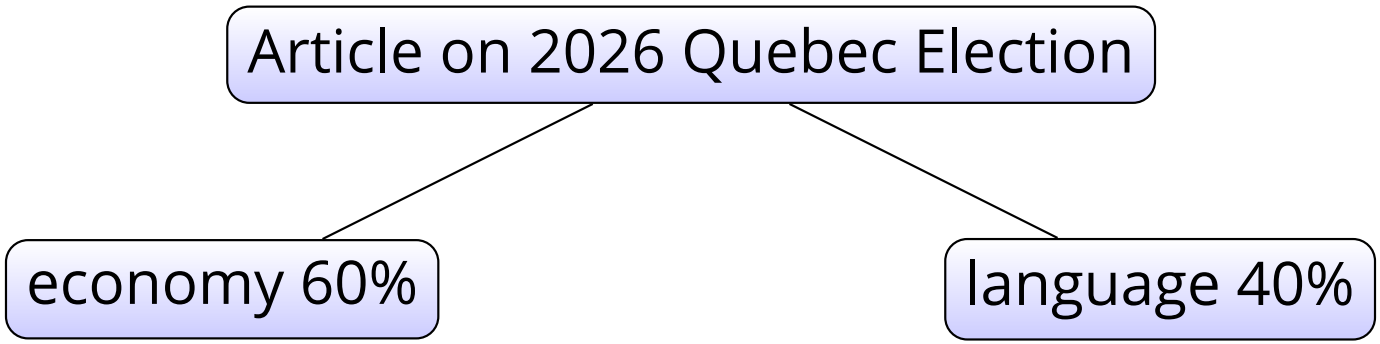
Latent Dirichlet Allocation (LDA)

- A text is a mixture of topics (weights that represent the prevalence of each topic).
- When writing a text the author first draws on a topic and then conditional on the topic draws the specific words.
- Allows for documents to have mixed membership in multiple topics.
- Bayesian Hierarchical Model.
- Based on how often words appear together.

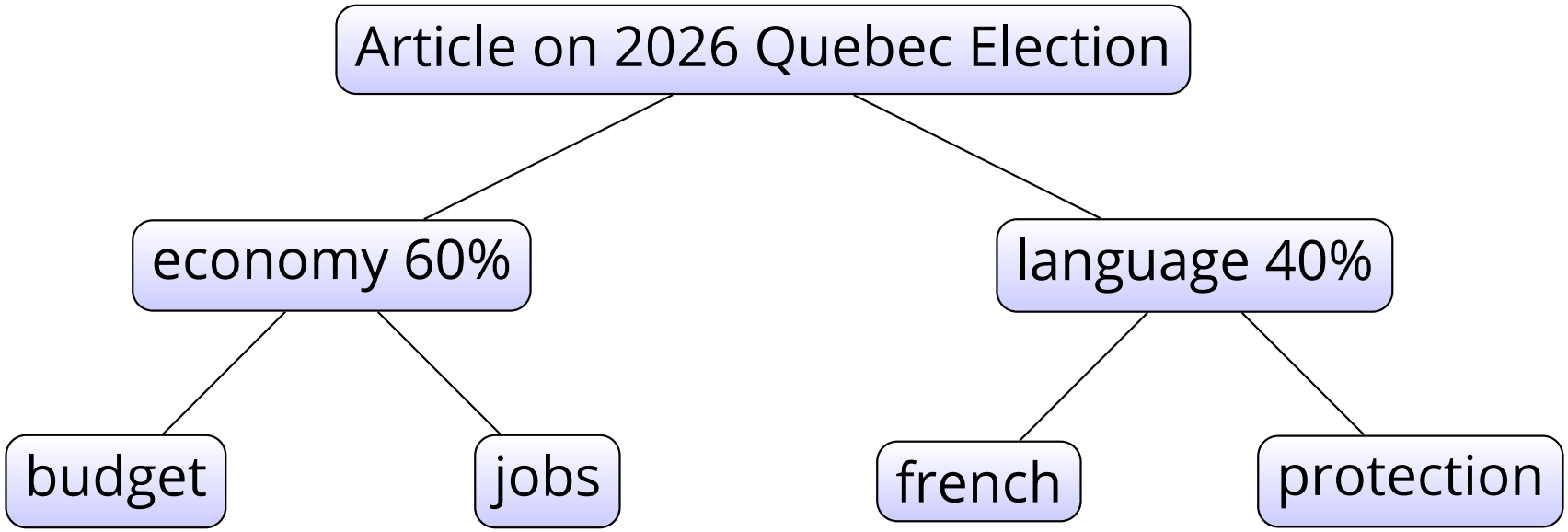
Visualization of LDA Theory

Article on 2026 Quebec Election

Visualization of LDA Theory



Visualization of LDA Theory



LDA in R

```
library(quanteda)

  ON_22$text <- as.character(ON_22$text)
ON_22_corp <- corpus(ON_22, text_field = "text")

toks <- tokens(ON_22_corp,
               what = "word",
               remove_punct = TRUE,
               remove_symbols = TRUE,
               remove_numbers = TRUE) %>%
  tokens_tolower() %>%
  tokens_remove(stopwords("english"))

dfm <- dfm(toks)
```

