Day 9: Text Analysis

ME314: Introduction to Data Science and Big Data Analytics

Jack Blumenau

12th August 2019

Day 10 Outline

Key Features of QTA

Basic QTA workflow: 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

	words										
docs	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

What role for "qualitative" analysis in QTA?

- Ultimately all reading of texts is qualitative, even when we count elements of the text or convert them into numbers
- QTA may involve human judgment in the construction of the feature-document matrix
- QTA may involve human judgment in the interpretation of the output of statistical models
- But QTA differs from more qualitiative approaches in that it:
  - Involves large-scale analysis of many texts, rather than close readings of few texts
  - Requires no *interpretation* of texts
- Uses a variety of statistical techniques to extract information from the document-feature matrix

Key feature of quantitative text analysis

- Conversion of textual features into a quantitative matrix
- A quantitative or statistical procedure to extract information from the quantitative matrix
- Summary and interpretation of the quantitative results

Key goals of quantitative text analysis

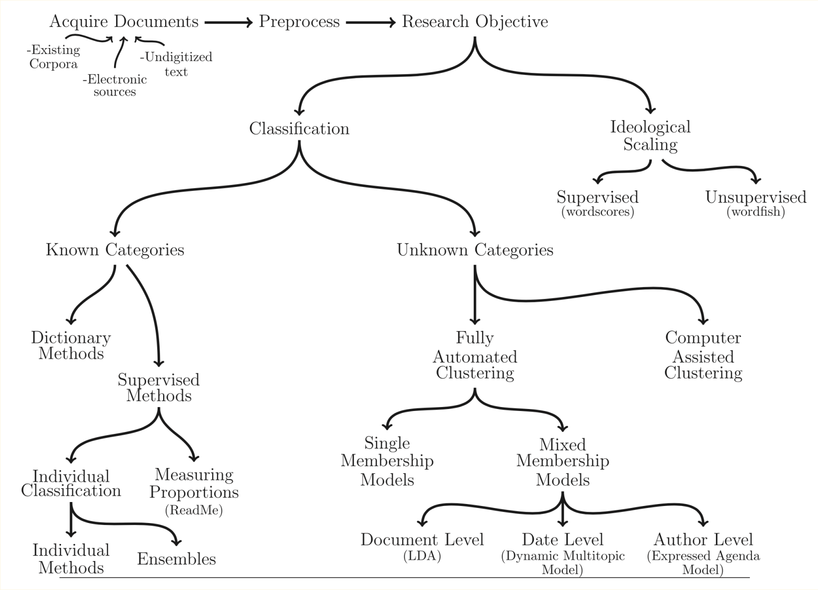
1. Prediction for 'downstream' tasks
   - Can we predict consumer behaviour from product reviews?
   - Can we predict football match outcomes using tweets?
2. Understanding of language use
   - Do men and women discuss political concepts differently?
   - How has the meaning of words changed over time?
3. Measurement of latent constructs
   - Can we infer student sophistication from the *complexity* of their writing?
   - Which set of *topics* characterises a corpus of texts?

Today focuses mostly on simple ways of approaching 3. You will cover 1 tomorrow, and bits of 2 and 3 on Wednesday.

3 guiding priciples for QTA

- All quantitative models for text are wrong, but some are useful
- Quatitative models for text augment, but do not replace, humans
- Validation is key

