

Exploratory Data Analysis and Data Visualization

Indira Sen

University of Konstanz
Research Projects in Computational Studies of Social Phenomena

Agenda

- ❖ Recap: Datasets for Projects
- ❖ Today:
 - Exploratory Data Analysis
 - Data Visualization
 - Hands-on EDA and Data Viz
 - Activity

recap

New data, new methods, new challenges...

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES |



TECHNOLOGY

Everything We Know About Facebook's Secret Mood-Manipulation Experiment

It was probably legal. But was it ethical?

By Robinson Meyer

ive-scale ocial

TECH • SOCIAL NETWORKING

The Author of a Controversial Facebook Study Says He's 'Sorry'

2 MINUTE READ

BY STEPHANIE BURNETT X JUNE 30, 2014 2:44 AM EDT



ne of the authors of a controversial Facebook study into emotional

Datasets

Suggested Ideas

1. **Reddit Posts and Comments about Politicians:** How do people discuss female politicians?
2. **Stance Detection Benchmark:** How well do current computational models perform at detecting stance towards different entities and topics?
3. **Annotator (Dis)agreement:** Characterizing differences in annotator perspective for subjective constructs
4. **Lost in Simplification? English vs. Simple Wikipedia:** How does content and framing diverge in Simplified Wikipedia?
5. **X (Twitter) and Reddit discussion about Football:** How do people talk about non-white players?
6. **One Day on Twitter:** What goes on in one day on Twitter?

Suggested Ideas

1. **Reddit Posts and Comments about Politicians:** How do people discuss female politicians?
2. **Stance Detection Benchmark:** How well do current computational models perform at detecting stance towards different entities and topics?
3. **Annotator (Dis)agreement:** Characterizing differences in annotator perspective for subjective constructs
4. **Lost in Simplification? English vs. Simple Wikipedia:** How does content and framing diverge in Simplified Wikipedia?
5. **X (Twitter) and Reddit discussion about Football:** How do people talk about non-white players?
6. **One Day on Twitter:** What goes on in one day on Twitter?

1. Reddit Posts and Comments about Politicians

Full data:

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YWRXEP&version=1.0>

Quantifying gender biases towards politicians on Reddit

Sara Marjanovic, Karolina Stańczak , Isabelle Augenstein

Published: October 26, 2022 • <https://doi.org/10.1371/journal.pone.0274317>

Article	Authors	Metrics	Comments	Media Coverage	Peer Review
⌵					

Abstract

- 1 Introduction
 - 2 Data
 - 3 Analyses
 - 4 Results
 - 5 Discussion
 - 6 Conclusion
 - Supporting information
 - Acknowledgments
 - References
-
- Reader Comments
 - Figures

Abstract

Despite attempts to increase gender parity in politics, global efforts have struggled to ensure equal female representation. This is likely tied to implicit gender biases against women in authority. In this work, we present a comprehensive study of gender biases that appear in online political discussion. To this end, we collect 10 million comments on Reddit in conversations *about* male and female politicians, which enables an exhaustive study of automatic gender bias detection. We address not only misogynistic language, but also other manifestations of bias, like benevolent sexism in the form of seemingly positive sentiment and dominance attributed to female politicians, or differences in descriptor attribution. Finally, we conduct a multi-faceted study of gender bias towards politicians investigating both linguistic and extra-linguistic cues. We assess 5 different types of gender bias, evaluating coverage, combinatorial, nominal, sentimental and lexical biases extant in social media language and discourse. Overall, we find that, contrary to previous research, coverage and sentiment biases suggest equal public interest in female politicians. Rather than overt hostile or benevolent sexism, the results of the nominal and lexical analyses suggest this interest is not as professional or respectful as that expressed about male politicians. Female politicians are often named by their first names and are described in relation to their body, clothing, or family; this is a treatment that is not similarly extended to men. On the now banned far-right subreddits, this disparity is greatest, though differences in gender biases still appear in the right and left-leaning subreddits. We release the curated dataset to the public for future studies.

[Quantifying gender biases towards politicians on Reddit](#)

2. Stance Detection Benchmark

One popular stance dataset:

<https://drive.google.com/drive/u/0/folders/1so8lY1XKpnhUtTvb15edEz6aeHt7CSuh>

How well do OTS models, especially LLM-based techniques identify stance towards targets?