LDA in R

```
library(topicmodels)
library(tidytext)

lda_model <- LDA(
                    # specify the pre-processed data
                    dfm,
                    # specify the number of topic
                    k = 10,
                    # seed for replicabilty
                    control = list(seed = 1998)
                 )
```

LDA in R

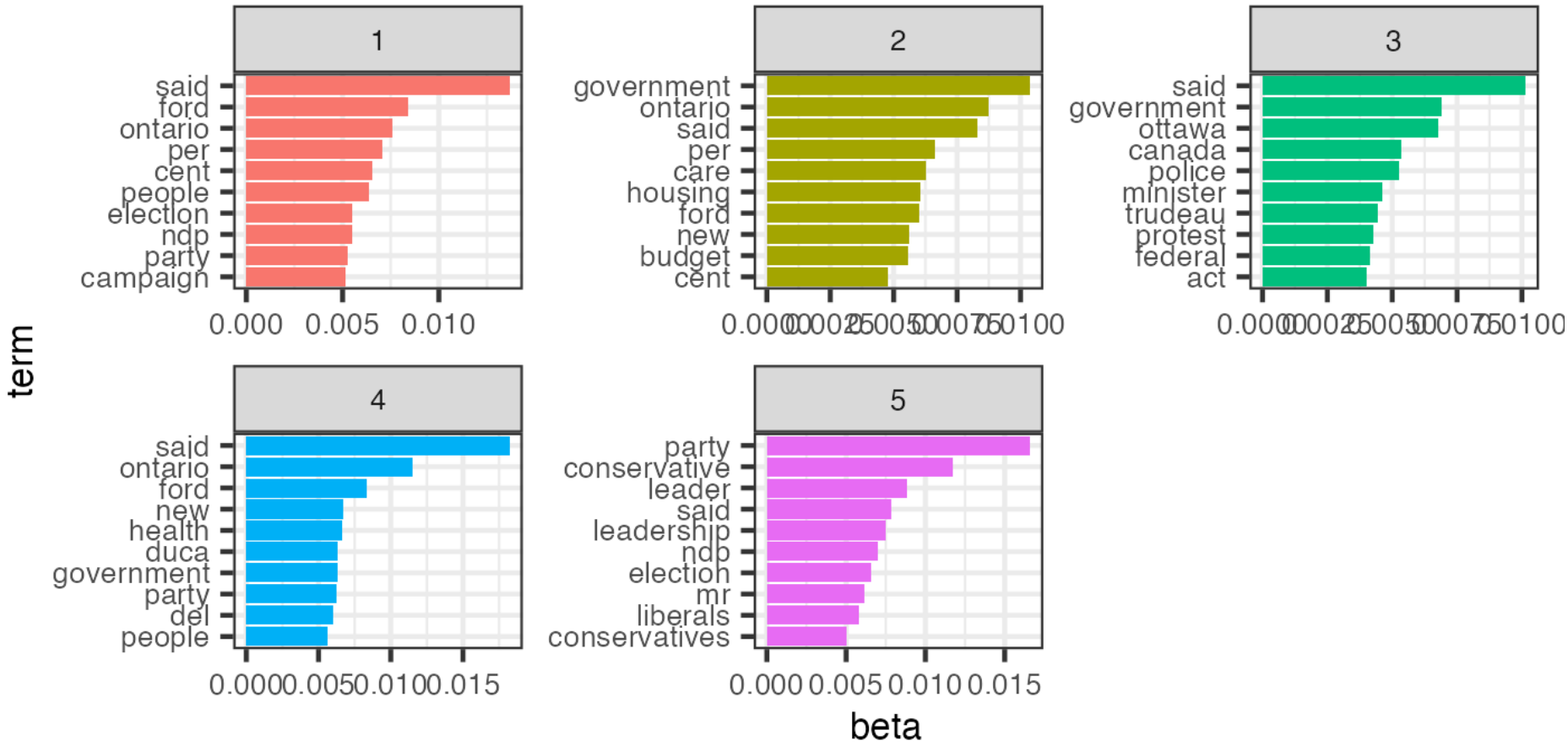
- Extract the per-topic word probabilities β

```
topics <- tidy(lda_model, matrix = "beta")
topics
```

topic	term	beta
1	ndp's	0.000609
2	ndp's	0.000008
3	ndp's	0.000000
4	ndp's	0.000280
5	ndp's	0.000388
6	ndp's	0.000000
7	ndp's	0.000039
8	ndp's	0.000000
1	sandy	0.000091
2	sandy	0.000000

