Quantitative text analysis: Automated Dictionary Methods

Pablo Barberá

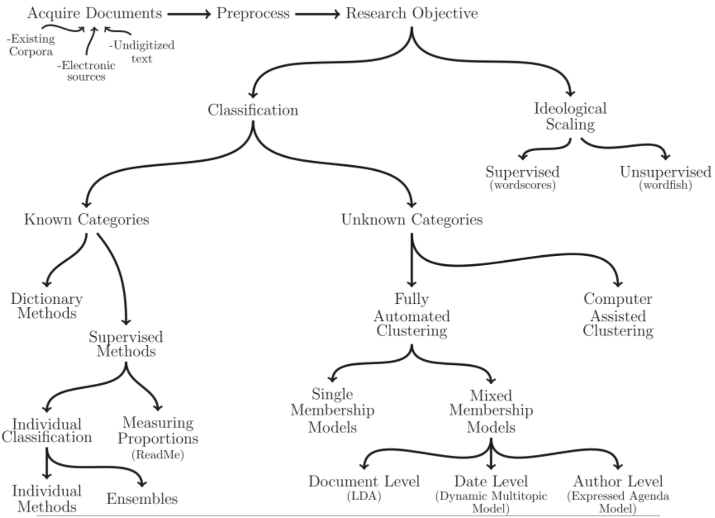
MY 459: Quantitative Text Analysis

January 28, 2019

Course website: lse-my459.github.io

1. Overview and Fundamentals
2. Descriptive Statistical Methods for Text Analysis
3. Automated Dictionary Methods
4. Machine Learning for Texts
5. Supervised Scaling Models for Texts
6. *Reading Week*
7. Unsupervised Models for Scaling Texts
8. Similarity and Clustering Methods
9. Topic models
10. Word embeddings
11. Working with Social Media

Overview of text as data methods



Outline for today

- Dictionary methods: an overview
- Some well-known dictionaries
- Advantages and disadvantages
- Dictionary construction
- Keyword detection
- Practical demo with quanteda

Dictionary methods

Classifying documents when categories are known:

- Lists of words that correspond to each category:
  - Positive or negative, for sentiment
  - Sad, happy, angry, anxious... for emotions
  - Insight, causation, discrepancy, tentative... for cognitive processes
  - Sexism, homophobia, xenophobia, racism... for hate speech

  many others: see LIWC, VADER, SentiStrength, LexiCoder...
- Count number of times they appear in each document
- Normalize by document length (optional)
- Validate, validate, validate.
  - Check sensitivity of results to exclusion of specific words
  - Code a few documents manually and see if dictionary prediction aligns with human coding of document

Bridging qualitative and quantitative text analysis

- A hybrid procedure between qualitative and quantitative classification at the fully automated end of the text analysis spectrum
- "Qualitative" since it involves identification of the concepts and associated keys/categories, and the textual features associated with each key/category
- Dictionary construction involves a lot of contextual interpretation and qualitative judgment
- Perfect reliability because there is no human decision making as part of the text analysis procedure

Rationale for dictionaries

- Rather than count words that occur, pre-define words associated with specific meanings
- Two components:

  key the label for the equivalence class for the concept or canonical term

  values (multiple) terms or patterns that are declared equivalent occurrences of the key class

- Frequently involves stemming/lemmatization: transformation of all inflected word forms to their "dictionary look-up form"

"Dictionary": a misnomer?

- A *dictionary* is really a thesaurus: a canonical term or concept (a "key") associated with a list of equivalent synonyms
- But dictionaries tend to be exclusive: they single out features defined as keys, selecting the terms or patterns linked to each key
- An alternative is a "thesaurus" concept: a tag of key equivalency for an associated set of terms, but non-exclusive
  - marriage = `engage, ring, wedding, spouse, husband, wife`
  - interest = `engage, appeal, excite, attract, entertain`

Outline for today

- Dictionary methods: an overview
- Some well-known dictionaries
- Advantages and disadvantages
- Dictionary construction
- Keyword detection
- Practical demo with quanteda

Well-known dictionaries: General Inquirer

- General Inquirer (Stone et al 1966)
- Example: self = *I*, *me*, *my*, *mine*, *myself*
  selves = *we*, *us*, *our*, *ours*, *ourselves*
- Latest version contains 182 categories – the "Harvard IV-4" dictionary, the "Lasswell" dictionary, and five categories based on the social cognition work of Semin and Fiedler
- Examples: "self references", containing mostly pronouns; "negatives", the largest category with 2291 entries
- Also uses disambiguation, for example to distinguishes between *race* as a contest, *race* as moving rapidly, *race* as a group of people of common descent, and *race* in the idiom "rat race"
- Output example:
  `http://www.wjh.harvard.edu/~inquirer/Spreadsheet.html`

