SOSC 4300/5500: Text Analysis Basics

Han Zhang

Outline

Assignment 1

- After the assignment, you will be more familiar about using statistical models to make predictions with rectangular data (e.g., survey data)
  - With these experiences, we can move on to machine learning/prediction on more complex data types (e.g., type)
  - The data type changes, but algorithms are very much the same
- There is no unique solution; you can solve the problem in tons of different ways
  - Improve through trial and error:
    - try a simple solution, submit to Kaggle, get a baseline
    - think and try ways to improve algorithms (e.g., find better tuning parameters)

Text as data

- Policy documents by governments
- Newspaper articles
- Social media text
- Patent's content
- Scientific articles
- Historical archive
- Else?

A complete workflow (Grimmer and Stewart, 2013)

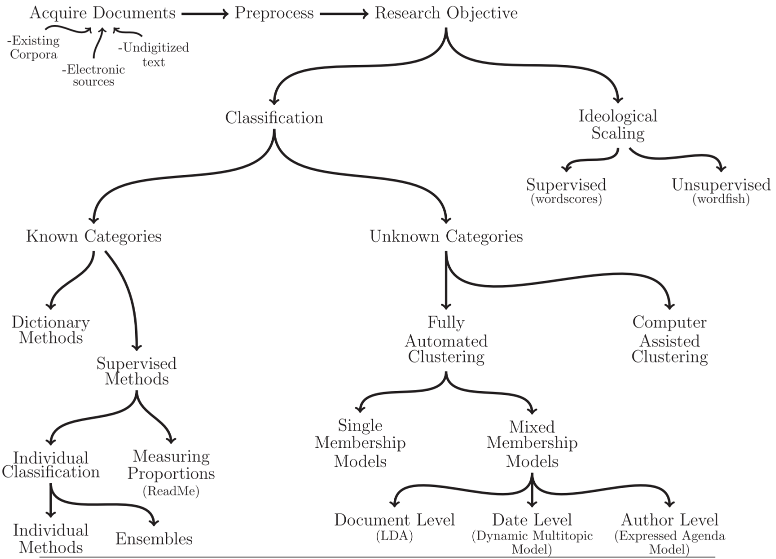


Fig. 1 An overview of text as data methods.

Corpora

- Corpora: a collection of text document
  - A list of text files
  - Or a big CSV/TXT file, with a text column
- Word, term, token (often used interchangeably)
- Vocabulary is all unique words in your corpus

Acquiring texts: electronic sources

- Electronic sources:
    - Electronic searchable newspaper databases
    - Websites; blogs
    - Social media

Acquiring data: electronic sources

- Collecting data from electronic sources
  - Manual approach: download and save each page,
  - Automatic approach: web scraping/crawling
    - More in tutorials
- Now, we assume that we have a cleaned corpora to work with

Acquiring data: undigitized text

- E.g., PDF of scanned books
- Need a lot more work
- Often some OCR (Optical character recognition) is required
  - OCR: Given an image containing texts, predict texts in it.

Stemming and Lemmatization

- Stemming: words with suffixes removed (using set of rules)
  - E.g., "family, families, families, familial" → `famili`
  - Stemming may be problematic, because the not all base form can be obtained by removing suffixes. This is called over-stemming.
    - E.g., university and universe -> univers
- Lemmatization: a more complex version that "seeks to reduce words to their base forms".

word	win	winning	wins	won	winner
stem	win	win	win	won	winner
lemma	win	win	win	win	win

Remove stop words

- Stop words: common words that may not be relevant to your task.
- E.g., a, the, these, not
  - `https://www.aclweb.org/anthology/W18-2502.pdf`
- For certain tasks, such as sentiment analysis, be careful of the stop word list choices!!
  - E.g., if you are doing sentiment analysis, removing word not can be very wrong
    - Cannot distinguish happy vs. not happy

