Text as Data: Computational Text Analysis

Week 1: Introduction to Computational Text Analysis and NLP

Ashrakat Elshehawy

Department of Politics and International Relations,
University of Oxford

April 26, 2021

Organizational

- All course material will be posted on **GitHub**.
- For our coding sessions we will use Google Colab Notebooks, coding sheets and their solutions are hosted on the GitHub Page of the class.
- You can easily open any sheet from within GitHub by clicking on *"Open in Colab"* button, remember to always save a copy of your work.
- You should have a Google Drive account to easily work with Colab and load data.
- The Teams Website will serve our needs for communication.
- The Syllabus will be hosted on Canvas.

Class Times and Office Hours

We will meet on the following days on Zoom:

- Week 1: Monday the 26th of April 14.00-16.00
- Week 2: Tuesday the 4th of May 14.00-16.00
- Week 3: Monday the 10th of May 14.00-16.00
- Week 4: Monday the 17th of May 14.00 16.00

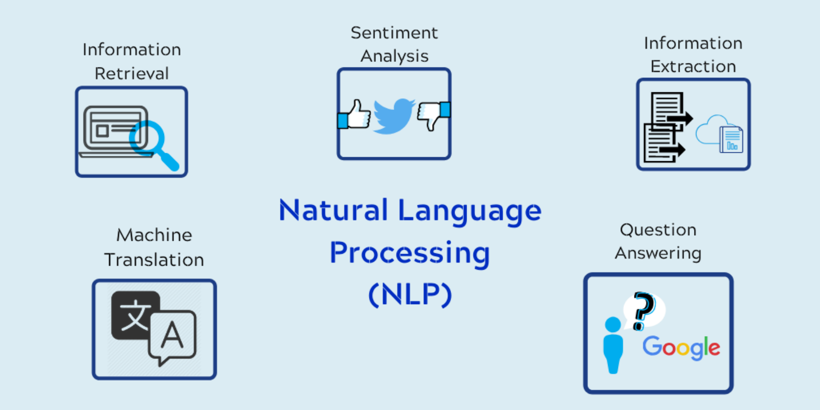
Office hours take place from week 1 to 4 on Fridays 4-6 pm. You need to book a slot beforehand (link in the Syllabus). Please come in the zoom room to attend the office hour sessions in the time you have booked.

Class Assessment

You will receive a coding assignment on the last session of the class, due by the Friday of the 5th week 24.00 - via email to the instructor (ashrakat.elshehawy@politics.ox.ac.uk).

What is Natural Language Processing?[1]

The application of computational techniques for the the analysis of human language.



Source: Cybiant

[1] Some of the slides are partly based on Federico Nanni's course of Computational Text Analysis at U Mannheim

Why is Natural Language Processing helpful for Political Science?

That is what we are learning in this class.

Most of data sets we use are already derived from Text

Examples:

- Understanding Political Behavior

Why is Natural Language Processing helpful for Political Science?

That is what we are learning in this class.

Most of data sets we use are already derived from Text

Examples:

- Understanding Political Behavior
- Coding Sentiment

Why is Natural Language Processing helpful for Political Science?

That is what we are learning in this class.

Most of data sets we use are already derived from Text

Examples:

- Understanding Political Behavior
- Coding Sentiment
- Frequency of Conflict

Why is Natural Language Processing helpful for Political Science?

That is what we are learning in this class.

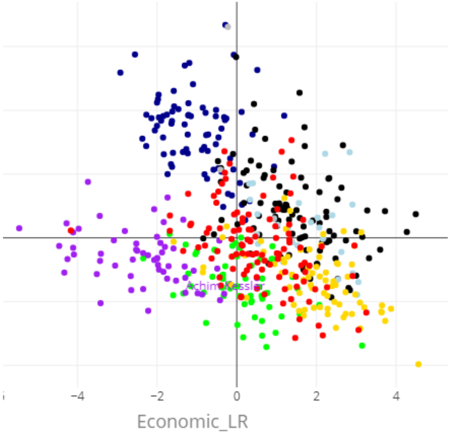
Most of data sets we use are already derived from Text

Examples:

- Understanding Political Behavior
- Coding Sentiment
- Frequency of Conflict
- Coding Topic Areas