Well-known dictionaries: Regressive Imagery Dictionary

- Consists of about 3,200 words and roots, assigned to 29 categories of primary process cognition, 7 categories of secondary process cognition, and 7 categories of emotions
- designed to measure primordial vs. conceptual thinking
  - Conceptual thought is abstract, logical, reality oriented, and aimed at problem solving
  - Primordial thought is associative, concrete, and takes little account of reality – the type of thinking found in fantasy, reverie, and dreams
- Categories were derived from the theoretical and empirical literature on regressive thought by Martindale (1975, 1990)

Regressive Imagery Dictionary categories

- Full listing of categories

1 orality	21 brink-passage	41 aggression	62 novelty
2 anality	22 narcissism	42 expressive behaviour	63 negation
3 sex	23 concreteness	43 glory	64 triviality
4 touch	24 ascend	44 female role	65 transmute
5 taste	25 height	45 male fole	
6 odour	26 descent	46 self	
7 general sensation	27 depth	47 related others	
8 sound	28 fire	48 diabolic	
9 vision	29 water	49 aspiration	
10 cold	30 abstract thought	50 angelic	
11 hard	31 social behaviour	51 flowers	
12 soft	32 instrumental behaviour	52 synthesize	
13 passivity	33 restraint	53 streight	
14 voyage	34 order	54 weakness	
15 random movement	35 temporal references	55 good	
16 diffusion	36 moral imperative	56 bad	
17 chaos	37 positive affect	57 activity	
18 unknown	38 anxiety	58 being	
19 timelessness	39 sadness	59 analogy	
20 counscious	40 affection	61 integrative con	

- More on categories:
`http://www.kovcomp.co.uk/wordstat/RID.html`

Linquistic Inquiry and Word Count

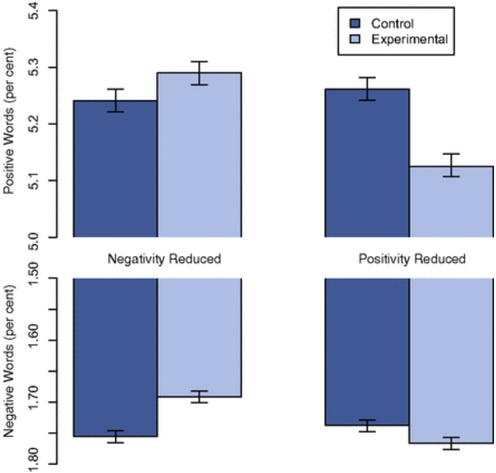
- Created by Pennebaker et al — see `http://www.liwc.net`
- Uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries
- For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. So observing the token *cried* causes each of these five subdictionary scale scores to be incremented
- Hierarchical: so "anger" are part of an *emotion* category and a *negative emotion* subcategory
- You can buy it here: `http://www.liwc.net/descriptiontable1.php`

Example: Terrorist speech (Pennebaker, 2008)

	Bin Ladin (1988 to 2006) N = 28	Zawahiri (2003 to 2006) N = 15	Controls N = 17	p (two-tailed)
Word Count	2511.5	1996.4	4767.5	
Big words (greater than 6 letters)	21.2a	23.6b	21.1a	.05
Pronouns	9.15ab	9.83b	8.16a	.09
I (e.g. I, me, my)	0.61	0.90	0.83	
We (e.g. we, our, us)	1.94	1.79	1.95	
You (e.g. you, your, yours)	1.73	1.69	0.87	
He/she (e.g. he, hers, they)	1.42	1.42	1.37	
They (e.g., they, them)	2.17a	2.29a	1.43b	.03
Prepositions	14.8	14.7	15.0	
Articles (e.g. a, an, the)	9.07	8.53	9.19	
Exclusive Words (but, exclude)	2.72	2.62	3.17	
Affect	5.13a	5.12a	3.91b	.01
Positive emotion (happy, joy, love)	2.57a	2.83a	2.03b	.01
Negative emotion (awful, cry, hate)	2.52a	2.28ab	1.87b	.03
Anger words (hate, kill)	1.49a	1.32a	0.89b	.01
Cognitive Mechanisms	4.43	4.56	4.86	
Time (clock, hour)	2.40b	1.89a	2.69b	.01
Past tense verbs	2.21a	1.63a	2.94b	.01
Social Processes	11.4a	10.7ab	9.29b	.04
Humans (e.g. child, people, selves)	0.95ab	0.52a	1.12b	.05
Family (mother, father)	0.46ab	0.52a	0.25b	.08
Content				
Death (e.g. dead, killing, murder)	0.55	0.47	0.64	
Achievement	0.94	0.89	0.81	
Money (e.g. buy, economy, wealth)	0.34	0.38	0.58	
Religion (e.g. faith, Jew, sacred)	2.41	1.84	1.89	