Stance Detection Datasets

[P-Stance: A Large Dataset for Stance Detection in Political Domain](#)

Authors	Target(s)	Source	Type	Size
Mohammad et al. (2016a)	Atheism, Climate change is a real concern, Feminist movement, Hillary Clinton, Legalization of abortion, Donald Trump	Twitter	Target-specific	4,870
Ferreira and Vlachos (2016)	Various claims	News articles	Claim-based	2,595
Sobhani et al. (2017)	Trump-Clinton, Trump-Cruz, Clinton-Sanders	Twitter	Multi-target	4,455
Derczynski et al. (2017)	Various claims	Twitter	Claim-based	5,568
Swami et al. (2018)	Demonetisation in India in 2016	Twitter	Target-specific	3,545
Gorrell et al. (2019)	Various claims	Twitter, Reddit	Claim-based	8,574
Conforti et al. (2020b)	Merger of companies: Cigna-Express Scripts, Aetna-Humana, CVS-Aetna, Anthem-Cigna, Disney-Fox	Twitter	Target-specific	51,284
Conforti et al. (2020a)	Merger of companies: Cigna-Express Scripts, Aetna-Humana, CVS-Aetna, Anthem-Cigna	News articles	Target-specific	3,291
P-STANCE	Donald Trump, Joe Biden, Bernie Sanders	Twitter	Target-specific	21,574

Table 2: Comparison of English stance detection datasets.

3. Annotator (Dis)agreement

Why is it important to study annotator disagreement?

- Annotating data is *interpretive*
- People's perceptions of constructs (especially toxicity, hate speech) is affected by their backgrounds
 - Demographics
 - Lived experience (e.g., if they have faced harassment in the past or not)
- It is important our computational measurement models are 'representative'

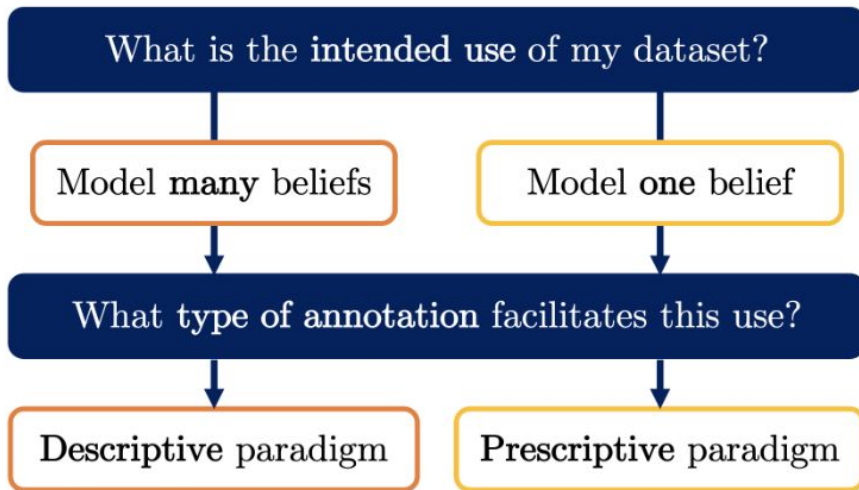


Figure 1: Two key questions for dataset creators.

[Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks](#)

We know that annotators perceive some constructs differently.

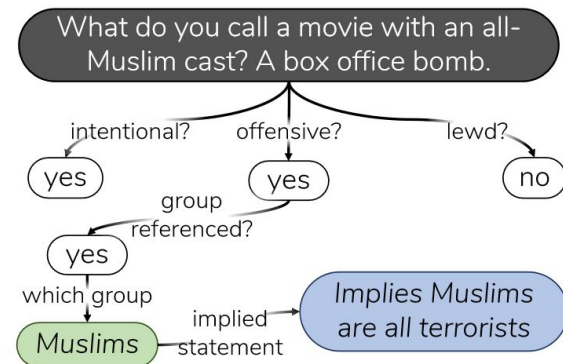
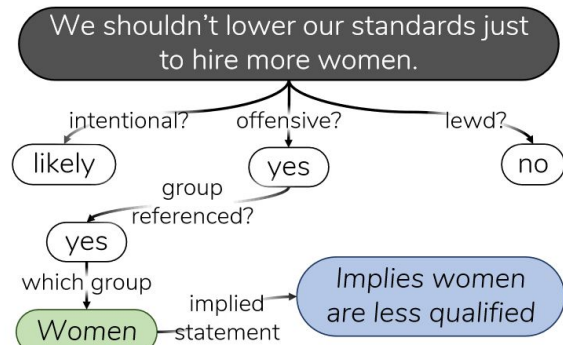
We know that annotators perceive some constructs differently.

