Intro to Computational Text Analysis

D-Lab training workshop

Caroline Le Pennec-Çaldichoury[1]

[1]Many thanks to the previous instructors who shared their teaching material: Geoff Bacon, Ben Gebre-Medhin and Laura Nelson.

Plan for today

- Introductions:
  - Who are you? (name, department, school)
  - Why are you here? (specific project? favorite programming language?)
- Some theory: how to use text as data
- Some practice: text pre-processing
- Next two sessions: learn how to implement simple methods and interpret the results
- Additional resources:
  - CTAWG (ask Caroline to be added to mailing list if interested)
  - Info 256: Applied NLP course (David Bamman)

This presentation

- Growing interest in using text as data in social sciences and humanities - and many other fields.
- In this presentation:
  - discuss what makes text so special... or not
  - overview of relevant statistical methods
  - examples of applications
- References:
  - Grimmer and Stewart (2013)
  - Gentzkow, Shapiro and Taddy (2017)

Outline

1. Introduction
2. Conceptual framework
3. Supervised methods
   - Dictionary methods
   - Text regression
   - Optional: Inverse regression
4. Unsupervised methods
   - PCA
   - Topic model
5. Conclusion

Outline

1. Introduction
2. Conceptual framework
3. Supervised methods
   - Dictionary methods
   - Text regression
   - Optional: Inverse regression
4. Unsupervised methods
   - PCA
   - Topic model
5. Conclusion

Framework

- Corpus of *documents*
- Objective: map raw text of each document i to some attribute v_i
  - predict customer satisfaction from product reviews
  - predict hate speech from reddit comments
- The estimated attribute $\hat{v}_i$ can then be used for descriptive or causal analysis:
  - causal effect of racial animus predicted from Google search data on vote for Obama (Stephens-Davidowitz (2014))

A simple representation of text

- Need to reduce the complexity of raw text → turn into numbers
- Pre-processing to reduce the size of *vocabulary*:
  - remove stopwords
  - remove special characters
  - stemming/lemmatization
  - exclude words too frequent or too infrequent
- Bag-of-words representation:
  - document i as an array of (unordered) token frequencies $\mathbf{c}_i$
  - a token can be word, n-gram, phrase, etc
  - corpus as a *document-term matrix*

Document-term matrix

- Example: tweets from Trump during the 2016 campaign (7,300 tweets)

	00	000	007llisav	00phdhsb37	00pm	01	02	03	08	10	...	yrs	yuge	yup	zero	zilch	zinger	zogby	zones	zucker	zuckerman
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

10 rows × 4768 columns

$\rightarrow$ Too simple? Probably, but still relevant.

Alternative representations

- Binary or weighted bag-of-words
  - Tf-Idf: more weight to document-specific words
- Word embeddings:
  - each word represented by a latent vector informative about context
  - each document is a function of word vectors (e.g mean) or represented by a document-level context vector

Outline

1 Introduction

2 Conceptual framework

3 Supervised methods
- Dictionary methods
- Text regression
- Optional: Inverse regression

4 Unsupervised methods
- PCA
- Topic model

5 Conclusion

Supervised methods

- Attribute v_i is known i.e the researcher knows what she is looking for
  - E.g: sentiment analysis
- Most common supervised task: **classification**
  - is document *i* positive or negative?
- Scaling on a known 'intensity' scale:
  - how positive is document *i* from -1 to 1?

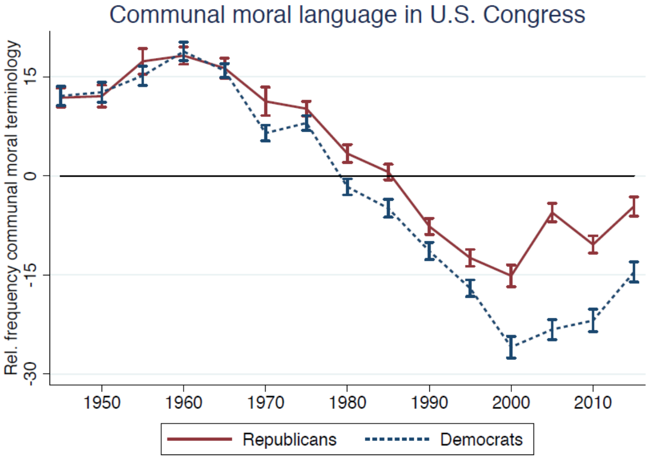
Dictionary methods

- Most widely used, very simple and intuitive
- v_i is a function of the frequency of key words from dictionary $\mathbf{d}$ in document i:

$$v_i = f(\mathbf{c}_i'\mathbf{d})$$

- Enke (2018): use the Moral Foundation Dictionary to study the supply and demand of moral values in US presidential elections
  - relative frequency of words related to 'communal' moral values vs 'universal' ones
  - study trends of communal values over time in political discourse and correlation with county-level vote shares at primary elections

