Text Analysis I

PSC 8185: Machine Learning for Social Science

Iris Malone

April 18, 2022

Materials adapted from Rochelle Terman

Announcements

- Problem Set 7 Released: Due April 27
- April 27: Designated Monday
- Sign-up for Poster Session Slot

Recap

Where We've Been:

- Use R for most statistical analysis
- Python creates opportunities for webscraping
- BeautifulSoup and Pandas help us acquire text as data

New Terminology:

- BeautifulSoup
- HTML
- DOM
- APIs

Agenda

1. Text as Data

2. Pre-Processing

Text as Data

Computational Text Analysis Growing Popularity

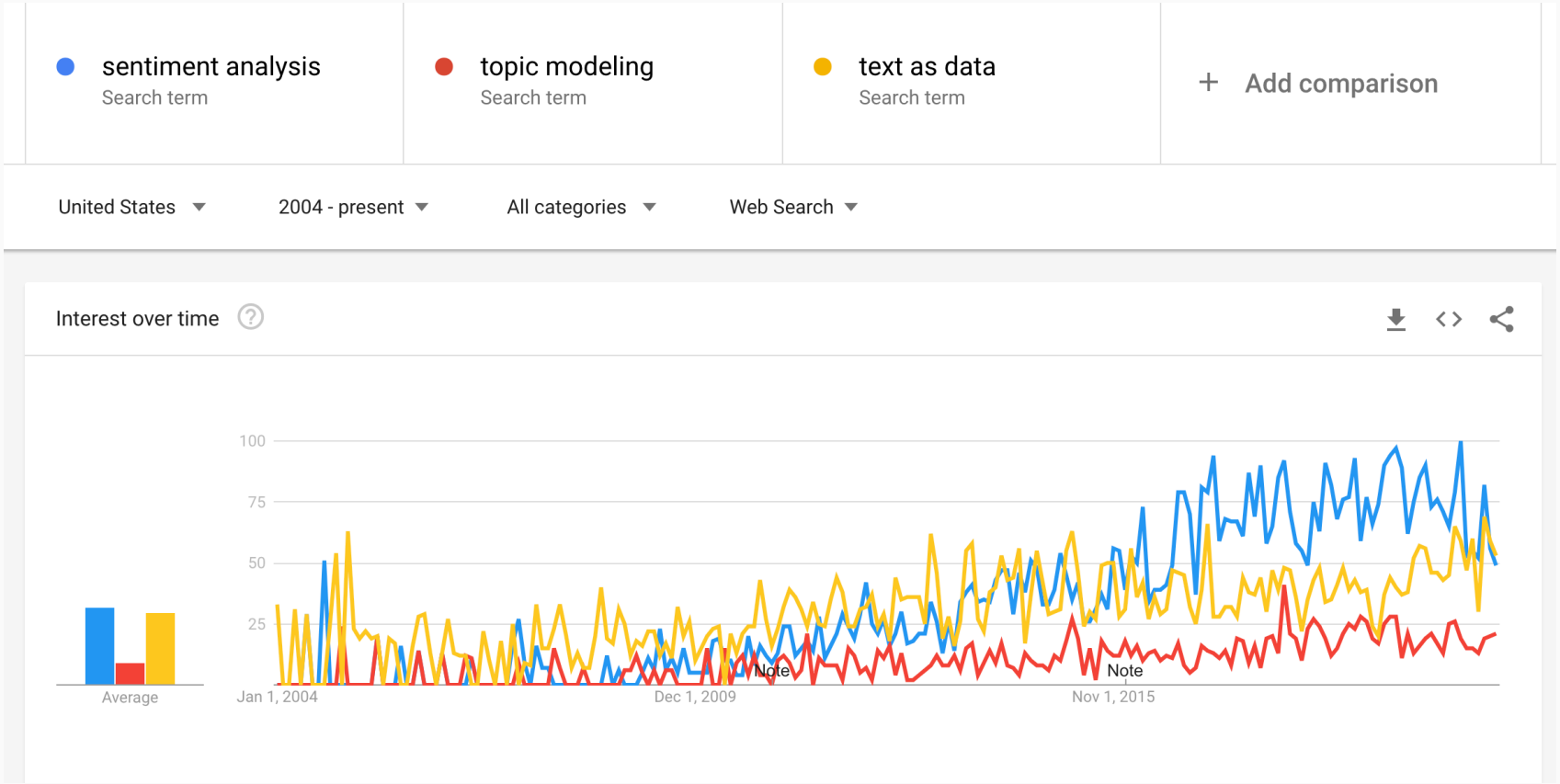

Figure 1: Trends, 2004-2022

ML Applications

- Political Science
  - Infer extent and strategy of Chinese censorship (King, Pan, and Roberts 2014)
  - Classify treaty provisions and language (Spirling 2012)