LDA in R

- We can use the `slice_max()` function to extract the terms with the highest probability of being generated by each topic

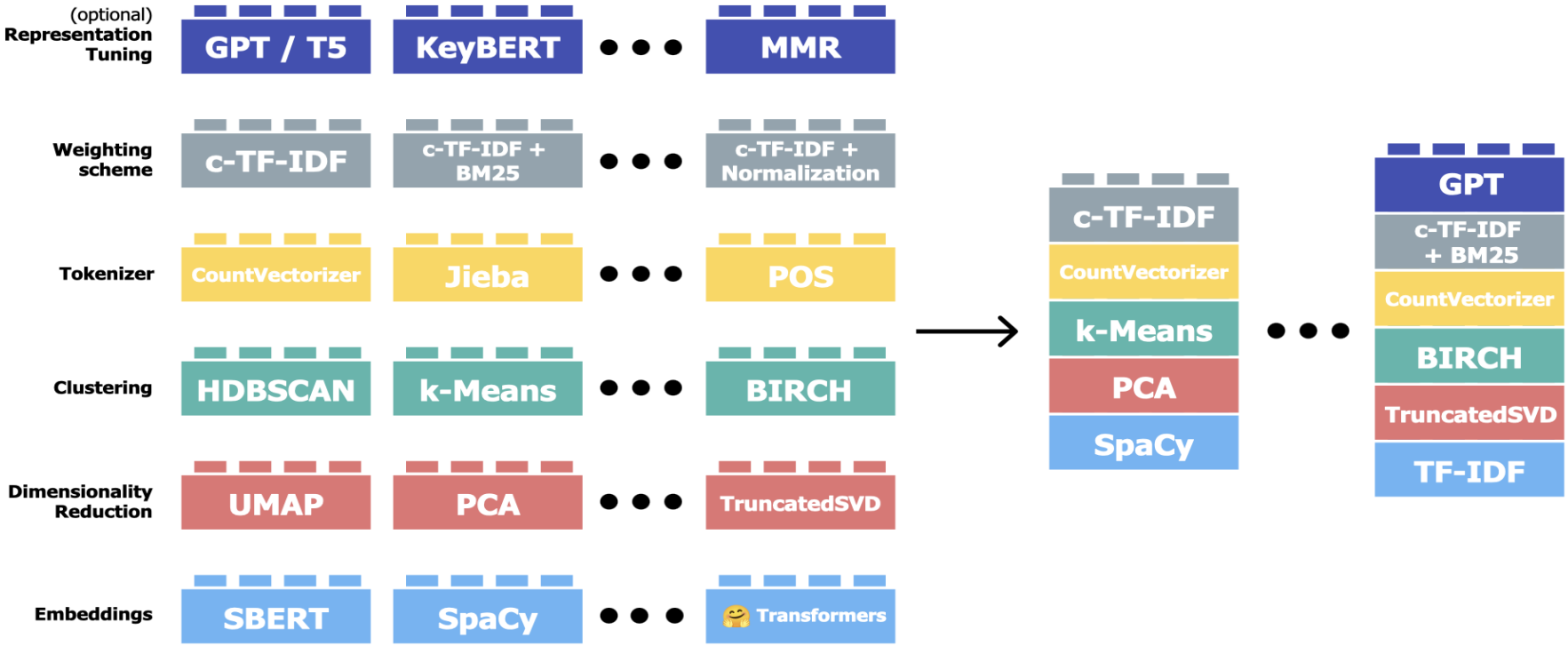


Weaknesses of LDA

- LDA requires the number of topics to be pre-specified.
- Often relies on “Bag-of-Words” representation.
- Struggles with short texts where there is limited variations in words.

What is BERTopic?

- A sequence of steps that create topic representations.
- usually based on Bidirectional Encoder Representations from Transformers (BERT), for document embedding.
  - Used for most English language google searches.



How does BERTopic work?

- BERTopic is highly flexible so here I only present the method recommended by Grootendorst (2022).