Ben Enke (2018)



Source: Enke, Benjamin. Moral values and voting: Trump and beyond. No. w24268. National Bureau of Economic Research, 2018. Gentzkow (2016)

Dictionary methods

- Limitation: need a **good** dictionary
- Context is important
- Mixing dictionary approach with word embeddings:
  - learn vector representations of key words or entire concepts
  - compare document representation to dictionary representation
  - Garten et al. (2016) use pre-trained distributed representations of words in MFD and compute the cosine similarity between moral category-level vector and document-level vector
  - they find that vector representations give a cleaner signal of moral rhetoric in tweets and political speeches than simple key word count

Text regression

- Standard prediction approach
- Requires a training set: $\mathbf{v}^{train}$ and $\mathbf{C}^{train}$
  $\rightarrow$ dataset already available or need humans to map the training documents into v
- Discriminative model: predict v_i conditional on observed $\mathbf{c}_i$
- Estimate a regression model:

$$E[v_i \,|\, \mathbf{c}_i] = f(\eta_i)$$

  - $\eta_i = \alpha + \mathbf{c}_i' \boldsymbol{\beta}$ is a linear function of word frequencies
  - $f(.)$ can be linear or logistic if v is categorical

- Get fitted values $\hat{v}_i = \hat{f}(\mathbf{c}_i)$ from the training set
- Use test set, $\mathbf{v}^{test}$ and $\mathbf{C}^{test}$, to validate the estimated model: measure accuracy between $\hat{v}_i$ and v_i

Penalized estimators

- How is it different from using non-text data?
  - The high-dimensionality!
- Finite sample and sparsity of the DTM can lead to over-fitting and severe bias because of infrequent words
- Common solution: **penalized** negative log-likelihood minimization

Penalized estimators

- With L1 (Lasso) regularization, $(\alpha, \boldsymbol{\beta})$ is the solution to:

$$min \left\{ l(\alpha, \boldsymbol{\beta}) + n\lambda \sum_j |\beta_j| \right\}$$

  - $l(\alpha, \boldsymbol{\beta}) = -\sum_i (\eta_i v_i - log(1 + e^{\eta_i}))$ for $v_i \in (0, 1)$ if $f(.)$ is binomial logistic
  - λ penalizes deviations from zero $\rightarrow$ sparse solution

- Intuitively: noisy coefficients that are too big are shrunk to zero so this method reduces dimensionality by selecting only 'good' covariates (tokens)

Penalized estimators

- How to choose λ?
  - Cross-validation (K-fold)
  - AIC or other criterion
- In practice: easy to implement with gmlnet in R or scikit-learn in Python
  - E.g. LogisticRegressionCV (sklearn.linear_model) implements both L1 and L2 penalty regularization if using linear solver

Generative text models

- Text regression estimates a model of $p[v_i \mid \mathbf{c}_i]$ but a more natural model of speech is $\mathbf{p}[\mathbf{c}_i \mid v_i]$
  - given attribute v_i, individual i is more likely to choose some words over others - not the reverse!
- Generative text models:
  - Naive Bayes
  - Inverse regression

Multinomial Inverse Regression

- Why 'inverse' regression?
  - Use training set to estimate token *loadings* - the sensitivity of each token to observed attribute v_i
  - Then use estimated loadings to project each document onto the v_i space to obtain $\tilde{v}_i$