But, what characteristics of the subjective instance drive this disagreement?

Datasets

Paper	Dataset	NLP Task	Which demographics?
Social Bias Frames	https://huggingface.co/datasets/social_bias_frames	Offensiveness, lewdness, sexual content	Gender, minority, political leaning,
Annotators with Attitudes	Contact authors	Toxicity (Anti-black toxicity)	Race, gender, political leaning, other beliefs (empathy, altruism, attitudes towards free speech, traditionalism)
NLPPositionality	http://nlpositionality.cs.washington.edu/	Social acceptability, hate speech	Gender, age, religion, country (residence, longest), education, ethnicity, native language
Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application	https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech	Hate speech	Age, disability, religion, sexuality, race, origin, gender
Designing Toxic Content Classification for a Diversity of Perspectives	https://data.esrg.stanford.edu/study/toxicity-perspectives (encrypted, need to contact authors)	Toxicity	Gender, age, race/ethnicity, LGBTQ+ status, Religion importance, political attitude, parental status
POPQUORN	https://github.com/Jiaxin-Pei/Potato-Prolific-Dataset/tree/main/dataset	Offensiveness, politeness, <i>email writing, question answering</i>	Gender, age, race, education
DICES Dataset: Diversity in Conversational AI Evaluation for Safety	https://github.com/google-research-datasets/dices-dataset/	Safety risk	Race, gender
Don't Take It Personally: Analyzing Gender and Age Differences in Ratings of Online Humor	No link to data	Humor and offense	Age, gender

Social Bias Frames



category_type	category	count	percentage
gender	woman	74337	51.98%
gender	man	68661	48.02%
race	white	115506	83.43%
race	hisp	8905	6.43%
race	asian	8597	6.21%
race	black	5444	3.93%
mixed	white man	57272	39.59%
mixed	white woman	58227	40.25%
mixed	black man	6	0.00%
mixed	black woman	5435	3.76%
mixed	asian man	5049	3.49%
mixed	asian woman	3548	2.45%
mixed	hisp man	3667	2.54%
mixed	hisp woman	5234	3.62%

Today: Introduction to Data Analysis

Various ways to analyze data

- our focus is on digital trace data
 - e.g. community detection, topic detection, sentiments
- but in general much more options related to
 - image analysis, multimedia data
 - timelines
 - and more

Data Preprocessing and Analysis

- sometimes, there isn't a clear separation between the two steps and preprocessing and analysis may go hand in hand
- steps may range from text cleaning (parsing the data, removing stopwords) to complex analysis (create text networks or get stance of the post)

Importance of Data Cleaning

PA

Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It

Matthew J. Denny¹ and Arthur Spirling²

¹ 203 Pond Lab, Pennsylvania State University, University Park, PA 16802, USA. Email: mdenny@psu.edu

² Office 405, 19 West 4th St., New York University, New York, NY 10012, USA. Email: arthur.spirling@nyu.edu

Abstract

Despite the popularity of unsupervised techniques for political science text-as-data research, the implications of preprocessing decisions in this domain have received scant systematic attention. As we show, such decisions have profound effects on the results of real models for real data. Substantive theory is typically too vague to be of use for feature selection, and that the supervisor is not necessarily a helpful source of advice. To aid researchers working in unsupervised settings, we develop a statistical procedure and software that examines the sensitivity of findings under alternate preprocessing regimes. This approach complements a researcher's substantive understanding of a problem by characterizing how the variability changes in preprocessing choices may induce when analyzing a dataset. In making scholars aware of the degree to which their results are likely to be sensitive to preprocessing decisions, it aids replication efforts.



[Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it](#)

[The impact of preprocessing on text classification.](#)

The impact of preprocessing on text classification

Alper Kursat Uysal  , Serkan Gunal 

Show more 

+ Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.ipm.2013.08.006> 

[Get rights and content](#) 

Highlights

- The impact of preprocessing on [text classification](#) in terms of various aspects is extensively examined.
- Experiments are conducted on two different domains and in two different languages.
- Choosing appropriate preprocessing tasks may improve classification

Data Cleaning

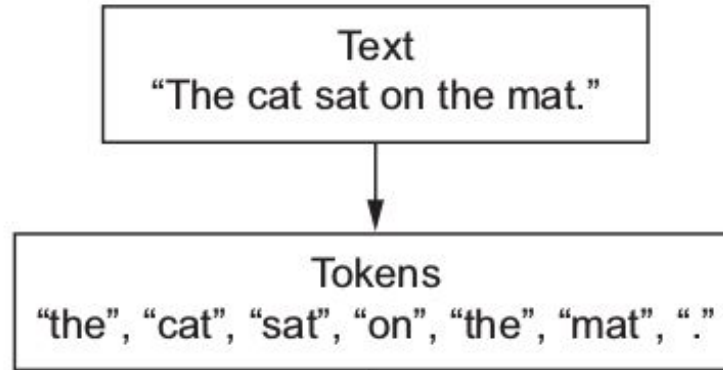
1. Remove unnecessary items from our dataset
 - a. function words ('the', 'on', etc)
2. Maintain order and consistency.
3. Standardization, e.g., time formats
4. Deduplication: not just exact match but also 'near-duplicates'

Data Cleaning: Typical Steps

- tokenization
- remove stopwords
- Stemming / Lemmatization
- remove numbers
- remove headers and footers
- remove rare words
- Beyond Text:
 - dropping or imputing missing values
 - dropping columns with missing values

Data Cleaning: Typical Steps

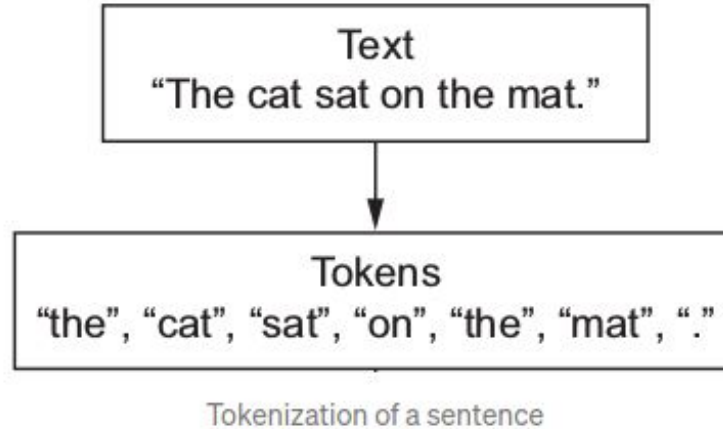
- **tokenization**
- remove stopwords
- Stemming ,
- remove nurl
- remove hec
- remove rar
- Beyond Tex
 - dro
 - dro



Tokenization of a sentence

Data Cleaning: Typical Steps

- tokenization: also digital trace data-based specific tokens such as @mentions, #hashtags, links, emojis, etc
- remove stopwords
- Stemming /
- remove ngrams
- remove hashtags
- remove rare terms
- Beyond Text
 - documents
 - documents



<https://github.com/jaredks/tweetokenize>

Data Cleaning: Typical Steps

- tokenization
- remove stopwords
- **Stemming / Lemmatization**
- remove punctuation
- remove numbers
- remove special characters
- Beyond

Rule		Example	
SSSES	→ SS	caresses	→ caress
IES	→ I	ponies	→ poni
SS	→ SS	caress	→ caress
S	→	cats	→ cat

Examples of stemming

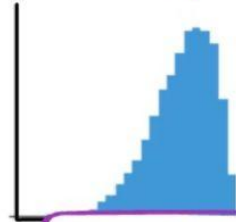
Exploratory Data Analysis

EDA (for Digital Trace Data)

- Summary statistics (mean, median, mode, quantiles, min, max)
- Computing distributions of various variables
- Creating new variables from existing ones
 - Length of content
 - # entities mentioned in the content (using NER)
 - # links in the content (using regex)
 - (and their domains...)
- Bivariate interactions:
 - Length of content across different user groups
-