1. Document Embedding using a pre-trained transformer (e.g., cBERT).
2. Dimensionality reduction using uniform manifold approximation and projection.
3. Cluster documents using a clustering algorithm hierarchical density-based clustering approach (HDBSCAN).
4. Extract topics using cluster based Term Frequency, Inverse Document Frequency (c-TF-IDF).

Document Embedding

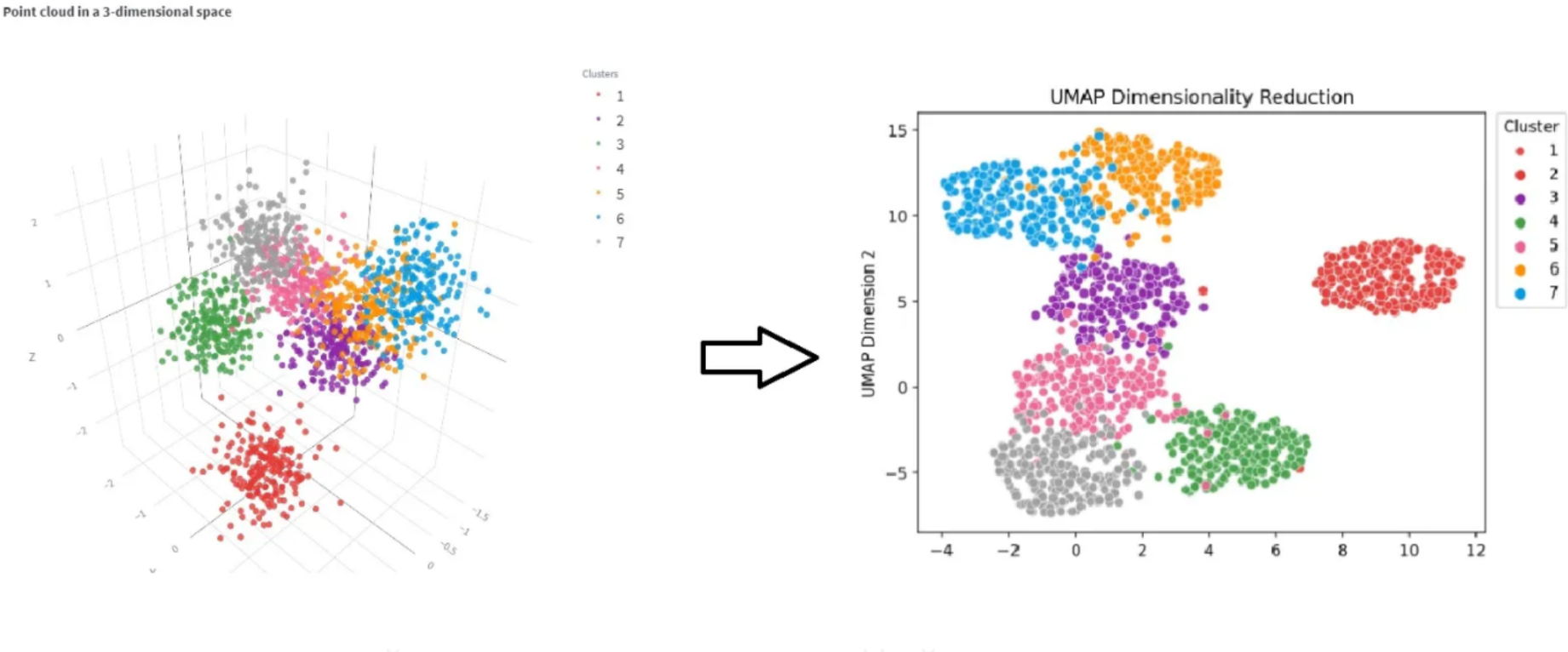
- Uses sentence BERT to create sentence embeddings that are compared using cosine similarity to locate sentences with similar meaning.
- Create vector space representations to be compared semantically.
- Assumes that documents on the same topic are semantically similar.

Dimensionality Reduction

- In high dimensional data the distance to the nearest data point approaches the distance to the furthest data point.
- The concept of spatial locality does not work well.
- Many methods: PCA, t-SNE, **UMAP**.

Uniform Manifold Approximation and Projection (UMAP)

- Preserves local and global features of high-dimensional data (McInnes et al. 2020).