Example: Infer from Twitter Text Ideological Positions of German Politicians

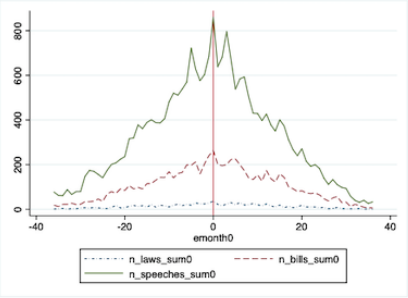
Marius Sälzer PhD Project, Universität Mannheim



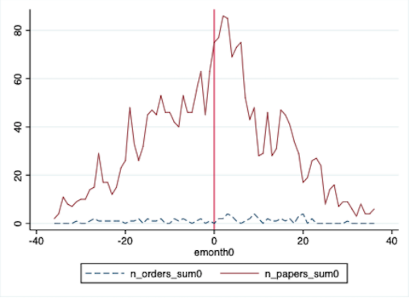
Source: Marius Sälzer Shiny app birdfish
Saeltzer, Marius (2020): Finding the Bird's Wings: Dimensions of Factional Conflict on Twitter. Party Politics (forthcoming).
DOI: 10.1177/1354068820957960

Example: How the US Electoral Cycle affects Elections Around the World?[2]

Documents Critical of a Country's Election: months -/+ election date



(a) Congressional Speeches, Bills, Laws



(b) Presidential Papers and Orders

[2] Source: Bubeck, J, Elshehawy, Ashrakat, Marinov, Nikolay, and Federico Nanni, 2021. How the US Electoral Cycle Affects Elections Around the World. Working Paper

Example: Women's Authority in Patriarchal Movements

Table 1 Men Use Hadith-Related Phrases More Than Women

Term	% of Documents Using Term		% of All Words	
	Men	Women	Men	Women
allāh	95	85	3.02	1.60
ṣalā allāh	67	32	0.50	0.21
raḍī allāh	50	20	0.18	0.10
ḥadīth	64	39	0.30	0.14
qāl/yaqūl (al-)rasūl	30	10	0.047	0.026
Ibn Taymiyya	23	3	0.030	0.006
al-Bukhari	41	10	0.098	0.030
al-Bayhaqi	17	1	0.013	0.001
Abu Hurayra	29	4	0.048	0.008
al-Tabarani	16	1	0.014	0.002
Abu Dawud	29	4	0.043	0.006

Note: This table counts the use of hadith-related phrases in 3,470 documents (1,306,641 words) by women and 17,854 documents (74,991,711 words) by men on saaid.net. The left two columns show the percentage of men's and women's documents that contain each phrase. The right two columns show the percentage of men's and women's words devoted to each phrase. All gender differences are statistically significant at the .05 percent level.

Source: Nielsen, R.A., 2020. Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers. American Journal of Political Science, 64(1), pp.52-66.

Principles of Computational Text Analysis - Grimmer and Stewart (2013)

- All quantitative models of language are wrong - but some are useful (Language is complex, "Time flies like an arrow. Fruit flies like a banana.")

Principles of Computational Text Analysis - Grimmer and Stewart (2013)

- All quantitative models of language are wrong - but some are useful (Language is complex, "Time flies like an arrow. Fruit flies like a banana.")
- The importance of *"the human in the loop"*
(Deep understanding of text is necessary)

Principles of Computational Text Analysis - Grimmer and Stewart (2013)

- All quantitative models of language are wrong - but some are useful (Language is complex, "Time flies like an arrow. Fruit flies like a banana.")
- The importance of *"the human in the loop"*
  (Deep understanding of text is necessary)
- No-free-lunch theorem
  (There is no single best algorithm)

Principles of Computational Text Analysis - Grimmer and Stewart (2013)

- All quantitative models of language are wrong - but some are useful (Language is complex, "Time flies like an arrow. Fruit flies like a banana.")
- The importance of *"the human in the loop"*
  (Deep understanding of text is necessary)
- No-free-lunch theorem
  (There is no single best algorithm)
- Always validate!