Word segmentation

- For digitized Latin-language families, words have boundary
- But for Chinese, there is no word boundary
- So word segmentation has to be used
  - jieba: easy and quick; precision is relatively low
  - pkuseg and THULAC : better precision; no R version

	Nanjing Yangtze River Bridge
Sequence	南京市长江大桥
Result1	南京 市长 江大桥
	Nanjing mayor Daqiao Jiang
Result2	南京市 长江大桥
	Nanjing City Yangtze River Bridge

From Words to Numbers

- Still, there is no easy way for us to use text as variables
- The next step is to turn a corpora into a matrix X with numeric values
  - Or, turn each document into a numeric vector
- Then we can feed this matrix representation of a corpora into a prediction model
  - regression, tree, forest, SVM, etc.,
- How to turn documents into matrices is one of the most unique aspect of text analysis
- We will first talk about document-term matrix
- In several weeks, we will talk about word embedding

Document-Term Matrix

- Turning corpus into a matrix is usually achieved by obtaining document-term matrix, which rely on word frequencies
- $W :< N \times M >$ matrix; N is the number of documents and M is the size of vocabulary
- W_{im}: the number of times the m-th word occurs in the i-th document.
- The matrix W then can be used as the variables in any ML algorithms

docs \ words	made	because	had	into	get	some	through	next	where	many	irish
t06_kenny_fg	12	11	5	4	8	4	3	4	5	7	10
t05_cowen_ff	9	4	8	5	5	5	14	13	4	9	8
t14_ocaolain_sf	3	3	3	4	7	3	7	2	3	5	6
t01_lenihan_ff	12	1	5	4	2	11	9	16	14	6	9
t11_gormley_green	0	0	0	3	0	2	0	3	1	1	2
t04_morgan_sf	11	8	7	15	8	19	6	5	3	6	6
t12_ryan_green	2	2	3	7	0	3	0	1	6	0	0
t10_quinn_lab	1	4	4	2	8	4	1	0	1	2	0
t07_odonnell_fg	5	4	2	1	5	0	1	1	0	3	0
t09_higgins_lab	2	2	5	4	0	1	0	0	2	0	0
t03_burton_lab	4	8	12	10	5	5	4	5	8	15	8
t13_cuffe_green	1	2	0	0	11	0	16	3	0	3	1
t08_gilmore_lab	4	8	7	4	3	6	4	5	1	2	11
t02_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

Document-Term Matrix: bag-of-words assumption

- Document-term matrix makes the bag-of-words assumption
- Word order do not matter, only presence maters
  - For some problems it's reasonable (e.g., whether an article mentions China or not)
  - For many other problems, it's clearly wrong (e.g., sentiment, "not happy")
- A remedy: n-gram approach
  - Adding concurrent words into vocabulary
- E.g., "I am the instructor"
- With 2-gram
- "I am", "am the", "the instructor" are added into vocabulary

Document-Term Matrix: weighting

- Another common problem: some words appear too often
- Instead of just removing them
- We can add weights to document-term matrix, by penalizing words that appear in too many documents
- This is called inverse-document frequency (idf) score.
    - Low idf score suggest the word is common

$$idf_w = log \frac{\text{number of documnet}}{\text{number of documents in which the term w appears}}$$

- tf-idf matrix: combining document-term matrix (term frequency) and inverse-document frequency matrix together
    - Each cell in W multiplies the corresponding word's idf score

Document-Term matrix: curse of dimensionality

- $W : < N \times M >$ matrix; N is the number of documents and M is the size of vocabulary
- M is ofter larger than N, because:
  - Vocabulary size (on a scale of 10K to 100K) is often larger than the number of documents
  - If n-gram is used, the size of vocabulary can increase exponentially
- Therefore, by its design, document-term matrix suffers from the curse of dimensionality
  - Again, simple linear regression does not work well on high-dimensional data, with more columns than rows
- In two weeks we will introduce something called word embedding
  - It represents documents into low-dimensional matrix