Note. Numbers are mean percentages of total words per text file. Statistical tests are between Bin Ladin, Zawahiri, and Controls. Documents whose source indicates "Both" (n=3) or "Unknown" (n=2) were excluded due to their small sample sizes.

Example: Emotional Contagion on Facebook



Source: Kramer et al, PNAS 2014

VADER: an open-source alternative to LIWC

Valence Aware Dictionary and sEntiment Reasoner:

- Especially tuned for social media text
- Captures polarity and intensity of sentiments
- Includes emoticons, emoji, slang
- Feature-specific weights
- Python and R libraries: `https://github.com/cjhutto/vaderSentiment`

Other open-source sentiment dictionaries: LexiCoder (media text), SentiStrength (social media text)

Example: Laver and Garry (2000)

- A *hierarchical* set of categories to distinguish policy domains and policy positions – similar in spirit to the CMP
- Five domains at the top level of hierarchy
  - economy
  - political system
  - social system
  - external relations
  - a " 'general' domain that has to do with the cut and thurst of specific party competition as well as uncodable pap and waffle"
- Looked for word occurences within "word strings with an average length of ten words"
- Built the dictionary on a set of specific UK manifestos

Example: Laver and Garry (2000): Economy

Table 1 Abridged Section of Revised Manifesto Coding Scheme

1 ECONOMY
Role of state in economy

- 1 1 ECONOMY/+State+
  Increase role of state
  - 1 1 1 ECONOMY/+State+/Budget
    Budget
    - 1 1 1 1 ECONOMY/+State+/Budget/Spending
      Increase public spending
      - 1 1 1 1 1 ECONOMY/+State+/Budget/Spending/Health
      - 1 1 1 1 2 ECONOMY/+State+/Budget/Spending/Educ. and training
      - 1 1 1 1 3 ECONOMY/+State+/Budget/Spending/Housing
      - 1 1 1 1 4 ECONOMY/+State+/Budget/Spending/Transport
      - 1 1 1 1 5 ECONOMY/+State+/Budget/Spending/Infrastructure
      - 1 1 1 1 6 ECONOMY/+State+/Budget/Spending/Welfare
      - 1 1 1 1 7 ECONOMY/+State+/Budget/Spending/Police
      - 1 1 1 1 8 ECONOMY/+State+/Budget/Spending/Defense
      - 1 1 1 1 9 ECONOMY/+State+/Budget/Spending/Culture
    - 1 1 1 2 ECONOMY/+State+/Budget/Taxes
      Increase taxes
      - 1 1 1 2 1 ECONOMY/+State+/Budget/Taxes/Income
      - 1 1 1 2 2 ECONOMY/+State+/Budget/Taxes/Payroll
      - 1 1 1 2 3 ECONOMY/+State+/Budget/Taxes/Company
      - 1 1 1 2 4 ECONOMY/+State+/Budget/Taxes/Sales
      - 1 1 1 2 5 ECONOMY/+State+/Budget/Taxes/Capital
      - 1 1 1 2 6 ECONOMY/+State+/Budget/Taxes/Capital gains
    - 1 1 1 3 ECONOMY/+State+/Budget/Deficit
      Increase budget deficit
      - 1 1 1 3 1 ECONOMY/+State+/Budget/Deficit/Borrow
      - 1 1 1 3 2 ECONOMY/+State+/Budget/Deficit/Inflation

Example: Laver and Garry (2000)

```
ECONOMY / +STATE
    accommodation
    age
    ambulance
    assist
    ...

ECONOMY / -STATE
    choice*
    compet*
    constrain*
    ...
```

MFD (Graham and Haidt)

Moral Foundations dictionary:

- Moral foundations: dimensions of difference that explain human moral reasoning
- Measures the proportions of virtue and vice words for each foundation:
  1. Care/Harm
  2. Fairness/Cheating
  3. Loyalty/Betrayal
  4. Authority/Subversion
  5. Purity/Degradation
- Link: `https://www.moralfoundations.org/othermaterials`