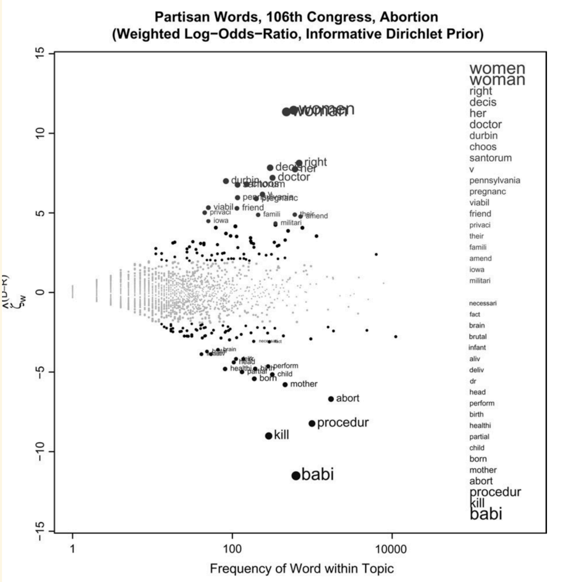
An overview of text-as-data-methods



Example: Wordclouds

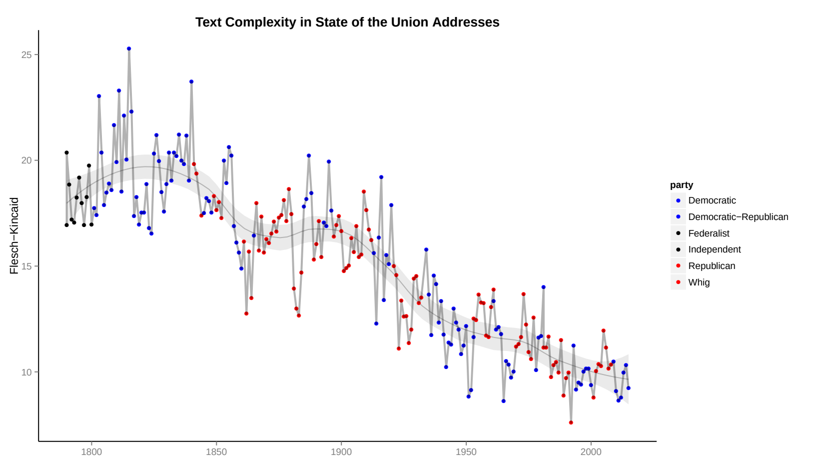
(from Herzog and Benoit EPSA 2013)

Example: Better Wordclouds

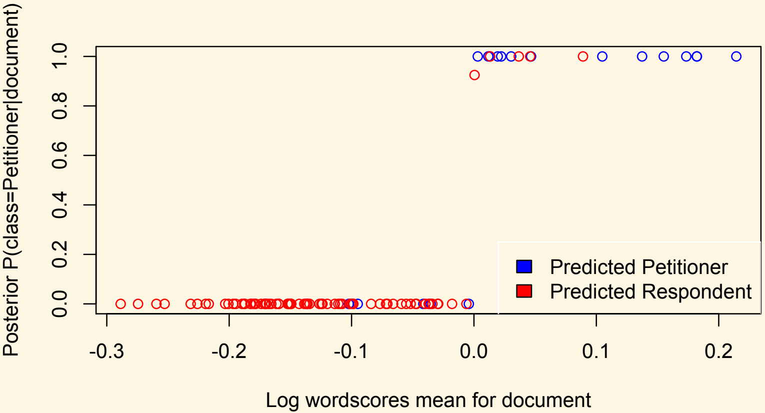


(from Munroe et al., 2009)

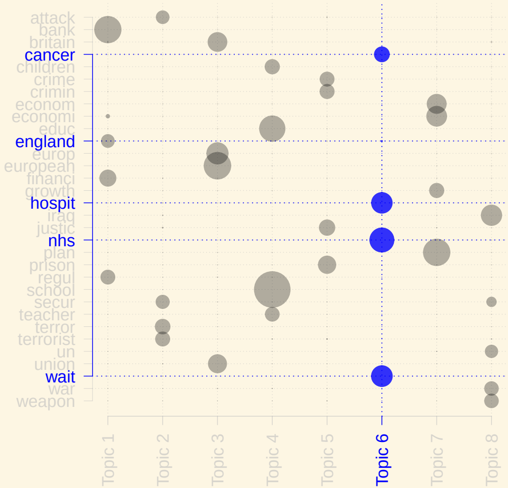
Example: Text complexity



Example: Document classification



Example: Exploring the topics of a group of texts



This 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
  - disregard grammar, disregard word order, just pay attention to word frequencies
  - 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