EDA

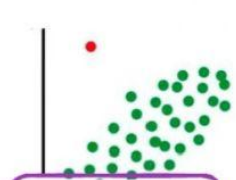
Exploratory Data Analysis



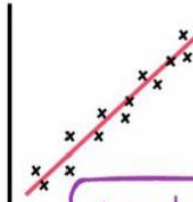
Data distribution

	F	G	H	I	J
A	0.620576	0.140053	1.352728	NaN	0.808078
B	NaN	0.526829	NaN	NaN	0.170902
C	NaN	0.458827	1.406713	0.071119	NaN
D	NaN	2.307197	NaN	NaN	NaN
E	0.203402	0.259913	NaN	0.505811	1.516755

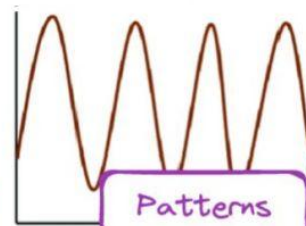
Missing data



Outliers



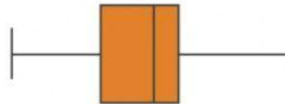
Correlation



Patterns

```
Cust_No      int64
Cust_Name    object
Product_id   int64
Product_cost float64
Purchase_Date datetime64[ns]
dtype: object
```

Data types



Data visualization

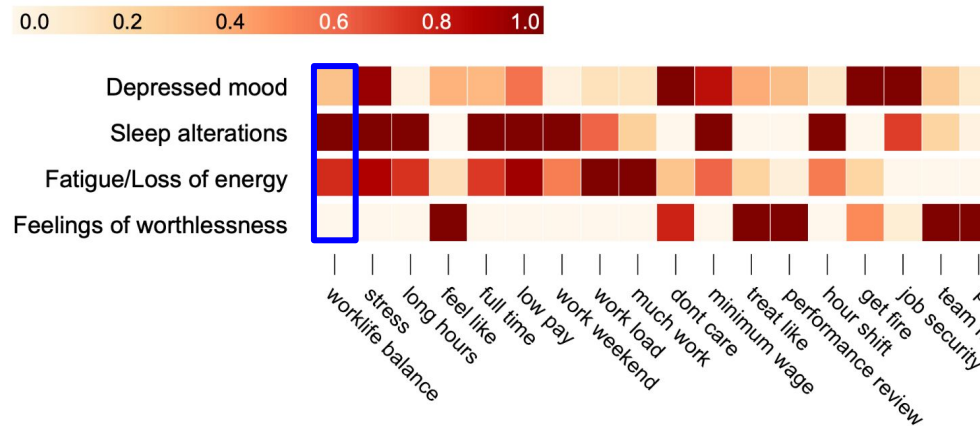


Data quality

<https://www.markovm.com/blog/exploratory-data-analysis>

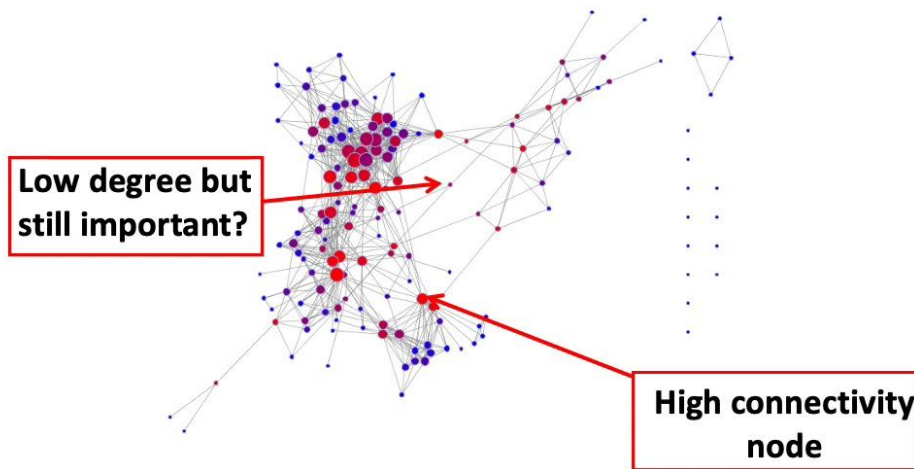
EDA for different types of data: Text

- Tokenize the text into words or n-grams (sequences of words).
- Word Frequency Analysis:
 - Calculate the frequency of each word in the corpus.
 - Term Frequency-Inverse Document Frequency (TF-IDF):
- Calculate TF-IDF scores to identify important terms in the corpus.
- Analyze the frequency and distribution of **n-grams (e.g., bigrams, trigrams)** to capture phrases or collocations.
- Topic Modeling