Measurement

- We have discussed the difference between prediction and explanation
  - and have seen an example of using texts for predicting polling
- A third approach is to use prediction for measurement
  - In other words, predictions are used to generate measures for some concepts we are interested in
  - You can then use the "predicted measures" for typical explanatory social science questions
- This prediction-for-measurement approach is arguably more popular no in social sciences than the prediction approach

Two perspectives of predictions

- Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart, *Machine Learning for Social Science: An Agnostic Approach*, Annual Review of Political Science **24** (2021), no. 1, null
- Prediction as a new paradigm of social science research
  - In contrast to explanation
- Prediction for measurement
  - new wine in old bottles
  - but integrates well with traditional explanatory social science research

Texts for measurement

- We will use text analysis to illustrate this prediction-for-measurement perspective
- What can be measured from texts?
    - Sentiments
    - Attitudes
    - Topics
    - Event occurrences
    - Many others

Three eras of survey research

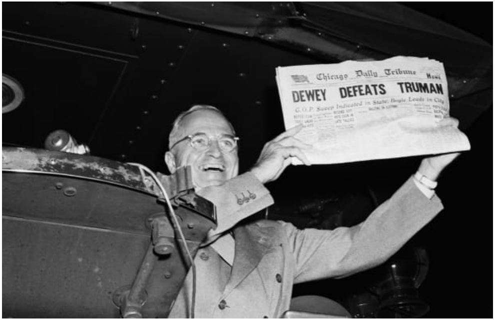
- Matthew Salganik, *Bit by Bit: Social Research in the Digital Age*, Princeton University Press, 2019
- Chapter 3, Table 3.1

Table 3.1: Three Eras of Survey Research Based on Groves (2011)

	Sampling	Interviewing	Data environment
First era	Area probability sampling	Face-to-face	Stand-alone surveys
Second era	Random-digit dialing (RDD) probability sampling	Telephone	Stand-alone surveys
Third era	Non-probability sampling	Computer-administered	Surveys linked to big data sources

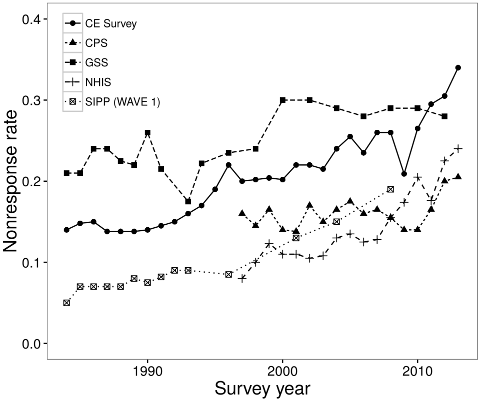
Traditional Surveys: probability sampling

- Why social scientists trust probability sampling more than non-probability sampling?
- Probability sampling on a smaller sample outperforms non-probability sampling on a much larger sample



Problems of traditional surveys

- Rising non-response rate
- Matthew Salganik, *Bit by Bit: Social Research in the Digital Age*, Princeton University Press, 2019
- Chapter 3, Figure 3.6



Problems of traditional surveys

- Trade-off between cost and heterogeneity
- Nicholas Beauchamp, *Predicting and Interpolating State-Level Polls Using Twitter Textual Data*, American Journal of Political Science **61** (2017), no. 2, 490–503
  - some states are poorly polled
  - some days, and sub-state regions, are not polled

Using social media texts to assist election polls

- Nicholas Beauchamp, *Predicting and Interpolating State-Level Polls Using Twitter Textual Data*, American Journal of Political Science **61** (2017), no. 2, 490–503
- Argument: there are many work (especially by computer scientists) stating that social media texts can be used to predict election polls
- But policy researchers still heavily rely on polls
- Can Twitter texts be used to predict vote share for Obama in 2012?