Multinomial Inverse Regression

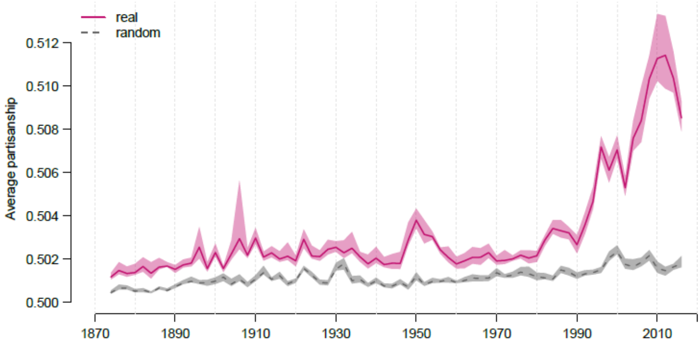
- $\tilde{v}_i$ is a one-dimensional 'summary' of high-dimensional $\mathbf{c}_i$
  - can be used subsequently for prediction, classification, scaling, etc.
  - allows to control for other attributes - more appropriate for inference
- Model estimated with distributed multinomial regression (Poisson approximation) and Gamma-lasso penalization in textir[2] R package.
  More details
- Gentzkow et al. (2016): use an inverse penalized regression approach to measure polarization in political speech from Congressional Records over time
  - estimate each phrase sensitivity to observed party affiliation (Rep or Dem) while controlling for additional covariates (state, gender, etc)
  - back-out the probability of 'guessing right' the party of politician i from observing a random spoken phrase

[2] `https://cran.r-project.org/web/packages/textir/textir.pdf`

Gentzkow et al. (2016)

Figure 3: Average Partisanship of Speech, Penalized Estimates

Panel A: Preferred Specification



Source: Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Measuring polarization in high-dimensional data: Method and application to congressional speech. No. w22423. National Bureau of Economic Research, 2016.

Enke (2018)

Other supervised methods

- Polarization/scaling: Wordscore, Pearson Chi-square
- Decision tree methods
  - random forest
  - typically not so useful with high-dimensional data and risk of over-fitting
  - can be combined with methods of dimensionality reduction
- Support Vector Machine
  - for categorical attributes
- Deep learning and neural networks

Outline

1. Introduction
2. Conceptual framework
3. Supervised methods
   - Dictionary methods
   - Text regression
   - Optional: Inverse regression
4. Unsupervised methods
   - PCA
   - Topic model
5. Conclusion

Unsupervised methods

- Attribute v_i is unknown, not explicitly defined
  - exploratory data analysis
  - similarity to other documents
  - scaling on a unknown scale
- Example: compare documents using cosine similarity between vectors of token frequencies
  - measures the *angle* between vectors
  - allows to compare documents of different lengths
- But very noisy because of sparsity of DTM[¡.-¿]
- Here again: the main point is to reduce dimensionality and the question is how

Principal Component Analysis

- PCA decomposition reduces dimensionality while conserving relevant variation
- Single value decomposition of the DTM in orthogonal principal components, where the first component explains more variance then the second, etc.
- Each component is a linear combination of covariates on which each document can be projected to get a *component score*
- Sparse DTM can be replaced by a dense matrix of PC scores in subsequent analysis

LDA topic modelling

- Close to cluster analysis (e.g K-mean)
- Latent Dirichlet Allocation (Blei et al. 2003): generative model where
  - each document is a mixture of topics ($\mathbf{v}_i$)
  - each topic is a mixture of tokens

$$E\left[\frac{\mathbf{c}_i}{m_i}\right] = q_{ij} = v_{i1}\boldsymbol{\theta}_1 + v_{i2}\boldsymbol{\theta}_2 + \ldots + v_{iT}\boldsymbol{\theta}_T$$

  - $\boldsymbol{\theta}_t$ is topic t's probability distribution over words
  - v_{it} is the *weight* of topic t in document i
  - $\boldsymbol{\theta}_t$ and $\mathbf{v}_i$ are generated from a Dirichlet-distributed prior
- Unsupervised: attribute $\mathbf{v}_i$ is a list of weights over T topics where topics are **not determined ex-ante** (only T is)

LDA topic modelling

- LDA can be useful for:
  - dimensionality reduction from sparse DTM to dense matrix of topic weights
  - combined with qualitative analysis, link topics generated without supervision to a known attribute and turn to supervised methods
- Catalinac (2015): use topic model to test for change in campaign strategy under different electoral rules
  - fit an LDA of 69 topics on 7,500 campaign manifestos from candidates at Japanese national elections
  - in-depth qualitative interpretation to identify 'particularistic' vs 'programmatic' topics
  - estimate change in topic prevalence before/after an electoral reform from single-member to multi-member districts
  - find that candidates from the dominant party talk more about programmatic and national security issues when intra-party competition decreases

Data as text?