Overview of the Course

- Corpus Creation

Overview of the Course

- Corpus Creation
- Text-Processing and Cleaning

Overview of the Course

- Corpus Creation
- Text-Processing and Cleaning
- Dictionaries and Keyword-lists

Overview of the Course

- Corpus Creation
- Text-Processing and Cleaning
- Dictionaries and Keyword-lists
- Term-doc Matrices, Vector Space Representation

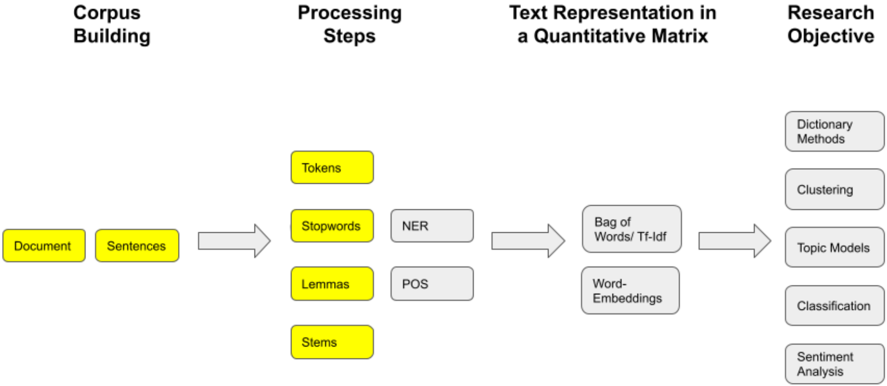
Overview of the Course

- Corpus Creation
- Text-Processing and Cleaning
- Dictionaries and Keyword-lists
- Term-doc Matrices, Vector Space Representation
- Unsupervised Techniques (e.g Topic-Models and Clustering)

Overview of the Course

- Corpus Creation
- Text-Processing and Cleaning
- Dictionaries and Keyword-lists
- Term-doc Matrices, Vector Space Representation
- Unsupervised Techniques (e.g Topic-Models and Clustering)
- Supervised Machine-Learning Techniques (e.g Classification)

From Text to Data



Overview of Data Collections Available

Newspaper Collections

- New York Times Corpus (1987-2007, 1.8 million articles)

Overview of Data Collections Available

Newspaper Collections

- New York Times Corpus (1987-2007, 1.8 million articles)
- Die Zeit

Overview of Data Collections Available

Newspaper Collections

- New York Times Corpus (1987-2007, 1.8 million articles)
- Die Zeit
- Historical Corpus of German Newspaper (1650-1800) https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2544

Overview of Data Collections Available

Newspaper Collections

- New York Times Corpus (1987-2007, 1.8 million articles)
- Die Zeit
- Historical Corpus of German Newspaper (1650-1800) https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2544
- Lexis Nexis - usually hard to scrape from

Overview of Data Collections Available to use as a Text Corpus

Political Speeches

- EuroParl: `http://www.talkofeurope.eu/data/`

Overview of Data Collections Available to use as a Text Corpus

Political Speeches

- EuroParl: `http://www.talkofeurope.eu/data/`
- UK: `https://www.hansard-corpus.org/`

Overview of Data Collections Available to use as a Text Corpus

Political Speeches

- EuroParl: http://www.talkofeurope.eu/data/
- UK: https://www.hansard-corpus.org/
- US Congress: https://www.congress.gov/

Overview of Data Collections Available to use as a Text Corpus

Political Speeches

- EuroParl: http://www.talkofeurope.eu/data/
- UK: https://www.hansard-corpus.org/
- US Congress: https://www.congress.gov/
- German Officials Political Speeches: https://politische-reden.eu/#data

Overview of Data Collections Available to use as a Text Corpus

Political Speeches

- EuroParl: http://www.talkofeurope.eu/data/
- UK: https://www.hansard-corpus.org/
- US Congress: https://www.congress.gov/
- German Officials Political Speeches: https://politische-reden.eu/#data
- United Nations General Debate Corpus - https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0TJX8Y

Overview of Data Collections Available to use as a Text Corpus

- Party Manifestos via Manifesto Project
  https://manifestoproject.wzb.eu/

Overview of Data Collections Available to use as a Text Corpus

- Party Manifestos via Manifesto Project
  `https://manifestoproject.wzb.eu/`
- US Presidential Proclamations, Memoranda:
  `http://www.presidency.ucsb.edu/`

Overview of Data Collections Available to use as a Text Corpus

- Party Manifestos via Manifesto Project
  `https://manifestoproject.wzb.eu/`
- US Presidential Proclamations, Memoranda:
  `http://www.presidency.ucsb.edu/`
- Social Media Datasets

Overview of Data Collections Available to use as a Text Corpus

- Party Manifestos via Manifesto Project
  `https://manifestoproject.wzb.eu/`
- US Presidential Proclamations, Memoranda:
  `http://www.presidency.ucsb.edu/`
- Social Media Datasets
  - Brexit social media dataset
    `https://aifb-ls3-kos.aifb.kit.edu/projects/BreXLiMe/`

