Text Mining and Extracting Value from Text as Big Data

Kenneth Benoit

Big Data Economics Summer School

7 September 2016

Outline of my talk

- Definition and motivation of text mining
- Basic process and assumptions
- Examples
- Tools
- Assumptions and Process
- Defining Features
- Key Words in Context
- Dictionaries
- Topic Models

Challenges and opportunities

- Text is ubiquitous
- Text is semi-structured
- Text is unstructured
- Text can be used for machine learning and statistics, but follows a different data-generating process

Basic Text Mining Process: Texts → Feature matrix → Analysis

When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will

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

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

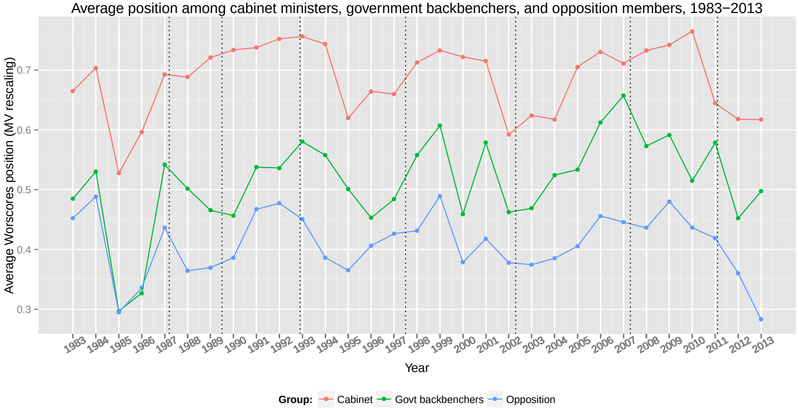
Vocabulary analysis

Sentiment analysis

Sources of text

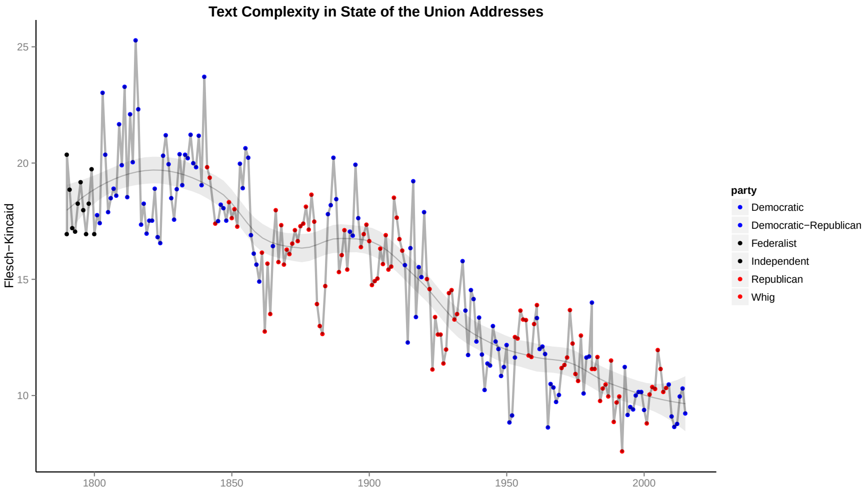
- Electronic publication: the Internet
- Private assets – especially for information technology companies
- Self-generated through research
- Social media: 400 million Tweets per day

Government v. Opposition in yearly budget debates



(from Herzog and Benoit EPSA 2013)

Reading level of US State-of-the-Union addresses over time



Reading level of Trump speech

POLITICO MAGAZINE

OUR LATEST | SEARCH | EMAI

6.2K
SHARES

Facebook
Twitter
Google +
Email
Comment
Print

FOURTH ESTATE

Donald Trump Talks Like a Third-Grader

By **JACK SHAFER** | August 13, 2015

Share on Facebook | Share on Twitter

Donald Trump isn't a simpleton, he just talks like one. If you were to market Donald Trump's vocabulary as a toy, it would resemble a small box of Lincoln Logs. Trump resists

Example: Adams v. Trump

James Adams, 1791

Numerous as are the providential blessings which demand our grateful acknowledgments, the abundance with which another year has again rewarded the industry of the husbandman is too important to escape recollection. (19.3 FK)

Donald J. Trump, 2015

Now, we have to build a fence. And it's got to be a beauty. Who can build better than Trump? I build; it's what I do. I build; I build nice fences, but I build great buildings. Fences are easy, believe me. (0.9 FK)

Analysis of European Parliament election candidates 2014

Contents lists available at ScienceDirect

Electoral Studies

journal homepage: www.elsevier.com/locate/electstud

Social media and political communication in the 2014 elections to the European Parliament☆

Paul Nulty [a, *], Yannis Theocharis [b], Sebastian Adrian Popa [b], Olivier Parnet [c], Kenneth Benoit [a]

[a] London School of Economics and Political Science, UK
[b] Mannheim Centre for European Social Research (MZES), Germany
[c] TNS Europe, UK

A R T I C L E I N F O

Article history:
Received 12 October 2015
Accepted 24 April 2016
Available online xxx

Keywords:
Electoral participation
Political communication
Social networks
European elections
Content analysis
Social media

A B S T R A C T

Social media play an increasingly important part in the communication strategies of political campaigns by reflecting information about the policy preferences and opinions of political actors and their public followers. In addition, the content of the messages provides rich information about the political issues and the framing of those issues during elections, such as whether contested issues concern Europe or rather extend pre-existing national debates. In this study, we survey the European landscape of social media using tweets originating from and referring to political actors during the 2014 European Parliament election campaign. We describe the language and national distribution of the messages, the relative volume of different types of communications, and the factors that determine the adoption and use of social media by the candidates. We also analyze the dynamics of the volume and content of the communications over the duration of the campaign with reference to both the EU integration dimension of the debate and the prominence of the most visible list-leading candidates. Our findings indicate that the lead candidates and their televised debate had a prominent influence on the volume and content of communications, and that the content and emotional tone of communications more reflects preferences along the EU dimension of political contestation rather than classic national issues relating to left-right differences.