EDA for different types of data: Network

- **Degree Distribution:**
 - Plot the degree distribution to understand how node degrees are distributed in the network. This can help identify important nodes or hubs.
- **Centrality Measures:**
 - degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality
 - Use centrality scores on the network to identify key nodes.



Data Visualization

General Data Visualization

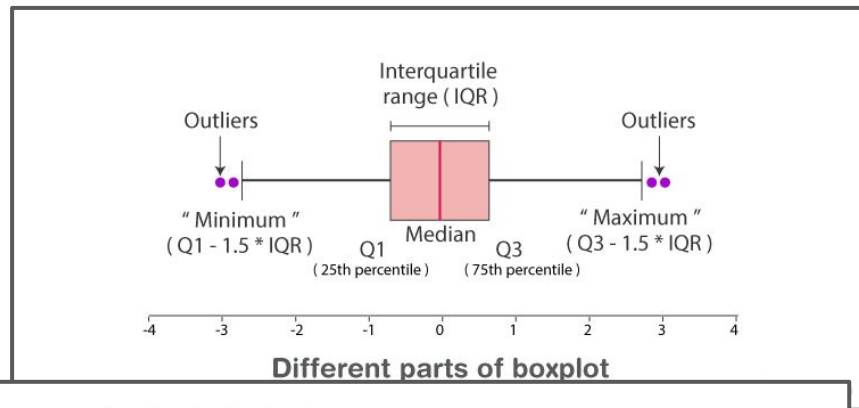
Histograms

Bar Charts

Box Plots

Pie Charts

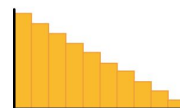
...



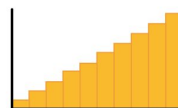
Symmetric (normal) vs skewed and uniform distributions



Normal distribution
(unimodal, symmetric,
the "bell curve")



**Right-skewed
distribution**
(Positively-skewed)



**Left-skewed
distribution**
(Negatively-skewed)

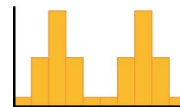


Uniform distribution
(equal spread,
no peaks)

Unimodal vs bimodal distributions



Normal distribution
(unimodal, symmetric,
the "bell curve")