Cleaning Text Data

- Raw data: 40M tweets between Sep 1, 2012 to Nov 4, 2012 (the election day)
- Each tweet contain at least one political words:

  obama, romney, pelosi, reid, biden, mc- connell, cantor, boehner, liberal, liberals, conservative, conservatives, republican, republicans, democrat, democrats, democratic, politics, political, president, election, voter, voters, poll, polls, mayor, governor, congress, congressional, representatives, senate, senator, rep., sen., (D),

- And each tweet was geolocated using keywords (e.g., they contain location words)
- Resulted in 850GB raw data

Turning Text into Variables

- Beauchamp further reduced the data dimension
- By selecting 10,000 most common words
- And calculate the word percentage, w_{kjt}, for word k at state j at day t
  - Number of tweets containing word k for state j at day t
  - Divided by number of total tweets for state j at day t
- End up with 500 MB data; 50 states $\times$ 67 days $\times$ 10,000 variables
- Turning text into variables is the key to most machine learning using text data;
  - More on this shortly

Selecting training and test data

- Training data: for each day t, training data are
  - 3 previous weeks's vote share for Obama based on polling
  - And/or day t's tweets
- Test data:
  - vote share for Obama on day t
  - across 42 days before the election and in 24 states
    - Other states have a few polls; shortcoming of polls if you want to study some detailed patterns

Selecting model

- 9 different models for training
- Simpler regression based models: (M1 - M5)
  - Fixed effects: each state has its' own intercept β_j
  - Time trends: capture time changes (if there is any)
  - Words: each words has its own coefficient
    - but only maintain the coefficient if its p value < 0.001

$$p_{jt} = \beta_j + \tau t + \beta_k w_{kjt} + \epsilon_{kjt}, \quad \text{for } k \text{ in } [1 \ldots 10,000], \tag{1}$$

Selecting models

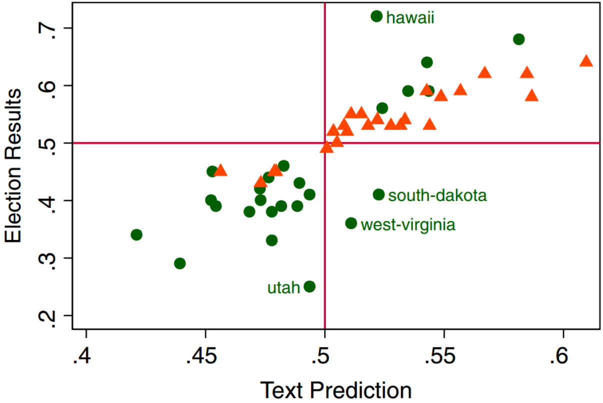
Table 1 Accuracy in Matching Out-of-Sample Text-Predicted Polls to True Polls

	M1	M2	M3	M4	M5	Random Forest	SVM	Elastic Net[c] $\lambda_1 = 0.001$	Elastic Net[c] $\lambda_1 = 0.1$
Twitter text	×		×		×	×	×	×	×
State fixed effects		×	×	×	×	×	×	×	×
Time trend				×	×	×	×	×	×

- The elastic net reduces to LASSO regression since they set $\lambda_2 = 0$
- Random forests, SVM, and elastic net are generally regarded as better than simpler regression models

Prediction Performances: visualization of predictions from M1

- Triangles: states with better polls
- Circles: states with worse polls
- Is this good enough?



Selecting error evaluation criteria

- RMSE
- R^2
  - pooled: variance explained across all cases
  - within: variance explained within states
- And visualization! Simple but powerful

Prediction Performances: quantitative measures

Table 1 Accuracy in Matching Out-of-Sample Text-Predicted Polls to True Polls

								Elastic Net[c]	
	M1	**M2**	**M3**	**M4**	**M5**	**Random Forest**	**SVM**	**$\lambda_1 =$ 0.001**	**$\lambda_1 =$ 0.1**
Twitter text	×		×		×	×	×	×	×
State fixed effects		×	×	×	×	×	×	×	×
Time trend				×	×	×	×	×	×
MAE (smoothed)[a]	1.91	0.60	0.53	0.54	**0.51**	1.53	3.53	0.88	3.76
MAE (real)[a]	2.16	1.38	1.32	1.30	**1.27**	1.81	2.76	1.53	3.21
R^2 Pooled[b]	0.77	0.98	0.98	0.98	**0.98**	0.90	0.19	0.95	0.01
R^2 Within[b]	0.03	0.19	0.36	0.37	**0.40**	0.09	0.07	0.08	0.22