© 2016 Published by Elsevier Ltd.

Table 1
Candidates and election-related twitter communication during the 2014 EP Elections, by country (updating candidates accounts).

Country	Total parties	Total candidates	Cands w/Twitter	% Using Twitter	Total tweets
By country					
Ireland	7	41	30	73.2	7300
Sweden	12	373	249	66.8	36,483
Finland	9	249	166	66.7	16,797
Netherlands	10	345	229	66.4	42,109
Italy	8	653	355	54.4	70,414
Denmark	8	100	54	54	5513
United Kingdom	9	749	341	45.5	66,921
Latvia	6	170	64	37.6	4220
Slovenia	10	118	44	37.3	4150
Luxembourg	8	54	19	35.2	14
Cyprus	5	48	15	31.2	587
Estonia	7	88	26	29.5	1115
Austria	7	348	78	22.4	19,876
Greece	9	544	118	21.7	7460
Belgium	13	182	38	20.9	2345
Poland	8	1286	249	19.4	13,696
Germany	7	946	163	17.2	16,772
Lithuania	9	257	33	12.8	507
Spain	9	2105	266	12.6	76,784
France	7	3735	411	11	38,361
Slovakia	10	334	36	10.8	1193
Croatia	7	275	26	9.5	876
Hungary	6	322	29	9	218
Romania	10	580	48	8.3	411
Bulgaria	7	286	23	8	830
Portugal	5	336	22	6.5	4482
Czech Republic	9	829	48	5.8	1867
Total	222	15,353	3180		441,301
By incumbency status					
Non-incumbent		14,607	2641	18%	
Incumbent		746	539	72%	
Total		15,353	3180	21%	

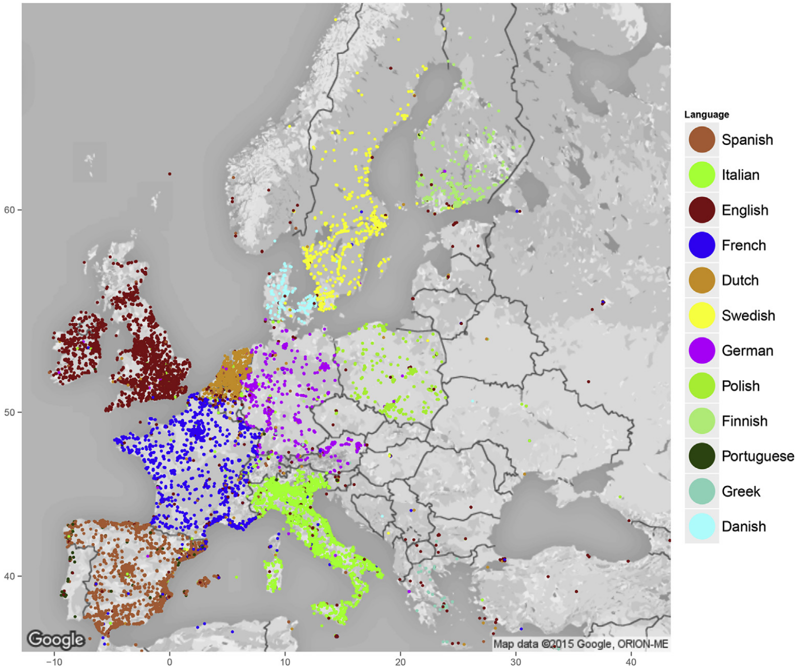


Fig. 1. Location of tweets with co-ordinate information enabled, colored by the language of the tweet.

Table 2
Predicting MEP Candidates' Adoption of Twitter. Multilevel logistic regression with exponentiated coefficients and confidence intervals.