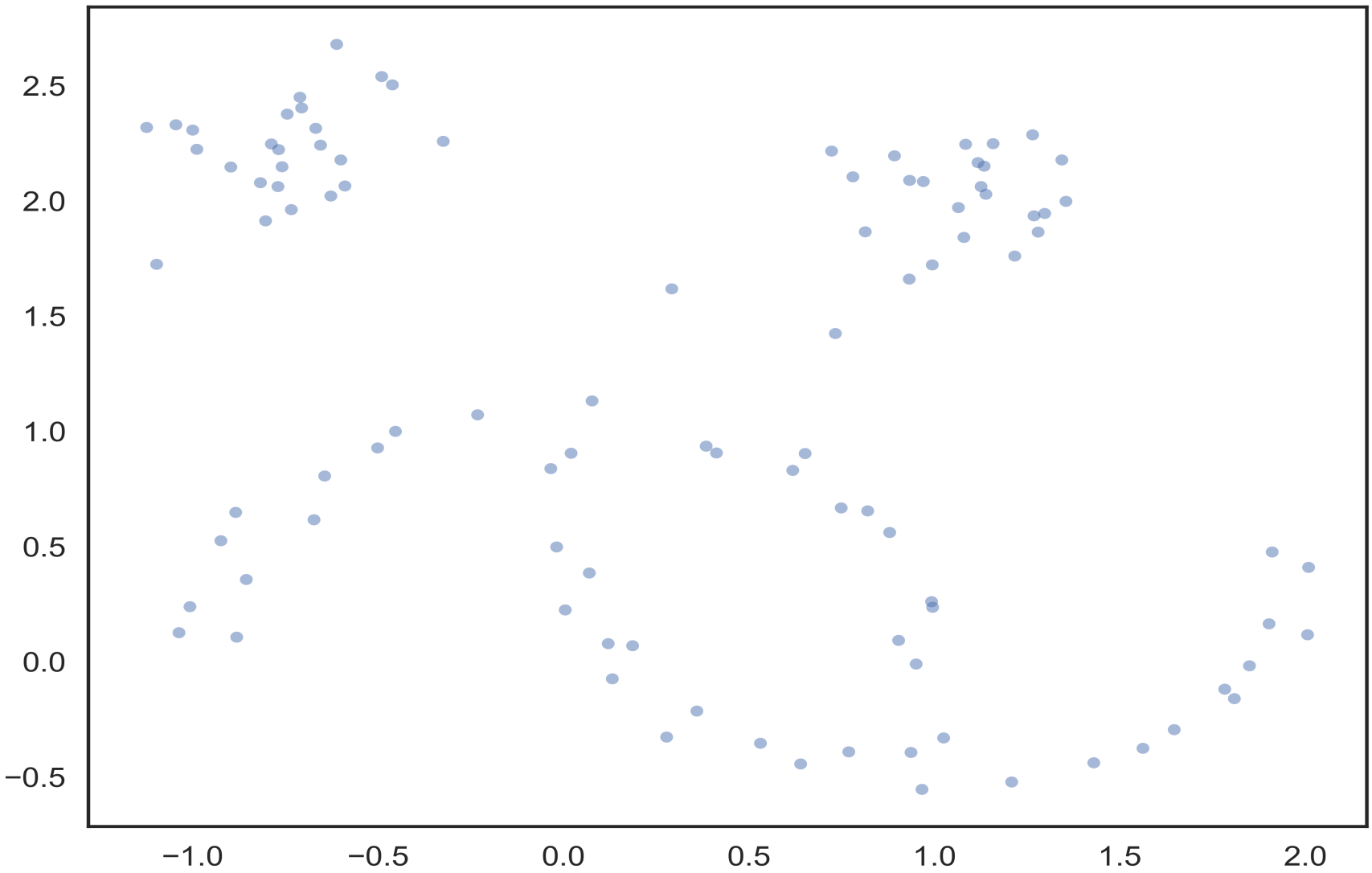


(Luna 2023)