Bag of words assumption

- Consider two sentences:
  1. Time flies like an arrow.
  2. Fruit flies like a banana.
- Convert these into a bag-of-words feature matrix:

	time	flies	fruit	like	an	a	banana	arrow
Sentence 1	1	1	0	1	1	0	0	1
Sentence 2	0	1	1	1	0	1	1	0

- The dependency structure between words in each sentence is lost
- The word “flies” has a different meaning in the two sentences (metaphorical versus literal), but both sentences score a 1 here
- The “joke” is no longer funny

Key features of quantitative text analysis

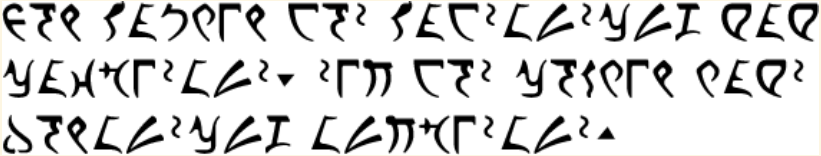
1. Selecting texts: Defining the *corpus*
2. Conversion of texts into a common electronic format
3. Defining documents: deciding what will be the unit of analysis (document, paragraph, sentence, etc)

Key features of quantitative text analysis

4. Defining features. These can take a variety of forms, including tokens, equivalence classes of tokens (dictionaries), selected phrases, human-coded segments (of possibily 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

Extreme forms of QTA

- Fully automated technique with minimal human intervention or judgment calls – only with regard to reference text selection
- Methods can "discover" topics with little human supervision
- Language-blind: can scaling anything that occurs with regular patterns (even without knowing what these mean)
- Could potentially work on texts like this:



http://www.kli.org

- We will focus on these methods tomorrow (not for klingon)

Some key basic concepts

(text) corpus a large and structured set of texts for analysis

types for our purposes, a unique word

tokens any word – so token count is total words

stems words with suffixes removed

lemmas canonical word form

keys such as dictionary entries, where the user defines a set of equivalence classes that group different word types

Some more key basic concepts

"key" words Words selected because of special attributes, meanings, or rates of occurrence

stop words Words that are designated for exclusion from any analysis of a text

readability provides estimates of the readability of a text based on word length, syllable length, etc.

complexity A word is considered "complex" if it contains three syllables or more

diversity (lexical diversity) A measure of how many types occur per fixed word rate (a normalized vocabulary measure)

Documents and Features

Strategies for selecting units of textual analysis

- Words
- n-word sequences
- pages
- paragraphs
- Natural units (a speech, a poem, a manifesto)
- Key: depends on the research design

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.

  ```
  [\emph{Rindfleischetikettierungs\"uberwachungsaufgaben\"ubert
  ```

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.

  [\emph{Rindfleischetikettierungs\"uberwachungsaufgaben\"uber

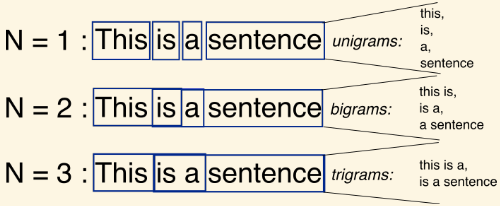
(the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef)

Defining Features

- "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18岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

- n-grams: contiguous sequence of words from document (1-gram, unigram; 2-gram, bigram, etc)



Defining Features

- (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

POS tagging in R

```r
library(spacyr)
text <- "Jack Blumenau is currently Lecturer in Political Science and Quantitative Methods
        at the Department of Political Science at UCL"
spacy_parse(text)
```

```
##           token        lemma   pos   entity
## 1          Jack         Jack PROPN PERSON_B
## 2      Blumenau     Blumenau PROPN PERSON_I
## 3            is           be  VERB
## 4     currently    currently   ADV
## 5      Lecturer     Lecturer PROPN    ORG_B
## 6            in           in   ADP    ORG_I
## 7     Political    Political PROPN    ORG_I
## 8       Science      Science PROPN    ORG_I
## 9           and          and CCONJ
## 10 Quantitative Quantitative PROPN PERSON_B
## 11      Methods      Methods PROPN PERSON_I
## 12           at           at   ADP
## 13          the          the   DET    ORG_B
## 14   Department   Department PROPN    ORG_I
## 15           of           of   ADP    ORG_I
## 16    Political    Political PROPN    ORG_I
## 17      Science      Science PROPN    ORG_I
## 18           at           at   ADP
## 19          UCL          UCL PROPN    ORG_B
## 20            .            . PUNCT
```