ML Applications

- Political Science
  - Infer extent and strategy of Chinese censorship (King, Pan, and Roberts 2014)
  - Classify treaty provisions and language (Spirling 2012)
- Economics/Finance
  - Identify Optimal Advertising Slogans
  - Measure Effect of Regulatory Comments

ML Applications

- Political Science
  - Infer extent and strategy of Chinese censorship (King, Pan, and Roberts 2014)
  - Classify treaty provisions and language (Spirling 2012)
- Economics/Finance
  - Identify Optimal Advertising Slogans
  - Measure Effect of Regulatory Comments
- Government and Industry
  - Identify extremist chatter
  - Counter Disinformation Campaigns

Why use Text as Data?

- We care about language
- Text as (qualitative) data is well-established norm
- Analyzing large texts is time-consuming, but computers can lower these costs

Problems with Text Analysis

Speech is:

- Ironic

  Thanks, Obama.

Problems with Text Analysis

Speech is:

- Ironic

  Thanks, Obama.

- Subtle

  They have not succeeded and will never succeed in breaking the will of the people.

Problems with Text Analysis

Speech is:

- Ironic

  Thanks, Obama.

- Subtle

  They have not succeeded and will never succeed in breaking the will of the people.

- Informal

  She's so extra.

Problems with Text Analysis

Speech is:

- Ironic

  Thanks, Obama.

- Subtle

  They have not succeeded and will never succeed in breaking the will of the people.

- Informal

  She's so extra.

- Order Dependent

  Peace, no more war. War, no more peace.

Problems with Text Analysis

Speech is:

- Ironic

  Thanks, Obama.

- Subtle

  They have not succeeded and will never succeed in breaking the will of the people.

- Informal

  She's so extra.

- Order Dependent

  Peace, no more war. War, no more peace.

- Multi-Language

Problems with Text Analysis

Speech is:

- Ironic

  Thanks, Obama.

- Subtle

  They have not succeeded and will never succeed in breaking the will of the people.

- Informal

  She's so extra.

- Order Dependent

  Peace, no more war. War, no more peace.

- Multi-Language

Overall: Validate, validate, validate.

Text Analysis Procedures

Supervised:

Unsupervised:

Text Analysis Procedures

Supervised:

- Hand code set of documents
- Train model on handcoded documents
- Predict content of unlabeled documents

Unsupervised:

Text Analysis Procedures

Supervised:

Unsupervised:

- **Sentiment Analysis:** Measure content of documents
- **TF-IDF:** Identify distinctive words
- **Topic Modeling:** Cluster text into categories

Pre-Processing

Procedure

Motivation: Need to prepare texts for computational text analysis by removing ‘noise’ in the documents:

Procedure

Motivation: Need to prepare texts for computational text analysis by removing 'noise' in the documents:

1. Acquire Text
2. Assemble into a **corpus**
3. Remove capitalization and punctuation
4. Discard word order
5. Combine similar terms
    - Stemming
    - Lemmatization
6. Create a count vector
7. Create a **Document Term Matrix**

Procedure

1. **Acquire Text**
2. Assemble into a **corpus**
3. Remove capitalization and punctuation
4. Discard word order
5. Combine similar terms
    - Stemming
    - Lemmatization
6. Create a count vector
7. Create a **Document Term Matrix (DTM)**

Acquiring Texts

Main Idea: Computational text analysis requires **machine readable text** meaning

- plain text (.txt or .csv) file
- common language
- encoded in UTF-8 or ASCII
- metadata (e.g., author, data, unique label)

Acquiring Texts

Popular Sources:

- Online databases, e.g. LexisNexis, Comparative Manifesto Project, Foreign Broadcast Information Service
- Websites
  - Scraping
  - APIs
- Archives
  - Pre-Digitized, e.g. FRUS or Wilson Center
  - OCR-Compatible, e.g. high quality scanner + optical character recognition

Procedure

1. Acquire Text
2. **Assemble into a corpus**
3. Remove capitalization and punctuation
4. Discard word order
5. Discard stop words
6. Combine similar terms
    - Stemming
    - Lemmatization
7. Create a count vector
8. Create a **Document Term Matrix**

Preparing a Corpus

Def. **corpus**: a collection of texts, ususally stored electronically, and from which we perform our analysis

Preparing a Corpus

Def. **corpus**: a collection of texts, ususally stored electronically, and from which we perform our analysis