Outline for today

- Dictionary methods: an overview
- Some well-known dictionaries
- Advantages and disadvantages
- Dictionary construction
- Keyword detection
- Practical demo with quanteda

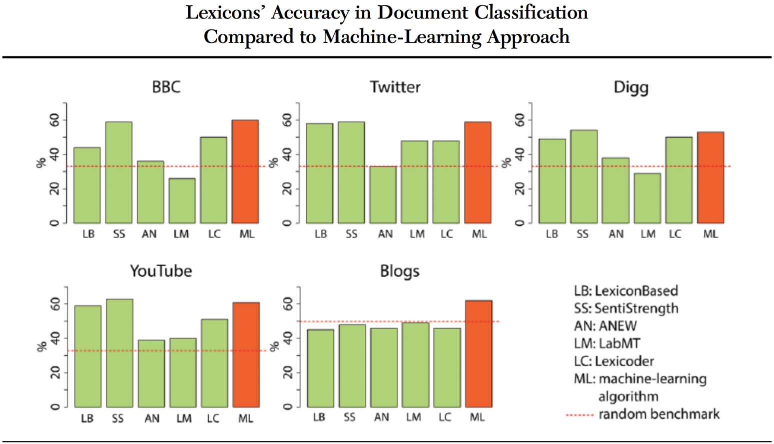
Potential advantage: Multi-lingual

APPENDIX B
DICTIONARY OF THE COMPUTER-BASED CONTENT ANALYSIS

	NL	UK	GE	IT
Core	elit*	elit*	elit*	elit*
	consensus*	consensus*	konsens*	consens*
	ondemocratisch*	undemocratic*	undemokratisch*	antidemocratic*
	ondemokratisch*			
	referend*	referend*	referend*	referend*
	corrupt*	corrupt*	korrupt*	corrot*
	propagand*	propagand*	propagand*	propagand*
	politici*	politici*	politiker*	politici*
	bedrog	*deceit*	täusch*	ingann*
	bedrieg	*deceiv*	betrüg*	
			betrug*	
	verraa	*betray*	*verrat*	tradi*
	verrad			
	schaam*	shame*	scham*	vergogn*
			schäm*	
	schand*	scandal*	skandal*	scandal*
	waarheid*	truth*	wahrheit*	verità
	oneerlijk*	dishonest*	unfair*	disonest*
			unehrlich*	
Context	establishm*	establishm*	establishm*	partitocrazia
	heersend*	ruling*	*herrsch*	
	capitul*			
	kapitul*			
	kaste*			
	leugen*		lüge*	menzogn*
	lieg*			mentir*

(from Rooduijn and Pauwels 2011)

Potential disadvantage: Context specific



Source: González-Bailón and Paltoglou (2015)

Disadvantage: Highly specific to context

- Example: Loughran and McDonald used the Harvard-IV-4 TagNeg (H4N) file to classify sentiment for a corpus of 50,115 firm-year 10-K filings from 1994–2008
- found that almost three-fourths of the "negative" words of H4N were typically not negative in a financial context e.g. *mine* or *cancer*, or *tax*, *cost*, *capital*, *board*, *liability*, *foreign*, and *vice*
- Problem: polysemes – words that have multiple meanings
- Another problem: dictionary lacked important negative financial words, such as *felony*, *litigation*, *restated*, *misstatement*, and *unanticipated*

Potential disadvantage: sensitive to frequent words

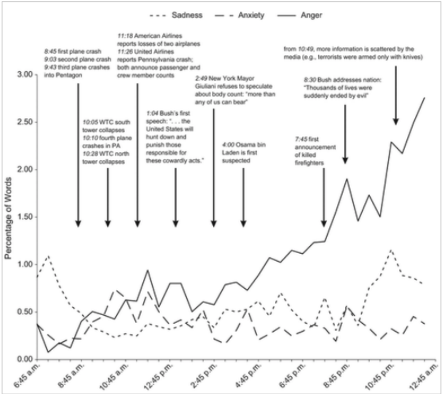


Fig. 1. The timeline of sadness, anxiety, and anger on September 11 as expressed in messages sent to text pagers. Each data point represents the mean percentage of words related to the specific negative emotion, averaged across 30 min. The time slots start at 6:45 a.m. to 7:14 a.m. on September 11, 2001, and end at 12:15 a.m. to 12:44 a.m. on September 12, 2001. Exact times and brief descriptions of the most important events of September 11 are included above the timelines. WTC = World Trade Center

(from Back et al, Psychological Science, 2010)

Potential disadvantage: sensitive to frequent words

Automation can lead to confounds in text analysis: Back, Küfner, and Egloff (2010) and the not-so-angry Americans.