	Dependent variable: Candidate has a twitter account				
	(1)	(2)	(3)	(4)	(5)
Fixed effects					
Constant	0.251***	0.321***	0.178***	0.134***	0.158***
	(0.175, 0.361)	(0.234, 0.440)	(0.100, 0.318)	(0.061, 0.296)	(0.070, 0.360)
MEP 2014	7.191***	4.492***	3.432***	3.391***	3.415***
	(5.847, 8.843)	(3.660, 5.512)	(2.725, 4.324)	(2.693, 4.271)	(2.711, 4.300)
MEP 2009	5.713***	4.261***	3.388***	3.430***	3.408***
	(4.415, 7.392)	(3.309, 5.487)	(2.540, 4.519)	(2.572, 4.572)	(2.554, 4.546)
MEP gender		1.201***			
		(1.087, 1.328)			
EU position (party)			1.200***		1.148***
			(1.105, 1.303)		(1.044, 1.264)
Left-right (party)				1.791***	1.211
				(1.344, 2.388)	(0.824, 1.780)
Left-right^2 (party)				0.944***	0.981
				(0.920, 0.970)	(0.946, 1.018)
Party size			2.924	2.996	1.999
			(0.729, 11.734)	(0.801, 11.201)	(0.565, 7.072)
Internet penetration (country)	1.071***	1.062***	1.078***	1.074***	1.076***
	(1.040, 1.104)	(1.031, 1.094)	(1.042, 1.114)	(1.039, 1.109)	(1.041, 1.112)
Random effects (variance)					
Intercept (party)			2.012	1.592	2.823
EU position (party)			0.060		0.032
Left-right (party)				0.017	0.015
Party size			8.47	6.331	4.282
Intercept (country)	0.825	0.602	0.733	0.694	0.745
Internet penetration	0.001	0.001	0.001	0.001	0.001
Observations (Candidates)	15,361	9,335	6298	6298	6298
Observations (Party)			174	174	174
Observations (Country)	27	27	27	27	27
Log likelihood	−6145.901	−5021.652	−3404.106	−3404.993	−3399.572
Akaike Inf. Crit.	12305.800	10059.310	6838.212	6841.987	6841.144
Bayesian Inf. Crit.	12359.280	10116.440	6939.432	6949.955	6982.852

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

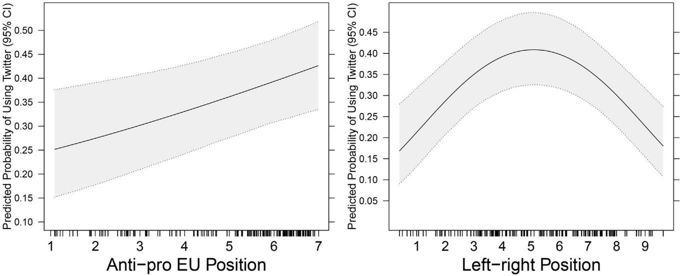


Fig. 2. Effect of candidate party's left-right position on the predicted probability of having a Twitter account. Predicted probabilities computed based on Model 5 respectively Model 4 in Table 2. Predicted values computed while holding all continuous variable at the mean and all categorical variables at zero.

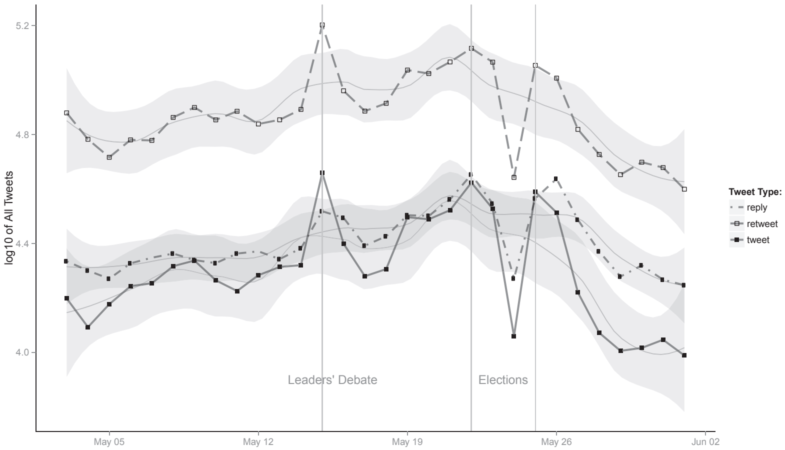


Fig. 4. Overall tweet volume throughout the campaign, by tweet type.

Fig. 7. Hashtags co-occurring with names of lead candidates. Each hashtag is positioned and colored by the candidate with which it has the highest relative co-occurrence.

Table 6
OLS regression of log ratio of positive to negative emotion as measured by the LIWC on tweets aggregated by candidate, for English, Spanish, German, Italian, French, and Dutch. Policy data from Chapel Hill Survey.

	Dependent variable: log (positive/negative)
EU Position	0.041^{***}
	(0.011)
Left-right	0.008
	(0.036)
Left-right^2	-0.001
	(0.004)
English	0.328^{***}
	(0.047)
French	-0.188^{***}
	(0.052)
German	-0.099^{*}
	(0.053)
Italian	-1.082^{***}
	(0.056)
Spanish	-0.097^{*}
	(0.055)
Constant	0.385^{***}
	(0.070)
Observations	4269
R^2	0.205
Adjusted R^2	0.203
Residual Std. Error	0.855 (df = 4260)
F Statistic	137.144^{***} (df = 8; 4260)