Key Components:

- **Documents:** elements within a corpus, e.g. chapter
- **Segments:** elements within a document, e.g. paragraph
- **Tokens:** elements within a segment, e.g. word

Preparing a Corpus

Rule of Thumb:

- Make sure text is machine readable
- Use for loops to merge and append data as necessary
- Each document is a row, one column for text, and other columns for metatadata

Preparing a Corpus

	date	user_loc	followers	friends	message	bbox_coords
0	2018-04-13 08:14:22	NaN	61	367	This is Paul Ryan. Exactly. 100%. Fuck Paul Ryan. https://t.co/MYxN9jOas8	[[[-74.988897, 39.810025], [-74.908642, 39.810025], [-74.908642, 39.87514], [-74.988897, 39.87514]]]
1	2018-04-13 08:01:58	Long Island, NY	3656	3549	There's a video from the Daily Show (Comedy Central) you'll want to see.\r\nhttps://t.co/wjKUW9wBXo	[[[-79.76259, 40.477383], [-71.777492, 40.477383], [-71.777492, 45.015851], [-79.76259, 45.015851]]]
2	2018-04-13 07:48:16	Alameda Ca	2304	2917	@SallyAlbright First of all, discrediting the FBI is disgusting and borderline treasonous. Secondly, I wonder if t… https://t.co/BdnZrBCKhq	[[[-122.332411, 37.720367], [-122.224562, 37.720367], [-122.224562, 37.797229], [-122.332411, 37.797229]]]
3	2018-04-13 07:45:37	San Francisco, CA	52	876	Paul Ryan is a coward and a piece of shit! Good riddance! Now if the rest of the pieces of shit in that party would…	[[[-121.6919801, 36.643802], [-121.5905572, 36.643802], [-121.5905572, 36.7344965]

Procedure

1. Acquire Text
2. Assemble into a **corpus**
3. **Remove capitalization and punctuation**
4. Discard word order
5. Discard stop words
6. Combine similar terms
    - Stemming
    - Lemmatization
7. Create a count vector
8. Create a **Document Term Matrix**

Remove Capitalization and Punctuation

Main Idea: We are interested in the meaning and frequency of different words in a document. Documents have words, but also lots of extraneous stuff like ...

Remove Capitalization and Punctuation

Main Idea: We are interested in the meaning and frequency of different words in a document. Documents have words, but also lots of extraneous stuff like ...

- Capitalization
- Punctuation
- Numbers
- Emojis
- Slang
- Dates

Remove Capitalization and Punctuation

Main Idea: We are interested in the meaning and frequency of different words in a document. Documents have words, but also lots of extraneous stuff like ...

- Capitalization
- Punctuation
- Numbers
- Emojis
- Slang
- Dates

Consequence: Including these elements (1) reduces comparability and (2) does not provide useful information

Remove Capitalization and Punctuation

Main Idea: We are interested in the meaning and frequency of different words in a document. Documents have words, but also lots of extraneous stuff like ...

- Capitalization
- Punctuation
- Numbers
- Emojis
- Slang
- Dates

Consequence: Including these elements (1) reduces comparability and (2) does not provide useful information

Caution: "Turkey" = "turkey"

Application: Gettysburg Address

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated can long endure.

Application: Gettysburg Address

now we are engaged in a great civil war testing whether that nation or any nation so conceived and so dedicated can long endure

Procedure

1. Acquire Text
2. Assemble into a **corpus**
3. Remove capitalization and punctuation
4. **Discard word order**
5. Discard stop words
6. Combine similar terms
    - Stemming
    - Lemmatization
7. Create a count vector
8. Create a **Document Term Matrix**

Discard Word Order

Main Idea: We assume that word order doesn't matter in order to facilitate **tokenization**

Discard Word Order

Main Idea: We assume that word order doesn't matter in order to facilitate **tokenization**

Tokenization:

Discard Word Order

Main Idea: We assume that word order doesn't matter in order to facilitate **tokenization**

Tokenization:

- Treats words in a document as a **"bag of words"**
- Ignores long sequencing (otherwise RNN)
- Transforms words into word vector
- Different word lengths:
  - Unigram
  - Bigram
  - Trigram
  - ...

Application

now we are engaged in a great civil war testing whether that nation or any nation so conceived and so dedicated can long endure

[now, we, are, engaged, in, a, great, civil, war, testing, whether, that, nation, or, any, nation, so, conceived, and, so, dedicated, can, long, endure]

[a, and, any, are, can, conceived, dedicated, endure, engaged, great, in, long, nation, now, or, so, so, testing, that, war, we, whether]