Strategies for feature selection

- This can lead to a lot of features!
- An example (small) corpus:
  - 17,129 speeches made in the final month of 2016 in the House of Commons
  - $\approx$ 3 million total words
  - 46998 unique words
  - 468244 unique 1-gram and 2-gram sequences

Strategies for feature selection

- document frequency How many documents in which a term appears
- term frequency How many times does the term appear in the corpus
- deliberate disregard Use of "stop words": words excluded because they represent linguistic connectors of no substantive content
- purposive selection Use of a *dictionary* of words or phrases

Common English stop words

```r
library(quanteda)
cat(paste0(stopwords("en"), collapse = "; "))
```

i; me; my; myself; we; our; ours; ourselves; you; your; yours; yourself; yourselves; he; him; his; himself; she; her; hers; herself; it; its; itself; they; them; their; theirs; themselves; what; which; who; whom; this; that; these; those; am; is; are; was; were; be; been; being; have; has; had; having; do; does; did; doing; would; should; could; ought; i'm; you're; he's; she's; it's; we're; they're; i've; you've; we've; they've; i'd; you'd; he'd; she'd; we'd; they'd; i'll; you'll; he'll; she'll; we'll; they'll; isn't; aren't; wasn't; weren't; hasn't; haven't; hadn't; doesn't; don't; didn't; won't; wouldn't; shan't; shouldn't; can't; cannot; couldn't; mustn't; let's; that's; who's; what's; here's; there's; when's; where's; why's; how's; a; an; the; and; but; if; or; because; as; until; while; of; at; by; for; with; about; against; between; into; through; during; before; after; above; below; to; from; up; down; in; out; on; off; over; under; again; further; then; once; here; there; when; where; why; how; all; any; both; each; few; more; most; other; some; such; no; nor; not; only; own; same; so; than; too; very; will

- But no list should be considered universal...

Common English stop words

```r
library(quanteda)
cat(paste0(stopwords("smart"), collapse = "; "))
```