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

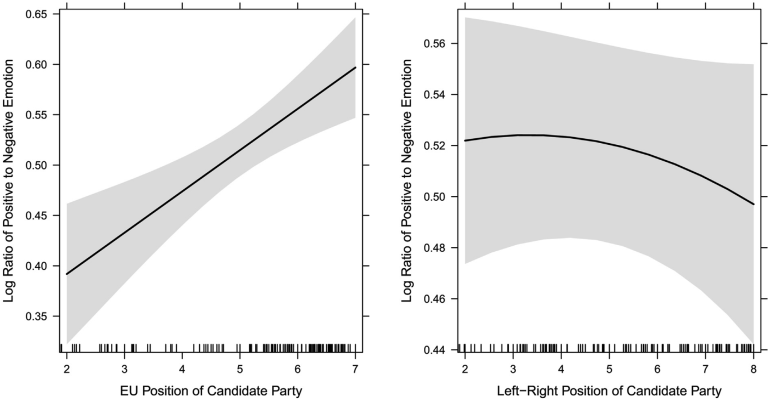
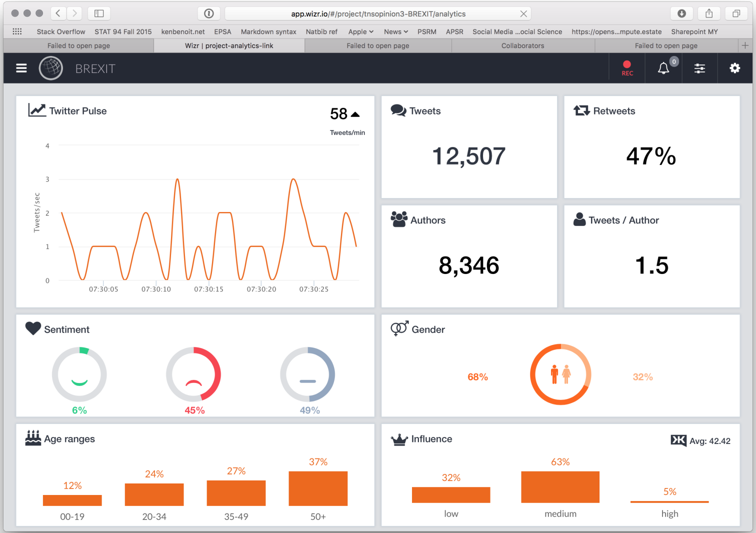
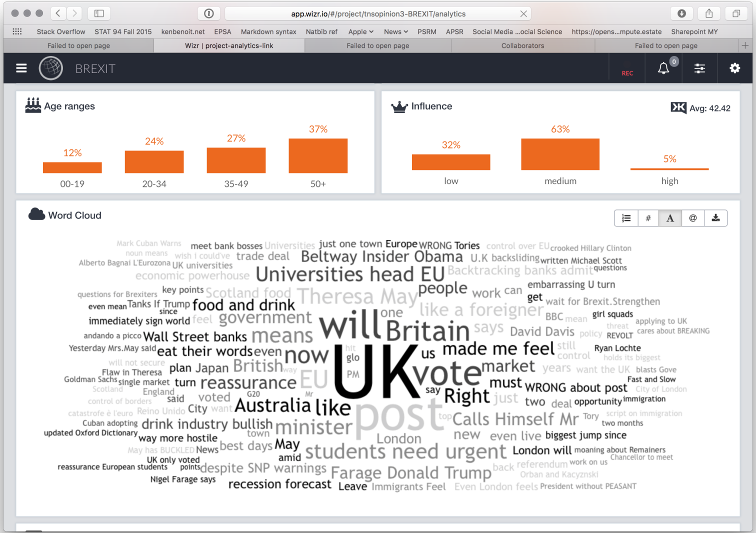


Fig. 11. Marginal effects on emotional tone of EU and general left-right positions of the candidate's party. From Table 6.

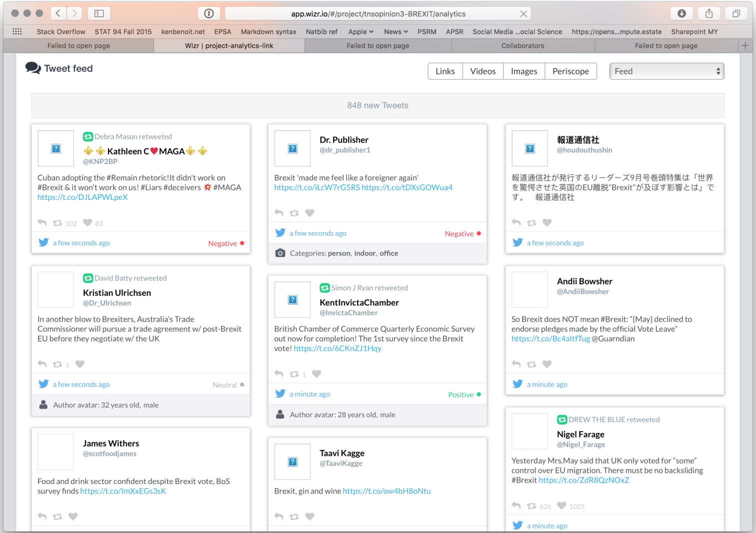
Twitter data from the "Brexit" debate in the UK



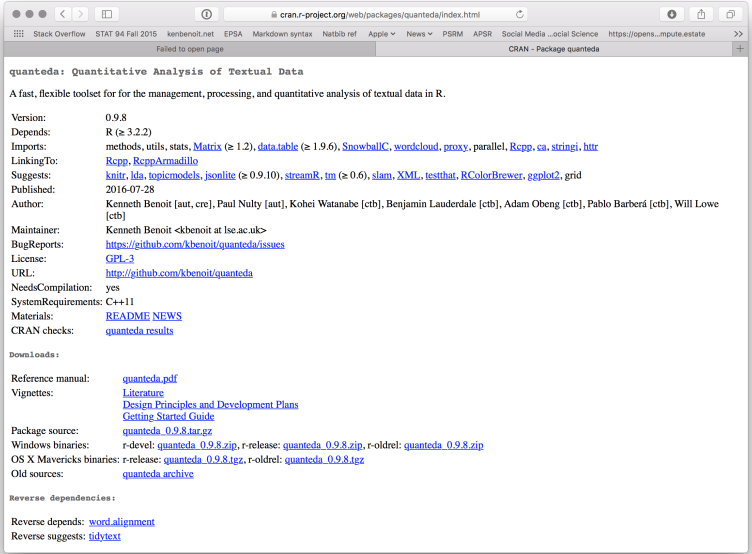
Twitter data from the “Brexit” debate in the UK



Twitter data from the “Brexit” debate in the UK



Tools for performing text analytics: R



cran.r-project.org/web/packages/quanteda/index.html

quanteda: Quantitative Analysis of Textual Data

A fast, flexible toolset for for the management, processing, and quantitative analysis of textual data in R.