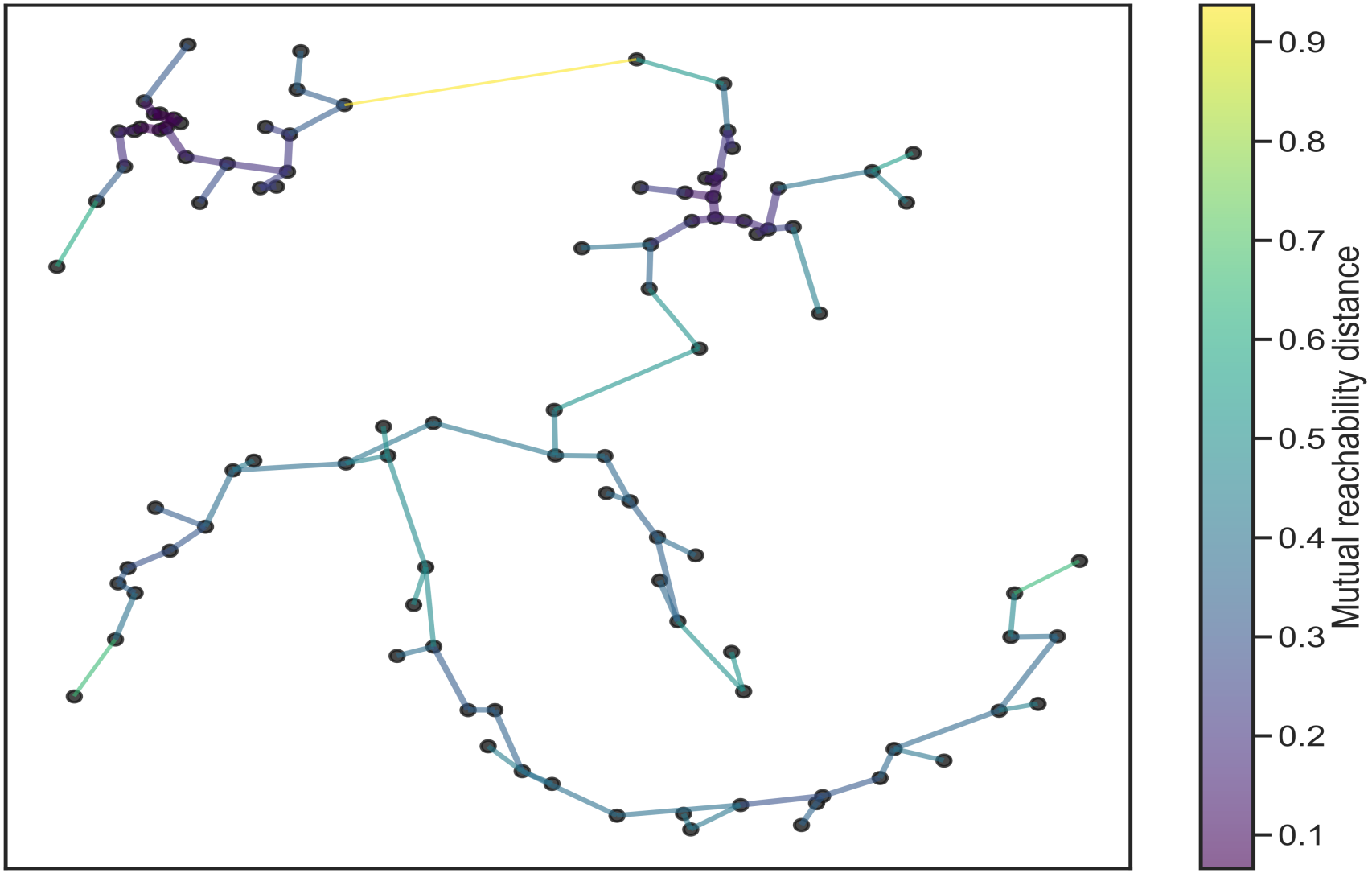
Cluster Documents

- Hierarchical density-based clustering approach (HDBSCAN)
- Allows noise to be modelled as outliers.