Findings

- Simply using Twitter texts (M1) are worse than simpler regression models (M2, M4)
- Best model (M5) combines Twitter texts and considers the cross-section times-series nature of the data
- Simply taking some machine learning models may not be the best
- But ultimately, draw your conclusions based on prediction evaluation metrics, based on out-of-sample algorithms

A complete workflow (Grimmer and Stewart, 2013)

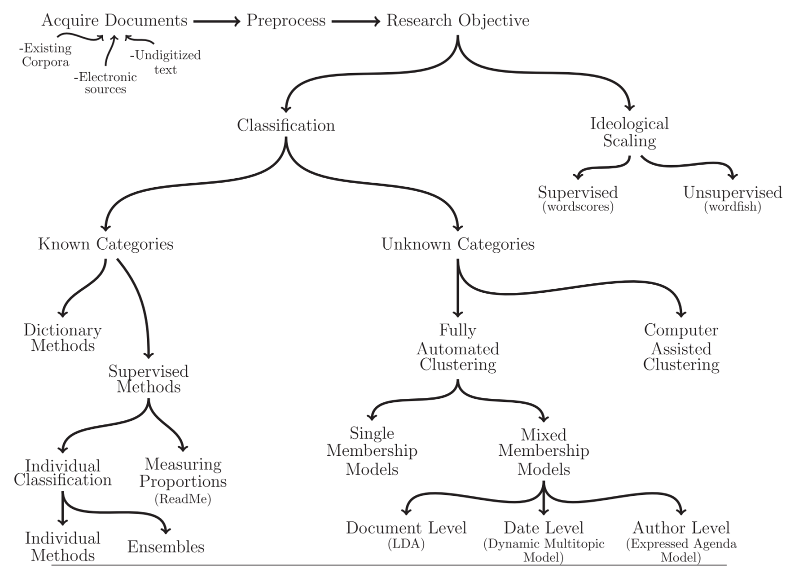


Fig. 1 An overview of text as data methods.

Research objectives

- Supervised: known categories/outcomes
  - Example: sentiment analysis; each document is mapped to either of the three category:
    - positive
    - negative
    - neutral
  - Supervised machine learning:
    - linear/logistic regression
    - decision tree/random forests/boosting trees
    - SVM
    - Neural networks and deep learning (the state of art)
  - Dictionary methods: deterministic
    - Easier than supervised ML; great to start with
    - Typically performs worse

Dictionary method

- The simplest supervised method
  - Often the first step before you jump to some more complex methods
- Dictionary methods relies on curating a list of words
  - Each word is attached with one category
  - Documents with more words in a category is treated as belonging to that category

Dictionary method: one dictionary

- We have collected a bunch of newspaper articles worldwide
- E.g., our research question: whether more foreign news media are reporting more about China after the "Belt and Road Initiative"
- Dictionary: [China, Chine, ...]
- Outcome of each document can be:
  - or, whether a document mentions at least one word in the dictionary (0/1)
  - the number of times a document mentions at least one word in the dictionary (continuous numbers)
  - or, the proportion that a document contains China-related words (to control for document length)
- We have a mapping of document $\rightarrow$ to outcome

Dictionary method: two dictionaries

- Sentiment analysis
- Research question: whether the news report is positive or negative toward China?
- Two dictionaries
  - One for words with positive sentiments;
  - The other for words with negative sentiments;
- A binary measure of sentiment for each document:
  - Positive, if there are more positive words than negative words
  - Negative, vice versa
- A continuous measure of sentiment for each document is:

$$\frac{\text{(number of positive words in that document) - (number of negative words in that document)}}{\text{number of total words in that document}}$$