Version:	0.9.8
Depends:	R (≥ 3.2.2)
Imports:	methods, utils, stats, Matrix (≥ 1.2), data.table (≥ 1.9.6), SnowballC, wordcloud, proxy, parallel, Rcpp, ca, stringi, httr
LinkingTo:	Rcpp, RcppArmadillo
Suggests:	knitr, lda, topicmodels, jsonlite (≥ 0.9.10), streamR, tm (≥ 0.6), slam, XML, testthat, RColorBrewer, ggplot2, grid
Published:	2016-07-28
Author:	Kenneth Benoit [aut, cre], Paul Nulty [aut], Kohei Watanabe [ctb], Benjamin Lauderdale [ctb], Adam Obeng [ctb], Pablo Barberá [ctb], Will Lowe [ctb]
Maintainer:	Kenneth Benoit <kbenoit at lse.ac.uk>
BugReports:	https://github.com/kbenoit/quanteda/issues
License:	GPL-3
URL:	http://github.com/kbenoit/quanteda
NeedsCompilation:	yes
SystemRequirements:	C++11
Materials:	README NEWS
CRAN checks:	quanteda results

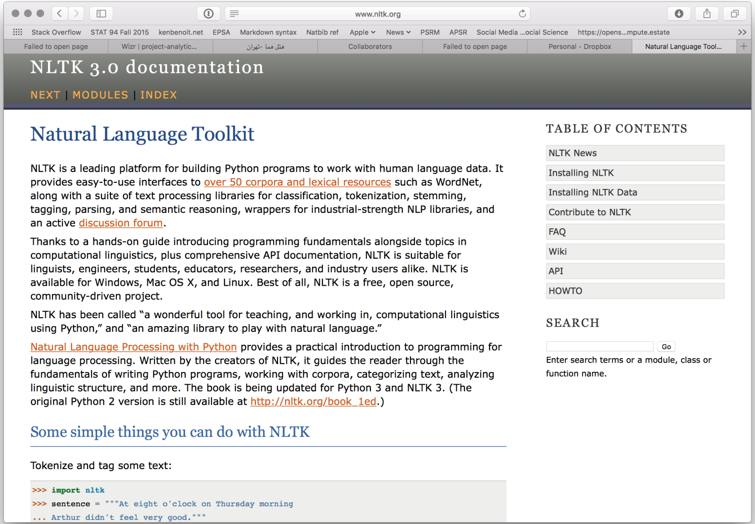
Downloads:

Reference manual:	quanteda.pdf
Vignettes:	Literature Design Principles and Development Plans Getting Started Guide
Package source:	quanteda_0.9.8.tar.gz
Windows binaries:	r-devel: quanteda_0.9.8.zip, r-release: quanteda_0.9.8.zip, r-oldrel: quanteda_0.9.8.zip
OS X Mavericks binaries:	r-release: quanteda_0.9.8.tgz, r-oldrel: quanteda_0.9.8.tgz
Old sources:	quanteda archive

Reverse dependencies:

Reverse depends:	word.alignment
Reverse suggests:	tidytext

Tools for performing text analytics: python



Quantitative text analysis requires assumptions

- That texts represent an observable implication of some underlying characteristic of interest (usually an attribute of the author)
- That texts can be represented through extracting their *features*
  - most common is the bag of words assumption
  - many other possible definitions of "features"
- A document-feature matrix can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

Key feature of quantitative text analysis

1. Selecting texts: Defining the *corpus*
2. Conversion of texts into a common electronic format
3. Defining documents: deciding what will be the doumentary unit of analysis

Key feature of quantitative text analysis (cont.)

4. Defining features. These can take a variety of forms, including tokens, equivalence classes of tokens (dictionaries), selected phrases, human-coded segments (of possibly variable length), linguistic features, and more.
5. Conversion of textual features into a quantitative matrix
6. A quantitative or statistical procedure to extract information from the quantitative matrix
7. Summary and interpretation of the quantitative results

Word frequencies and their properties

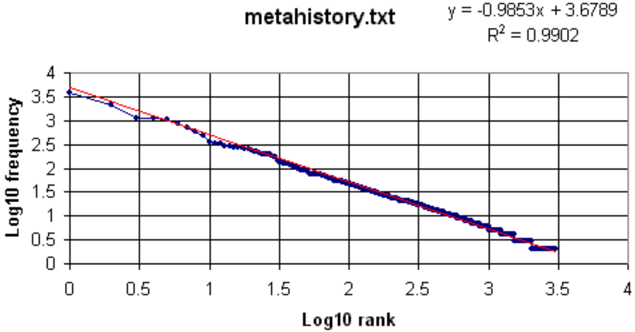
- Individual word usage tends to be associated with a particular degree of affect, position, etc. without regard to context of word usage
- Single tend to be the most informative, as n-grams are very rare
- Some approaches focus on occurrence of a word as a binary variable, irrespective of frequency: a binary outcome
- Other approaches use frequencies: Poisson, multinomial, and related distributions