Other unsupervised methods

- Structural topic modeling (STM)
  - LDA with covariates $\rightarrow$ allows topic content and topic prevalence to vary across document-level characteristics
- Unsupervised scaling (Wordfish)

Conclusion

- Wide range of available methods to treat text as data
- Choosing the 'right' method depends on the research question and the data:
  - is there an explicit attribute I want to map text into?
  - do I have a dictionary or a training set available?
  - technical concerns: sample size, document length, covariates...
- Validation is crucial
  - compare prediction to ground truth
  - robustness to alternative methods
- Understanding the method you apply is even more crucial

References

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.
- Catalinac, Amy. "From pork to policy: The rise of programmatic campaigning in Japanese elections." The Journal of Politics 78.1 (2016): 1-18.
- Draca, Mirko, and Carlo Schwarz. "How Polarized are Citizens? Measuring Ideology from the Ground-Up." (2018).
- Enke, Benjamin. Moral values and voting: Trump and beyond. No. w24268. National Bureau of Economic Research, 2018.
- Garten, Justin, et al. "Morality between the lines: Detecting moral sentiment in text." Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes, New York, NY. Retrieved from http://mortezadehghani.net/wp-content/uploads/morality-lines-detecting. pdf. 2016.
- Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Measuring polarization in high-dimensional data: Method and application to congressional speech. No. w22423. National Bureau of Economic Research, 2016.
- Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. Text as data. No. w23276. National Bureau of Economic Research, 2017.
- Stephens-Davidowitz, Seth. "The cost of racial animus on a black candidate: Evidence using Google search data." Journal of Public Economics 118 (2014): 26-40.
- Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." Political analysis 21.3 (2013): 267-297.
- Taddy, Matt. "Multinomial inverse regression for text analysis." Journal of the American Statistical Association 108.503 (2013): 755-770.
- Wen, Xuerong, and R. Dennis Cook. "Optimal sufficient dimension reduction in regressions with categorical predictors." Journal of Statistical Planning and Inference 137.6 (2007): 1961-1978.

More details on MNIR

- MNIR framework (Taddy 2013): each token frequency is an independent draw from a multinomial distribution

$$c_{ij} \sim \ MN(q_{ij}, m_i)$$

  - c_{ij} is the frequency of token j in document i
  - m_i is the number of tokens in document i
  - $$q_{ij} = \frac{exp(\alpha_j + \phi_j v_i)}{\sum_{k=1}^{W} exp(\alpha_k + \phi_k v_i)}$$
  - ϕ_j is a token *loading* that measures sensitivity to attribute $\mathbf{v}_i$
- Sufficient reduction projection (Cook 2007):

$$\tilde{v}_i = \sum_{w=1}^{W} \phi_j \cdot \frac{c_{ij}}{m_i}$$

is a sufficient reduction for v_i: conditional on $\tilde{v}_i$, v_i is independent from $\mathbf{c}_i$

Data as text?

- LDA is not about adapting an existing statistical model to text - it was specifically designed for text corpora.
- Can it be adapted to 'data treated as text'?
- Draca and Schwarz (2018): use LDA to reveal citizens' ideological types from coded answers to World Value Survey, not text!
  - a feature is a coded position on issue (e.g favor/oppose abortion) and a topic is a probability distribution over all possible issue positions
  - use most common issue positions by topic to identify 'ideological type' (from 2 to 5 types)
  - study prevalence of types over time and across countries

Draca and Schwarz (2018)

2 Type Model
Left
No problem Neighbours: Homosexuals
No problem Neighbours: People different race
No problem Neighbours: People AIDS
No problem Neighbours: Immigrants/foreign workers
Justifiable: divorce
Not Justifiable: someone accepting a bribe
Justifiable: euthanasia
Justifiable: homosexuality
Not Justifiable: claiming government benefits
Proud of nationality
Right
Not Justifiable: someone accepting a bribe
Not Justifiable: suicide
Proud of nationality
Not Justifiable: prostitution
Not Justifiable: avoiding a fare on public transport
Not Justifiable: claiming government benefits
Not Justifiable: cheating on taxes
Not Justifiable: abortion
Not Justifiable: homosexuality
No problem Neighbours: People different race

Source: Draca, Mirko, and Carlo Schwarz. "How Polarized are Citizens? Measuring Ideology from the Ground-Up." (2018).

Back