Or write it down mathematically (Grimmer and Stewart)

- We have a vocabulary of size M
- Document-term matrix: W_{im}, the number of times the m-th word occurs in the i-th document.
- And each word m has a weight s_m, which can take three values:
  - 0 (if it is irrelevant to sentiments)
  - 1 (if it shows positive sentiment)
  - -1 (if it shows negative sentiment)
- Each document i has a length of $N_i = \sum_{m=1}^{M} W_{im}$
- Then sentiment score for a document i can be calculated as:

$$t_i = \frac{1}{N_i} \sum_{m=1}^{M} s_m W_{im}$$

Off-the-shelf dictionaries

- Lots of off-the-shelf dictionaries are available
  - For different tasks
- Some commonly used dictionaries for sentiments
  - Minqing Hu and Bing Liu, *Mining and summarizing customer reviews*, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA), KDD '04, Association for Computing Machinery, 2004, pp. 168–177
    - 6800 words, collected from customer review of products on Amazon: careras, DVD player, MP3 and cellular phone, developed by computer scientists
    - `http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar`
  - LIWC is more complex collection (not free)
    - Developed by psychologists
    - `https://liwc.wpengine.com/`

Off-the-shelf dictionaries

- Another example: detecting political events from newspapers with dictionaries
- GDELT (`https://www.gdeltproject.org/data.html#intro`)
  - categories include
    - Making public statement
    - Appealing for help
    - Calling for cooperation
    - Threatening
    - Protesting
    - Military fight
    - And many many more
- Each category has its own dictionary
- If an newspaper article contains more words in a corresponding categories, it is assigned to that category

Construct your own dictionary

- Sometimes off-the-shelf dictionary are not satisfactory
  - Words that are meaningful for restaurant reviews may not be working for your problem
- Construct by yourself!
  - Read your documents closely
  - And pick it up by yourself

Some modern approaches of constructing dictionary

- William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky, *Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora*, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 595–605
- Intuition: human have recall biases when constructing dictionaries
- Snow-ball sampling:
  - Start with a set of seed words
  - Find similar words in the corpus, and decide whether to add them to dictionary
    - In the article they used word embeddings (more in several weeks)
    - but you can borrow their idea, check thesaurus and find similar words
  - Iterate the above process until you reach a satisfactory dictionary

Shortcomings: polysemy

- polysemy: multiple meanings in word

Sentences	Sentiment word	Part-of-speech	Sentiment polarity
Jane is patient to children.	patient	adjective	😄
Now there is a patient in the class.	patient	noun	😩

- well as noun vs. well as adjective
- other examples you can think of?

Shortcoming: word choice

- What words to keep?
  - Often arbitrary decisions; even experts do not agree with each other
- Size of dictionary:
  - How large the dictionary should be? Is 200 positive words enough? Or we need to have 2,000 positive words?
  - Often it's tempting to select more words
  - This choice will lead to high recall, but low precision
- On the other hand, select very few or very specific words result in high precision but low recall

Shortcomings: word choice (cont'd)

- Precision-recall tradeoff
- For instance, select keywords associated with Boston Marathon bombings in 2013
  - #prayforboston selects relevant results, but most tweets about Boston Bombing may not contain this hashtag
  - "Boston" do not miss too much, but the rate it hits an relevant post is very low
- Gary King, Patrick Lam, and Margaret E. Roberts, *Computer-Assisted Keyword and Document Set Discovery from Unstructured Text*, American Journal of Political Science **61** (2017), no. 4, 971–988

Summary

- Using texts for prediction, or for measurements
- Turn documents into numbers:
  - document-term matrix
- Other data cleaning steps: stemming, lemmatization, segmentation, removing stop words
- Dictionary methods