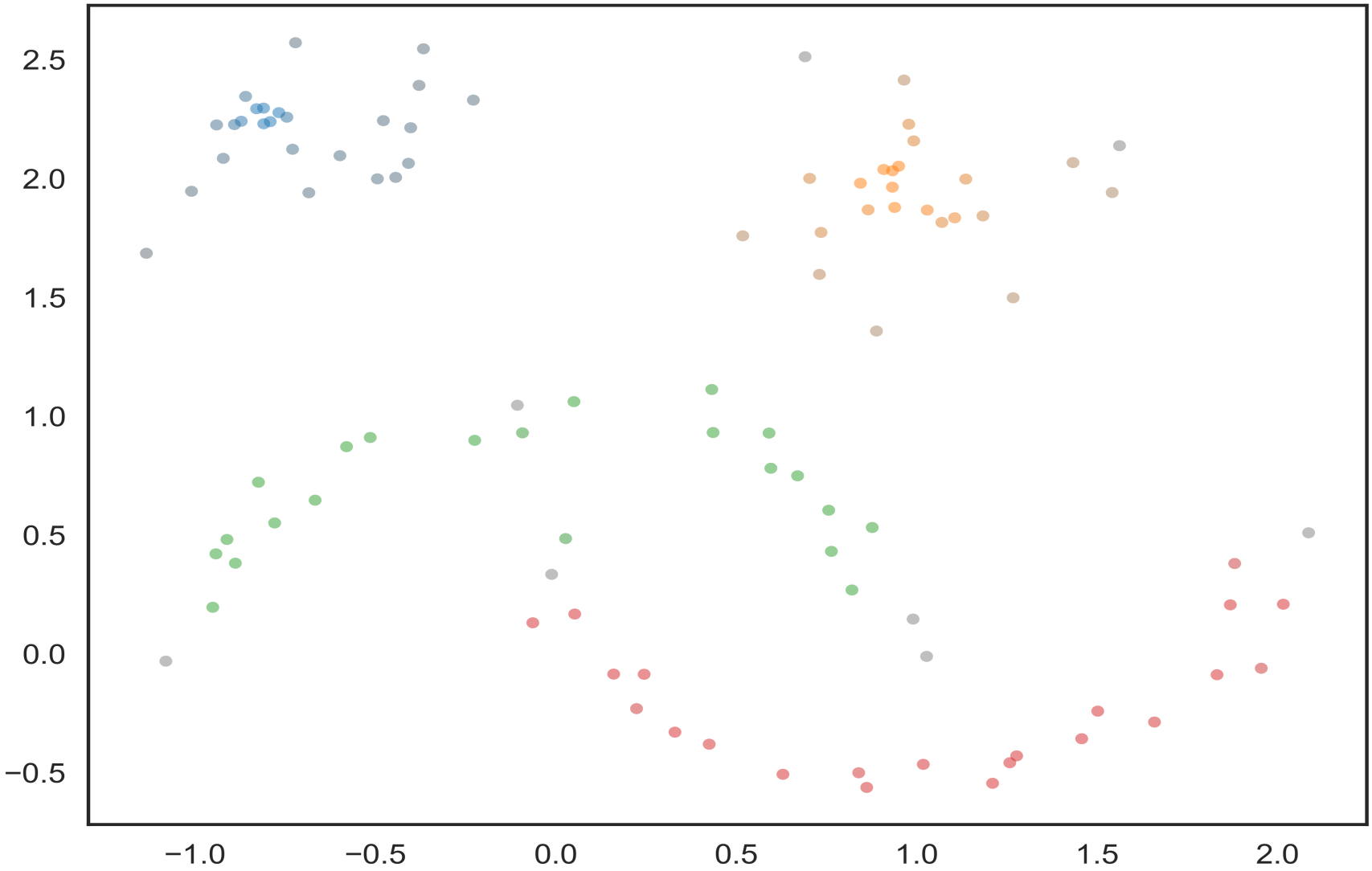
Visualizing HDBSCAN - Data



Minimum Spanning Tree - Mutually Reachability



HBDSCAN Clustered Data



Topic Representation

- TF-IDF measures the importance of a word to a document.
- How can we do this for topics (clusters)?

$$W_{t,d} = tf_{t,d} \cdot log(\frac{N}{df_t}) \tag{1}$$

$$W_{t,c} = tf_{t,c} \cdot log(1 + \frac{A}{tf_t}) \tag{2}$$

Where frequency is the frequency of a term t in a class (cluster) c. The inference class frequency is calculated as the logarithm of the average number of words in class A divided divided by the frequency of the term t across all classes. What makes one cluster different from another

TF-IDF in Practice

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.966
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

Table: Document frequency (df) and inverse document frequency (idf) by word (source Wikipedia)