**Symmetric bimodal
distribution**
(two modes)



**Non-symmetric
bimodal distribution**
(two modes)

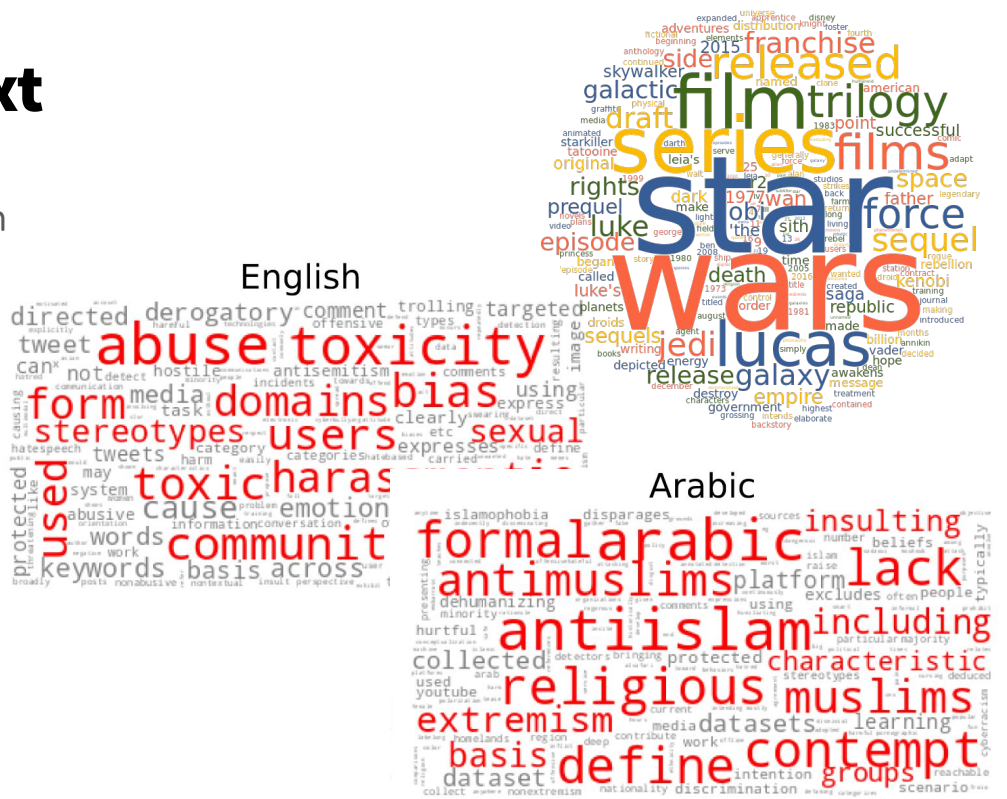
Data Visualization for Text

- Word clouds: easy to make, but often too vague.



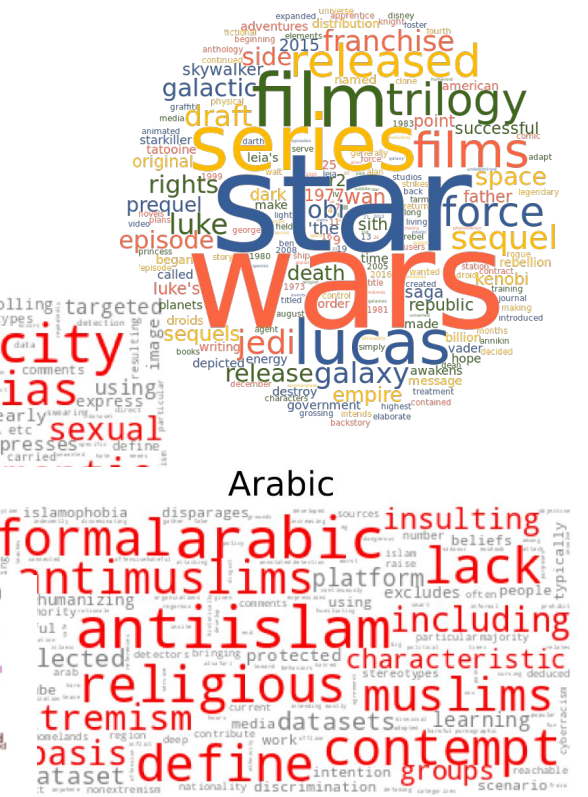
Data Visualization for Text

- Word clouds: easy to make, but often too vague.
- Variants of word clouds
 - Accentuated word clouds