EXPORT Add To My List Database: PsycINFO Comment/ Reply

Pury, Cynthia L. S.

Citation

Pury, C. L. S. (2011). Automation can lead to confounds in text analysis: Back, Küfner, and Egloff (2010) and the not-so-angry Americans. *Psychological Science, 22*(6), 835-836.
http://dx.doi.org/10.1177/0956797611408735

Abstract

Comments on an article by Mitja D. Back et al. (see record 2010-25035-010). The authors used Linguistic Inquiry and Word Count (LIWC) to analyze pager messages sent to more than 85,000 American pagers on September 11, 2001. They found that anger, as indexed by the words contained in those messages, rose steadily throughout the day. The data contained many technical codes; thus, Back et al. counted only words recognized by LIWC. However, this procedure did not exclude automatically generated messages. Consequently, LIWC words in such messages were counted, even if the words lacked emotional meaning in context. Furthermore, computers can send messages with superhuman frequency, turning an otherwise minor measurement error into a serious confound. This confound can be detected by treating individual text messages as primary units, reading samples of each key word in context, and looking for repeating false positives. Thus, it appears that much of the dramatic rise in anger reported by Back et al. was due to a repeated and emotionally neutral technical message associated with a single pager. Because today's e-mail, social media, and text messages can include automatically generated messages, future researchers of linguistic archives should consider ways to prevent similar confounds.

Potential disadvantage: sensitive to frequent words

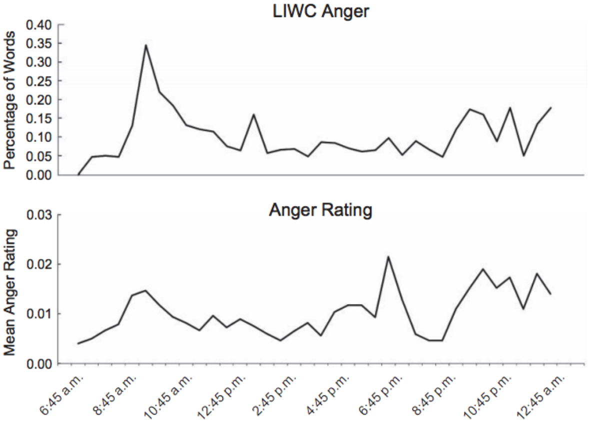


Fig. 1. A revised timeline of anger as expressed in 37,606 social messages sent to text pagers on September 11, 2001. The graphs show (a) the mean percentage of words related to anger (as classified by Linguistic Inquiry and Word Count; Pennebaker, Francis, & Booth, 2001) and (b) the mean anger rating (0 = *no anger*, 1 = *some anger*, 2 = *strong anger*; averaged across three raters for each message) across time slots starting at 6:45 a.m. to 7:14 a.m. on September 11, 2001, and ending at 12:15 a.m. to 12:44 a.m. on September 12, 2001.

(from Back et al, Psychological Science, 2011)

Outline for today

- Dictionary methods: an overview
- Some well-known dictionaries
- Advantages and disadvantages
- Dictionary construction
- Keyword detection
- Practical demo with quanteda

How to build a dictionary

- The ideal content analysis dictionary associates all and only the relevant words to each category in a perfectly valid scheme
- Three key issues:

Validity	Is the dictionary's category scheme valid?
Recall	Does this dictionary identify *all* my content?
Precision	Does it identify *only* my content?

- Imagine two logical extremes of including all words (too sensitive), or just one word (too specific)

How to build a dictionary

1. Identify “extreme texts” with “known” positions. Examples:
   - Tweets by populist vs mainstream parties (for populism dictionary)
   - Opposition leader and Prime Minister in a no-confidence debate (for opposition vs government dictionary)
   - Facebook comments to news about natural catastrophes vs football victories (for sentiment dictionary)
   - Subreddits for white nationalist groups vs regular politics (for racist rhetoric)
2. Search for differentially occurring words using word frequencies
3. Examine these words in context to check their precision and recall
4. Use regular expressions to see whether stemming or wildcarding is required

Outline for today

- Dictionary methods: an overview
- Some well-known dictionaries
- Advantages and disadvantages
- Dictionary construction
- Keyword detection
- Practical demo with quanteda

Detecting “keywords”