Overview of Data Collections Available to use as a Text Corpus

- Party Manifestos via Manifesto Project `https://manifestoproject.wzb.eu/`
- US Presidential Proclamations, Memoranda: `http://www.presidency.ucsb.edu/`
- Social Media Datasets
  - Brexit social media dataset `https://aifb-ls3-kos.aifb.kit.edu/projects/BreXLiMe/`
  - Reddit dataset `https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/`

Overview of Data Collections Available to use as a Text Corpus

- Party Manifestos via Manifesto Project https://manifestoproject.wzb.eu/
- US Presidential Proclamations, Memoranda: http://www.presidency.ucsb.edu/
- Social Media Datasets
  - Brexit social media dataset https://aifb-ls3-kos.aifb.kit.edu/projects/BreXLiMe/
  - Reddit dataset https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/
  - Trump Tweets Archive https://www.thetrumparchive.com/

Overview of Data Collections Available to use as a Text Corpus

Using an API:

An API is a tool that provides you with data, when querying a specific url. For example, you can "ask" Twitter for all the tweets with a certain hashtag. Important - always read the documentation and the limit of an API.

Preparing Text for Text Analysis: **Basic Pre-processig Steps**

Usual first steps

- Lower-case your words (Politics and politics)

Preparing Text for Text Analysis: **Basic Pre-processig Steps**

Usual first steps

- Lower-case your words (Politics and politics)
- Remove punctuation

Preparing Text for Text Analysis: **Basic Pre-processig Steps**

Usual first steps

- Lower-case your words (Politics and politics)
- Remove punctuation
- Remove numbers

Preparing Text for Text Analysis: **Basic Pre-processig Steps**

Usual first steps

- Lower-case your words (Politics and politics)
- Remove punctuation
- Remove numbers
- Remove stopwords: function words that do not convey meaning but primarily serve grammatical function

Common Stopwords in English

i, me, my, myself, we, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, don, should, now

Preparing Text for Text Analysis: **Adding to our NLP Pipeline - Tokenization**

NLP Pipeline:

- Lower-case your words (Politics and politics)
- Remove punctuation
- Remove numbers
- Remove stopwords
- **Tokenization**

Separating single words, starting from a string (which could be a document, a sentence, a tweet) $\rightarrow$ they will become tokens.

Preparing Text for Text Analysis: **Adding to our NLP Pipeline - Tokenization**

"I don't know whether or not [the wall] is part of a D.A.C.A. equation"

1) Naive Approach: Split on White Spaces

I
don't $\leftarrow$ dont ? do not? don't?
know
...
D.A.C.A. $\leftarrow$ DACA? D.A.C.A.? Deferred Action for Childhood Arrivals

Tokenization

- Greys Anatomy - one token or two?
- JAY Z - one token or two?
- JLO - one token or two?
- IPhone 12 - one token or two?

Tokenization

1) Naive Approach: Split on White Spaces

2) Rule-based, language specific tools $\rightarrow$ we will use the NLTK tokenizer.

Next Class:

- Lemmatization and Stemming

Next Class:

- Lemmatization and Stemming
- Stemming

Next Class:

- Lemmatization and Stemming
- Stemming
- Pos-Tagging

Next Class:

- Lemmatization and Stemming
- Stemming
- Pos-Tagging
- Named Enitity Recognition

Next Class:

- Lemmatization and Stemming
- Stemming
- Pos-Tagging
- Named Enitity Recognition
- Dictionaries

Next Class:

- Lemmatization and Stemming
- Stemming
- Pos-Tagging
- Named Enitity Recognition
- Dictionaries
- The Power of Counting Words

References

- Bubeck, J, Elshehawy, Ashrakat, Marinov, Nikolay, and Federico Nanni, 2021. How the US Electoral Cycle Affects Elections Around the World. Working Paper
- Grimmer, J. and Stewart, B.M., 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis, 21*(3), pp.267-297.
- Monroe, Burt and Phil Schrodt, 2008. Introduction to the Special Issue: The Statistical Analysis of Political Text. *Political Analysis 16,* 4, 351-355
- Saeltzer, Marius (2020): Finding the Bird's Wings: Dimensions of Factional Conflict on Twitter. *Party Politics* (forthcoming). DOI: 10.1177/1354068820957960
- Wilkerson, J. and Casas, A., 2017. Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annual Review of Political Science, 20,* pp.529-544.