Data Visualization for Text

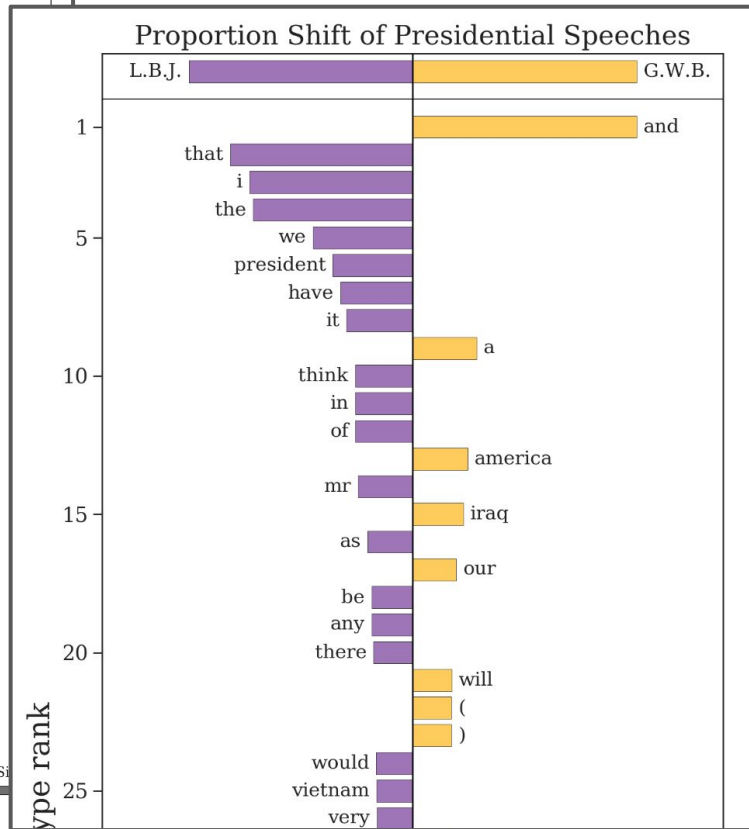
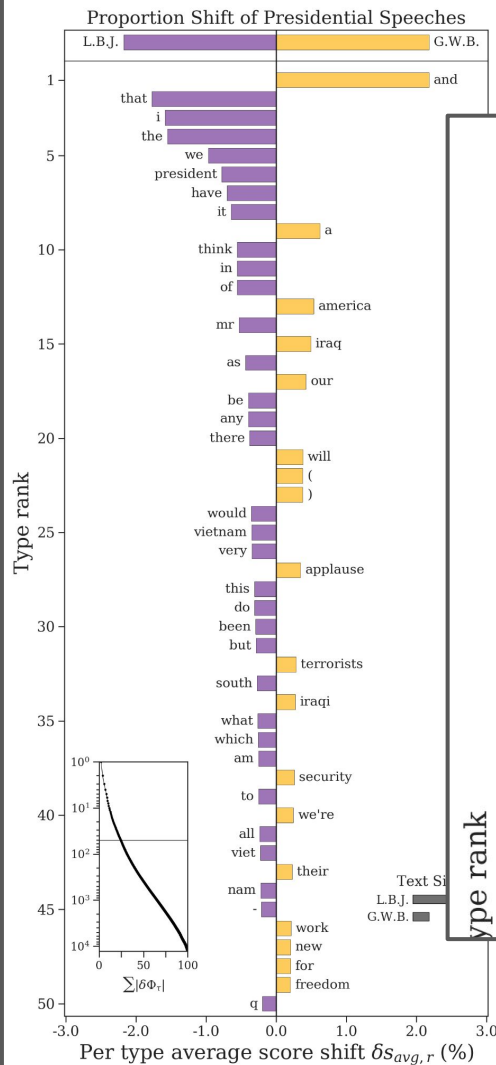
- Word clouds: easy to make, but often too vague.
- Variants of word clouds
 - Accentuated word clouds
 - Semantically grouped word clouds



Word Shift Graphs

- Used for visualizing **pairwise comparisons** between texts through word shifts
- Word shifts extract which words contribute to a difference between two texts
- Different options for computing text comparison, e.g., Shannon entropy, Kullback-Leibler divergence

<https://shifterator.readthedocs.io/en/latest/>



General tips, especially for your reports and presentations

- All axes should have labels
- Captions should be self-contained
- The text in the figure should be as big as the text in your paper
- Using colors to convey information is nice, but also good to use markers or textures for this
- [Optional] Also keep in mind accessibility:
 - <https://towardsdatascience.com/two-simple-steps-to-create-colorblind-friendly-data-visualizations-2ed781a167ec>

Now, let's explore some of these datasets

- Download and open the notebook:
- Create one exploratory data visualization based on one or more datasets
- Examples
 - Who are the most frequently mentioned politicians in the Reddit Politicians' dataset?
 - How is the conversation in r/the_Donald different from other subreddits (try word shift graphs for this)
 -

Further Readings and Resources

Exploratory Data Analysis

Data Visualization:

1. Semantically grouped word clouds: Hearst et al., [An Evaluation of Semantically Grouped Word Cloud Designs](#)
2. Word shift graphs: https://ryanjgallagher.github.io/code/word_shift/overview
3. Yong-Yeol (YY) Ahn's detailed data viz course: <https://yyahn.com/dviz-course/>
4. Visualizing text data: <https://textvis.lnu.se/>