cTF-IDF in Practice

$$W_{t,c} = tf_{t,c} \cdot log(1 + \frac{A}{tf_t}) \tag{3}$$

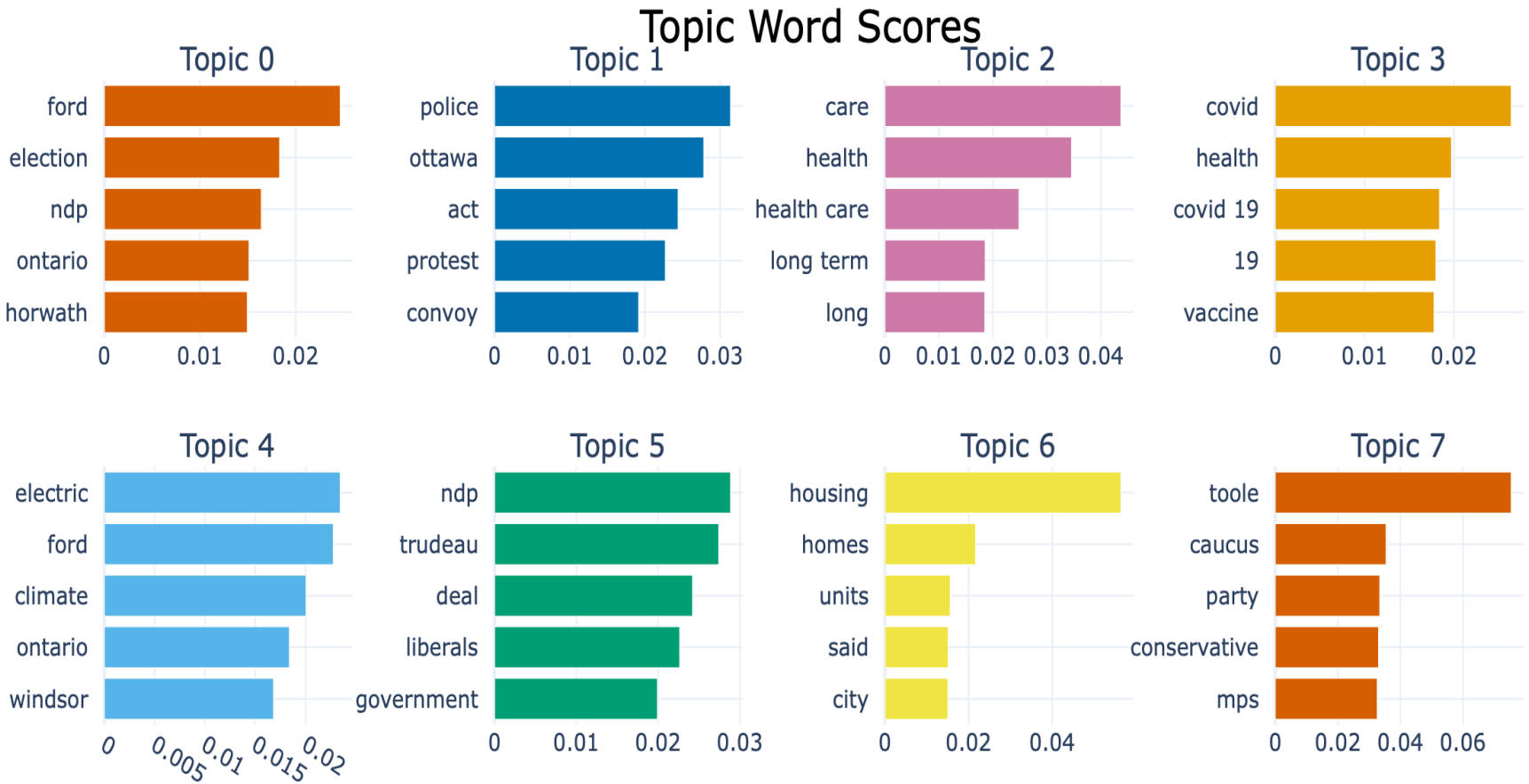
The teacher went to **school** and taught the students. Attending **school** is important for students learning and growth. **(A)**
The mayoral candidate promised to invest $20 billion to prevent **school** closures and $50 billion in hospitals. **(B)**

- **Cluster A:** $W_{t,c} = 2 \cdot log(1 + \frac{17.5}{3})$
- $W_{t,c} = 2 \cdot log(1 + 5.83)$
- $W_{t,c} = 2 \cdot 0.834$
- $W_{t,c} = 1.67$
- **Cluster B:** $W_{t,c} = 1 \cdot log(1 + \frac{17.5}{3})$
- $W_{t,c} = 0.834$

BERTopic in R

- `BERTopic` is designed for Python but there is an R version using `reticulate`.
- We will go over using BERT Topic in R in the lab component.

BERTopic in R



LLMs for Topic Modelling

- **Strengths:**
  - Recent research has shown that generative large language models produce topics that are more closely aligned with the human coded ground truth (e.g. `GPT` based models).
  - Topics have natural language labels.
  - More customizable than other methods.
- **Weaknesses**
  - Relies on closed models making replicability difficult.
  - "Zero-shot" topic modelling can often produce mis-specified topics.
  - The models do not provide information on the underlying process to generate topics.
  - Specific prompt engineering is required to ensure that models do not hallucinate.

One workflow `TopicGPT`

1. Topic Generation
   - Prompt an LLM to generate topics from a data set.
   - Provide new document and prompt LLM to assign it to an existing topic or create a new topic.
   - Refine topics by identifying duplicate topics and prompting LLMs to remove these topics.
2. Topic Assignment
   - Provide the LLM with the topics, descriptions and examples.
   - Ask the LLM to assign the one or more topics to each document.
   - Provide a quote to justify the assignment.

Lab Time

Rafael Campos-Gottardo
Rafael.campos-gottardo@mail.mcgill.ca
rafaelgottardo.github.io