Word frequency: Zipf's Law

- Zipf's law: Given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.
- The simplest case of Zipf's law is a "1/f function". Given a set of Zipfian distributed frequencies, sorted from most common to least common, the second most common frequency will occur $1/2$ as often as the first. The third most common frequency will occur $1/3$ as often as the first. The nth most common frequency will occur $1/n$ as often as the first.
- In the English language, the probability of encountering the the most common word is given roughly by $P(r) = 0.1/r$ for up to 1000 or so

Word frequency: Zipf's Law

- Formulaically: if a word occurs f times and has a rank r in a list of frequencies, then for all words $f = \frac{a}{r^b}$ where a and b are constants and b is close to 1
- So if we log both sides, $\log(f) = \log(a) - b\log(r)$
- If we plot $\log(f)$ against $\log(r)$ then we should see a straight line with a slope of approximately -1.



Defining Features

- words
- word stems or lemmas: this is a form of defining *equivalence classes* for word features
- word segments, especially for languages using compound words, such as German, e.g. *Rindfleischetikettierungsberwachungsaufgabenbertragungsgesetz* (the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef)

Defining Features (cont.)

- "word" sequences, especially when inter-word delimiters (usually white space) are not commonly used, as in Chinese

  莎拉波娃现在居住在美国东南部的佛罗里达。今年 4 月
  9 日，莎拉波娃在美国第一大城市纽约度过了 1 8 岁生
  日。生日派对上，莎拉波娃露出了甜美的微笑。

- linguistic features, such as parts of speech
- (if qualitative coding is used) coded or annotated text segments
- linguistic features: parts of speech

Parts of speech

- the Penn "Treebank" is the standard scheme for tagging POS

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Parts of speech (cont.)

- several open-source projects make it possible to tag POS in text, namely Apache's OpenNLP (and R package openNLP wrapper)

```
> s
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
> sprintf("%s/%s", s[a3w], tags)
 [1] "Pierre/NNP"      "Vinken/NNP"      ",/,"              "61/CD"
 [5] "years/NNS"       "old/JJ"          ",/,"              "will/MD"
 [9] "join/VB"         "the/DT"          "board/NN"         "as/IN"
[13] "a/DT"            "nonexecutive/JJ" "director/NN"      "Nov./NNP"
[17] "29/CD"           "./."             "Mr./NNP"          "Vinken/NNP"
[21] "is/VBZ"          "chairman/NN"     "of/IN"            "Elsevier/NNP"
[25] "N.V./NNP"        ",/,"             "the/DT"           "Dutch/JJ"
[29] "publishing/NN"   "group/NN"        "./."
```

Common English stop words

`a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, I, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your`

- But no list should be considered universal

Stemming words

Lemmatization refers to the algorithmic process of converting words to their lemma forms.

stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

both convert the morphological variants into stem or root terms

example: `produc` from `production, producer, produce, produces, produced`

Exploring Texts: Key Words in Context

Key words in context Refers to the most common format for concordance lines. A KWIC index is formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index.

Irish Budget Speeches KIWC in quanteda

R Console

```
> data(iebudgets)
> iebudgets2010 <- subset(iebudgets, year==2010)
> kwic(iebudgets2010, "christmas", regex=TRUE)
                                                                  preword       word                                    postword
     [2010_BUDGET_02_Richard_Bruton_FG.txt, 628]         and to see out this  Christmas in the hope of something
        [2010_BUDGET_03_Joan_Burton_LAB.txt, 371]        to suggest titles for a Christmas hit single. Fianna Fáil’s hit
        [2010_BUDGET_03_Joan_Burton_LAB.txt, 379]     Fianna Fáil’s hit single for Christmas will be, “I saw NAMA
        [2010_BUDGET_03_Joan_Burton_LAB.txt, 922]     women will say goodbye after Christmas  because they must take the
       [2010_BUDGET_03_Joan_Burton_LAB.txt, 1518]      in single golf clubs this Christmas. With a possible election next
       [2010_BUDGET_03_Joan_Burton_LAB.txt, 1726] Community faking its message this Christmas? Is the Society of St.
       [2010_BUDGET_03_Joan_Burton_LAB.txt, 3159]        bags. In previous years at Christmas  time people were laden down
       [2010_BUDGET_04_Arthur_Morgan_SF.txt, 346]             €204 per week or the Christmas  bonus. Of course, that is
      [2010_BUDGET_04_Arthur_Morgan_SF.txt, 3239]   to social welfare payments this Christmas. The loss of the Christmas
      [2010_BUDGET_04_Arthur_Morgan_SF.txt, 3244]        Christmas. The loss of the Christmas  bonus, a double payment which
      [2010_BUDGET_04_Arthur_Morgan_SF.txt, 3272]     streets on Santa presents and Christmas  food. The Government’s Scrooge measures
      [2010_BUDGET_04_Arthur_Morgan_SF.txt, 5899]         their jobs, who face this Christmas  in debt, in poverty and
         [2010_BUDGET_06_Enda_Kenny_FG.txt, 2629] to implement the reduction before Christmas. I do not know whether
    [2010_BUDGET_07_Kieran_ODonnell_FG.txt, 1365]           from the change in the  Christmas period. We suggested that the
      [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 550]        cut of €641, including the Christmas  payment. A couple on invalidity
      [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 638]        are on social welfare, the Christmas  payment is gone. Earnest lectures
      [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 998]     of emigration. Once again this Christmas, we will witness the scenes
     [2010_BUDGET_13_Ciaran_Cuffe_Green.txt, 911]       noted recently that over the  Christmas recess work will be done
[2010_BUDGET_14_Caoimhghin_OCaolain_SF.txt, 148]             will all be over by Christmas. If it is the last
>
```