a; a's; able; about; above; according; accordingly; across; actually; after; afterwards; again; against; ain't; all; allow; allows; almost; alone; along; already; also; although; always; am; among; amongst; an; and; another; any; anybody; anyhow; anyone; anything; anyway; anyways; anywhere; apart; appear; appreciate; appropriate; are; aren't; around; as; aside; ask; asking; associated; at; available; away; awfully; b; be; became; because; become; becomes; becoming; been; before; beforehand; behind; being; believe; below; beside; besides; best; better; between; beyond; both; brief; but; by; c; c'mon; c's; came; can; can't; cannot; cant; cause; causes; certain; certainly; changes; clearly; co; com; come; comes; concerning; consequently; consider; considering; contain; containing; contains; corresponding; could; couldn't; course; currently; d; definitely; described; despite; did; didn't; different; do; does; doesn't; doing; don't; done; down; downwards; during; e; each; edu; eg; eight; either; else; elsewhere; enough; entirely; especially; et; etc; even; ever; every; everybody; everyone; everything; everywhere; ex; exactly; example; except; f; far; few; fifth; first; five; followed; following; follows; for; former; formerly; forth; four; from; further; furthermore; g; get; gets; getting; given; gives; go; goes; going; gone; got; gotten; greetings; h; had; hadn't; happens; hardly; has; hasn't; have; haven't; having; he; he's; hello; help; hence; her; here; here's; hereafter; hereby; herein; hereupon; hers; herself; hi; him; himself; his; hither; hopefully; how; howbeit; however; i; i'd; i'll; i'm; i've; ie; if; ignored; immediate; in; inasmuch; inc; indeed; indicate; indicated; indicates; inner; insofar; instead; into; inward; is; isn't; it; it'd; it'll; it's; its; itself; j; just; k; keep; keeps; kept; know; knows; known; l; last; lately; later; latter; latterly; least; less; lest; let; let's; like; liked; likely; little; look; looking; looks; ltd; m; mainly; many; may; maybe; me; mean; meanwhile; merely; might; more; moreover; most; mostly; much; must; my; myself; n; name; namely; nd; near; nearly; necessary; need; needs; neither; never; nevertheless; new; next; nine; no; nobody; non; none; noone; nor; normally; not; nothing; novel; now; nowhere; o; obviously; of; off; often; oh; ok; okay; old; on; once; one; ones; only; onto; or; other; others; otherwise; ought; our; ours; ourselves; out; outside; over; overall; own; p; particular; particularly; per; perhaps; placed; please; plus; possible; presumably; probably; provides; q; que; quite; qv; r; rather; rd; re; really; reasonably; regarding; regardless; regards; relatively; respectively; right; s; said; same; saw; say; saying; says; second; secondly; see; seeing; seem; seemed; seeming; seems; seen; self; selves; sensible; sent; serious; seriously; seven; several; shall; she; should; shouldn't; since; six; so; some; somebody; somehow; someone; something; sometime; sometimes; somewhat; somewhere; soon; sorry; specified; specify; specifying; still; sub; such; sup; sure; t; t's; take; taken; tell; tends; th; than; thank; thanks; thanx; that; that's; thats; the; their; theirs; them; themselves; then; thence; there; there's; thereafter; thereby; therefore; therein; theres; thereupon; these; they; they'd; they'll; they're; they've; think; third; this; thorough; thoroughly; those; though; three; through; throughout; thru; thus; to; together; too; took; toward; towards; tried; tries; truly; try; trying; twice; two; u; un; under; unfortunately; unless; unlikely; until; unto; up; upon; us; use; used; useful; uses; using; usually; uucp; v; value; various; very; via; viz; vs; w; want; wants; was; wasn't; way; we; we'd; we'll; we're; we've; welcome; well; went; were; weren't; what; what's; whatever; when; whence; whenever; where; where's; whereafter; whereas; whereby; wherein; whereupon; wherever; whether; which; while; whither; who; who's; whoever; whole; whom; whose; why; will; willing; wish; with; within; without; won't; wonder; would; would; wouldn't; x; y; yes; yet; you; you'd; you'll; you're; you've; your; yours; yourself; yourselves; z; zero

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`

example II: saw
Lemmatization may covert to either see or saw depending on whether usage was as a noun or a verb

Feature selection in practice

debates18 includes 89416 speeches made in the 2018 in the House of Commons

```
library(quanteda)

# Construct DFM
debate_dfm <- dfm(debates18$texts)

# Stopwords
debate_dfm_stop <- dfm_remove(debate_dfm, pattern = stopwords("en"))

# Stem
debate_dfm_stem <- dfm_wordstem(debate_dfm)

# Trim (word frequency)
debate_dfm_trim1 <- dfm_trim(debate_dfm, min_termfreq = 5)

# Trim (document frequency)
debate_dfm_trim2 <- dfm_trim(debate_dfm, min_docfreq = 0.001,
                             docfreq_type = "prop")
```

Feature selection in practice

- 72404 unique words
  - After stopwords: 72232
  - ... and stemming: 49108
  - ... and removing features that appear fewer than 5 times: 29202
  - ... and removing features in fewer than 0.001 documents: 6482
- Feature selection matters! See Denny and Spirling, 2017

```
- Just seven (binary) preprocessing decisions leads to a tot
- These selection decisions can have substantive implication
```

Descriptive Text Analysis

Basic descriptive summaries of text

Length in characters, words, unique words, lines, sentences, paragraphs, pages, sections, chapters, etc.