- Detects words that *discriminate* between partitions of a corpus
- For instance, we could partition the Irish budget speech corpus into “government” and “opposition” speeches, and look for words that occur in one partition with higher relative frequency in opposition than in government speeches
- This is done by constructing a 2×2 table for each word, and testing association between that word and the partition categories

Detecting "keywords": Constructing the association table

	Target	~ Target	
Word 1	n_{11}	n_{12}	$n_{1.}$
~ (Word 1)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

- Once this is constructed, any standard measures of association (similar to those used to detect collocations) can be used to identify keyword associations with a class
- Same association measures are used as with collocation detection

statistical association measures

where m_{ij} represents the cell frequency expected according to independence:

G^2 likelihood ratio statistic, computed as:

$$2 * \sum_i \sum_j (n_{ij} * log \frac{n_{ij}}{m_{ij}})$$

χ^2 Pearson's χ^2 statistic, computed as:

$$\sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

pmi point-wise mutual information score, computed as $\log n_{11}/m_{11}$

Examples

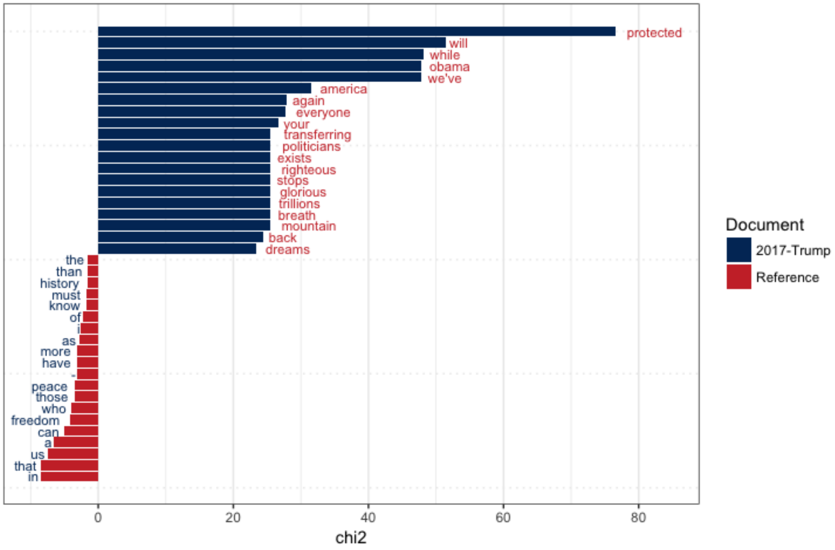
```
# compare Trump 2017 to other post-war presidents
period <- ifelse(docvars(data_corpus_inaugural, "Year") < 1945,
                 "pre-war", "post-war")
pwdfm <- dfm(corpus_subset(data_corpus_inaugural, period == "post-war"))

textstat_keyness(pwdfm, target = "2017-Trump") %>%
    head(n = 7)
#     feature     chi2            p n_target n_reference
# 1 protected 76.64466 0.000000e+00        5           1
# 2      will 51.44795 7.351897e-13       40         299
# 3     while 48.23022 3.790079e-12        6           7
# 4     obama 47.85727 4.584000e-12        3           0
# 5     we've 47.85727 4.584000e-12        3           0
# 6   america 31.45537 2.040775e-08       18         112
# 7     again 27.81145 1.337322e-07        9          33
```

Examples

```
# using the likelihood ratio method
textstat_keyness(dfm_smooth(pwdfm), measure = "lr", target = "2017-Trum
    head()
#    feature        G2            p n_target n_reference
# 1     will 24.604106 7.040156e-07       41         317
# 2  america 14.040255 1.789387e-04       19         130
# 3     your 10.435140 1.236402e-03       12          68
# 4    again  9.758516 1.784939e-03       10          51
# 5    while  9.504990 2.049139e-03        7          25
# 6 american  8.877690 2.886766e-03       12          76


textstat_keyness(pwdfm, target = "2017-Trump") %>%
    textplot_keyness()
```


protected
will
while
obama
we've
america
again
everyone
your
transferring
politicians
exists
righteous
stops
glorious
trillions
breath
mountain
back
dreams
the
than
history
must
know
of
i
as
more
have
-
peace
those
who
freedom
can
a
us
that
in
0
20
40
60
80
chi2
Document
2017-Trump
Reference

Outline for today

- Dictionary methods: an overview
- Some well-known dictionaries
- Advantages and disadvantages
- Dictionary construction
- Keyword detection
- Practical demo with quanteda