Dictionaries and why we might use them

- 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 occurences of the key class
- Frequently involves lemmatization: transformation of all inflected word forms to their "dictionary look-up form" — more powerful than stemming

“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
  - WC = `wc, toilet, restroom, bathroom, jack, loo`
  - vote = `poll, suffrage, franchis*, ballot*, ^vot$`

Bridging qualitative and quantitative text analysis

- A hybrid procedure between qualitative and quantitative classification the fully automated end of the text analysis spectrum
- "Qualitiative" 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

Linguistic Inquiry and Word Count: Positive Emotion

```
> liwc$posemo
 [1] ":)"           "(:"             "53 like*"      "accept"
 [5] "accepta*"     "accepted"       "accepting"     "accepts"
 [9] "active"       "actively"       "admir*"        "ador*"
[13] "advantag*"    "adventur*"      "affection*"    "agree"
[17] "agreeable"    "agreeableness"  "agreeably"     "agreed"
[21] "agreeing"     "agreement*"     "agrees"        "alright*"
[25] "amaze*"       "amazing"        "amazingly"     "amor*"
[29] "amus*"        "aok"            "appreciat*"    "approv*"
[33] "assur*"       "attract"        "attracted"     "attracting"
[37] "attraction"   "attracts"       "award*"        "awesome"
[41] "beautiful"    "beautify"       "beauty"        "beloved"
[45] "benefic*"     "benefit"        "benefits"      "benefitt*"
[49] "benevolen*"   "best"           "bestest"       "bestie"
[53] "besties"      "better"         "bless*"        "bliss*"
[57] "bold"         "bolder"         "boldest"       "boldly"
[61] "bonus*"       "brave"          "braved"        "braver"
[65] "bravery"      "braves"         "bravest"       "bright"
[69] "brilliance*"  "brilliant"      "brilliantly"   "calm"
[73] "calmer"       "calmest"        "calming"       "care"
[77] "cared"        "carefree"       "cares"         "caring"
[81] "certain*"     "challeng*"      "champ*"        "charit*"
```

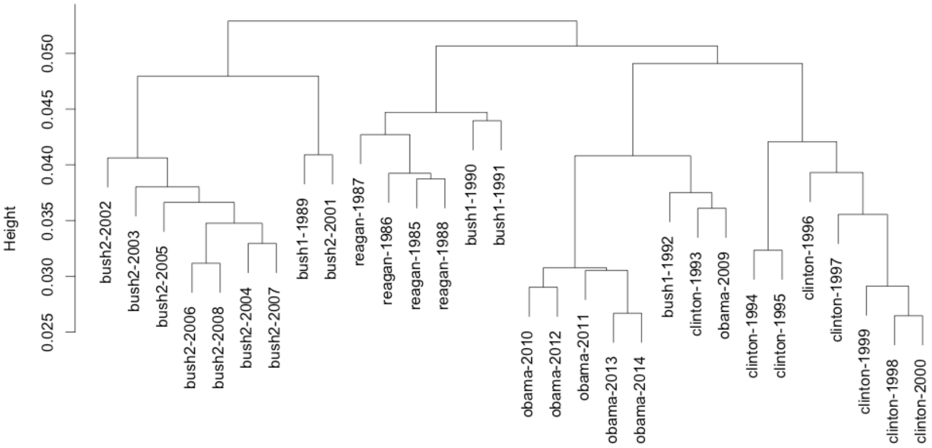
Linguistic Inquiry and Word Count: Negative Emotion

```
> liwc$negemo
 [1] ":("             "):"             "abandon*"       "abuse*"
 [5] "abusi*"         "ache*"          "aching*"        "advers*"
 [9] "afraid"         "aggravat*"      "aggress"        "aggressed"
[13] "aggresses"      "aggressing"     "aggression*"    "aggressive"
[17] "aggressively"   "aggressor*"     "agitat*"        "agoniz*"
[21] "agony"          "alarm*"         "alone"          "anger*"
[25] "angrier"        "angriest"       "angry"          "anguish*"
[29] "annoy"          "annoyed"        "annoying"       "annoys"
[33] "antagoni*"      "anxiety"        "anxious"        "anxiously"
[37] "anxiousness"    "apath*"         "appall*"        "apprehens*"
[41] "argh*"          "argu*"          "arrogan*"       "asham*"
[45] "assault*"       "asshole*"       "attack*"        "aversi*"
[49] "avoid*"         "awful"          "awkward"        "bad"
[53] "badly"          "bashful*"       "bastard*"       "battl*"
[57] "beaten"         "bereave*"       "bitch*"         "bitter"
[61] "bitterly"       "bitterness"     "blam*"          "bore*"
[65] "boring"         "bother*"        "broke"          "brutal*"
[69] "burden*"        "careless*"      "cheat*"         "coldly"
[73] "complain*"      "condemn*"       "confront*"      "confuse"
[77] "confused"       "confusedly"     "confusing"      "contempt*"
```