Key words in context provide how words or phrases are used in a corpus.

Readability statistics Use a combination of syllables and sentence length to indicate "readability" in terms of complexity

Vocabulary diversity At its simplest involves measuring a *type-to-token ratio* (TTR) where unique words are types and the total words are tokens

Word (relative) frequency Measures how often some word occurs relative to some other word

Describe your text data!

Speaker	Party	Tokens	Types
Brian Cowen	FF	5,842	1,466
Brian Lenihan	FF	7,737	1,644
Ciaran Cuffe	Green	1,141	421
John Gormley (Edited)	Green	919	361
John Gormley (Full)	Green	2,998	868
Eamon Ryan	Green	1,513	481
Richard Bruton	FG	4,043	947
Enda Kenny	FG	3,863	1,055
Kieran ODonnell	FG	2,054	609
Joan Burton	LAB	5,728	1,471
Eamon Gilmore	LAB	3,780	1,082
Michael Higgins	LAB	1,139	437
Ruairi Quinn	LAB	1,182	413
Arthur Morgan	SF	6,448	1,452
Caoimhghin O'Caolain	SF	3,629	1,035
All Texts		49,019	4,840
Min		919	361
Max		7,737	1,644
Median		3,704	991
Hapaxes with Gormley Edited		67	
Hapaxes with Gormley Full Speech		69	

Key Words in Context

KWIC *Key words in context* Refers to the most common format for concordance lines. A KWIC shows how a word or phrase is used across various texts in the corpus

lime (14)

79[C.10]	4	/Which was builded of	**lime**	and sand;/Until they came to
247A.6	4	/That was well biggit with	**lime**	and stane.
303A.1	2	bower,/Well built wi	**lime**	and stane,/And Willie came
247A.9	2	/That was well biggit wi	**lime**	and stane,/Nor has he stoln
305A.2	1	a castell biggit with	**lime**	and stane,/O gin it stands not
305A.71	2	is my awin,/I biggit it wi	**lime**	and stane;/The Tinnies and
79[C.10]	6	/Which was builded with	**lime**	and stone.
305A.30	1	a prittie castell of	**lime**	and stone,/O gif it stands not
108.15	2	/W*hi*ch was made both of	**lime**	and stone,/Shee tooke him by
175A.33	2	castle then,/Was made of	**lime**	and stone;/The vttermost
178[H.2]	2	near by,/Well built with	**lime**	and stone;/There is a lady
178F.18	2	built with stone and	**lime**	!/But far mair pittie on Lady
178G.35	2	was biggit wi stane and	**lime**	!/But far mair pity o Lady
2D.16	1	big a cart o stane and	**lime**	,/Gar Robin Redbreast trail it

The idea of a local “context” is central to more advanced QTA analyses such as word-embeddings.

Another KWIC Example: *Seale et al (2006)*

Table 3
Example of Keyword in Context (KWIC) and associated word clusters display

Extracts from Keyword in Context (KWIC) list for the word 'scan'

An MRI **scan** then indicated it had spread slightly

Fortunately, the MRI **scan** didn't show any involvement of the lymph nodes

3 very worrying weeks later, a bone **scan** also showed up clear.

The bone **scan** is to check whether or not the cancer has spread to the bones.

The bone **scan** is done using a type of X-ray machine.

The results were terrific, CT **scan** and pelvic X-ray looked good

Your next step appears to be to await the result of the **scan** and I wish you well there.