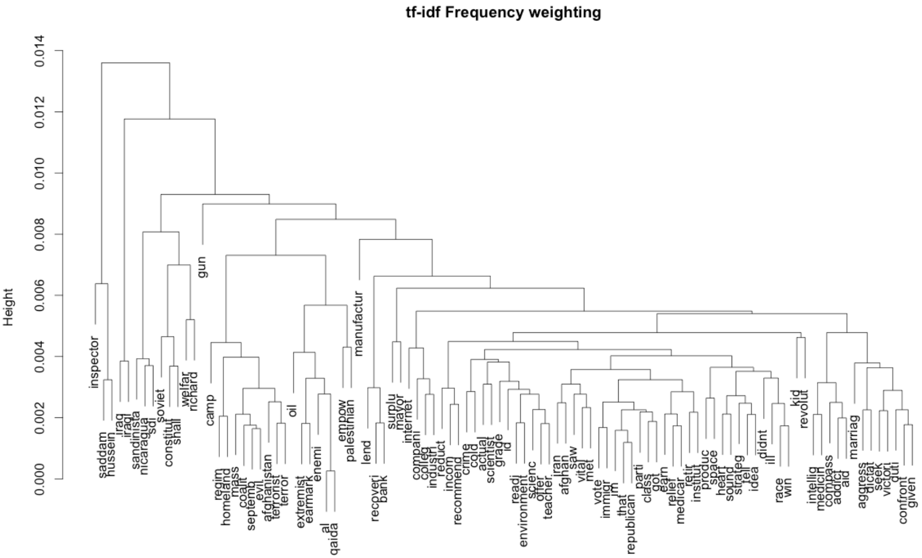
The idea of "clusters"

- Essentially: groups of items such that inside a cluster they are very similar to each other, but very different from those outside the cluster
- "unsupervised classification": cluster is not to relate features to classes or latent traits, but rather to estimate membership of distinct groups
- groups are given labels through post-estimation interpretation of their elements
- typically used when we do not and never will know the "true" class labels

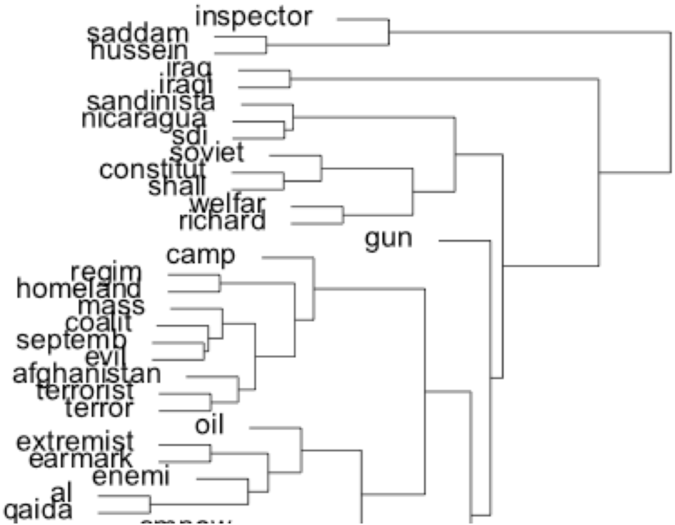
Dendrogram: Presidential State of the Union addresses



Dendrogram: Presidential State of the Union addresses



Dendrogram: Presidential State of the Union addresses



Topic Models

- Topic models are algorithms for discovering the main "themes" in an unstructured corpus
- Requires no prior information, training set, or special annotation of the texts
  – only a decision on K (number of topics)
- A probabalistic, generative advance on several earlier methods, "Latent Semantic Analysis" (LSA) and "probabalistic latent semantic indexing" (pLSI)

Uses and applications

- Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents
- Can be used to organize the collection according to the discovered themes
- Topic modeling algorithms can be applied to massive collections of documents
- Topic modeling algorithms can be adapted to many kinds of data. among other applications, they have been used to find patterns in genetic data, images, and social networks

Advantages over cruder methods

- parametric, so we get estimates of parameters for topic proportions in each document, and topic weights for each word
- can incorporate additional information hierarchically (e.g. using "structural" topic models)
- but we pay for these benefits in the form of far greater computational complexity

Latent Dirichlet Allocation

- The LDA model is a Bayesian mixture model for discrete data where topics are assumed to be uncorrelated (in "classic" LDA)
- LDA provides a generative model that describes how the documents in a dataset were created
- Each of the K *topics* is a distribution over a fixed vocabulary
- Each document is a collection of words, generated according to a multinomial distribution, one for each of K topics
- Inference consists of estimating a posterior distribution from a joint distribution based on the probability model from a combination of what is observed (words in documents) and what is hidden (topic and word parameters)

Latent Dirichlet Allocation

- So the process is, roughly:
  1. Choose a number of topics
  2. Choose a distribution of topics, and create a document from this distribution
  3. For each topic, generate words according to a distribution specific to that topic
- The goal of inference in LDA is to discover the topics from the collection of documents, and to estimate the relationship of words to these, *assuming this generative process*

Graphical model for LDA using plate notation

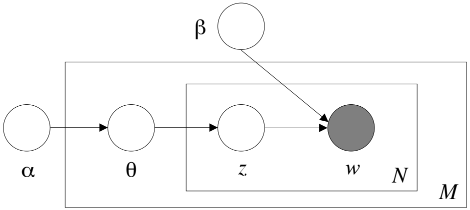


Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Example: Movie reviews

from Pang and Lee (2008)