I should go and have an MRI **scan** and a bone **scan**

Three-word clusters most frequently associated with keyword 'scan'

N	Cluster	Freq
1	A bone scan	28
2	Bone scan and	25
3	An MRI scan	18
4	My bone scan	15
5	The MRI scan	15
6	The bone scan	14
7	MRI scan and	12
8	And Mri scan	9

KIWC in quanteda

```r
library(quanteda)
debate_corpus <- corpus(debates18, text_field = "texts")
head(kwic(debate_corpus, "European"))
```

```
##
##  [text3, 102]  still in negotiations with the | European | Union in terms of delivering
##   [text5, 44]     day of consideration of the | European | Union Bill by the Committee
##   [text7, 55]           remain a party to the | European | convention on human rights after
##   [text7, 73]        is also reflected in the | European | Union Act 2018, which
##  [text15, 18] constituency voted to leave the | European | Union in the referendum.
##   [text37, 2]                             The | European | Union's negotiating position on the
```

Lexical Diversity

- Basic measure is the TTR: Type-to-Token ratio

$$TTR = \frac{\text{Number of Types}(V)}{\text{Number of Tokens}(N)}$$

- Problem 1: Very sensitive to overall document length, as shorter texts may exhibit fewer word repetitions
- Problem 2: length may relate to the introdution of additional subjects, which will also increase richness

Lexical diversity and corpus length

- In natural language text, the rate at which new types appear is very high at first, but diminishes with added tokens

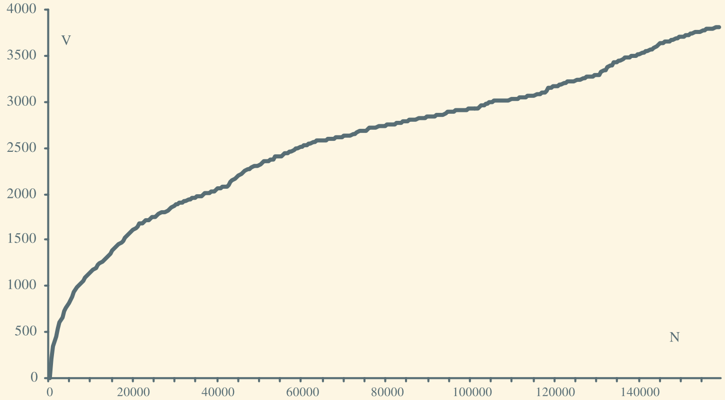
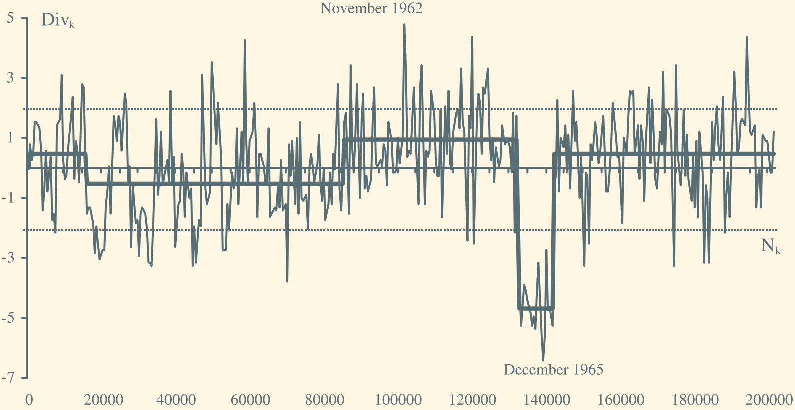


Fig. 1. Chart of vocabulary growth in the tragedies of Racine (chronological order, 500 token intervals).

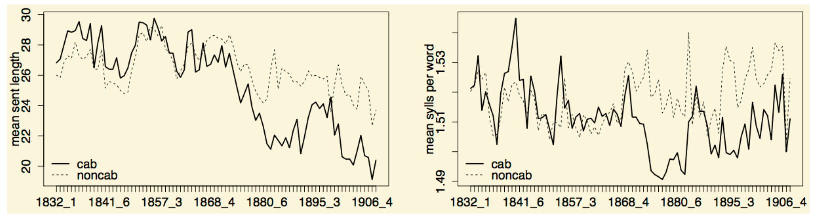
Lexical Diversity Example

- Variations use automated segmentation – here approximately 500 words in a corpus of serialized, concatenated weekly addresses by de Gaulle (from Labb'e et. al. 2004)
- While most were written, during the period of December 1965 these were more spontaneous press conferences



Readability Example

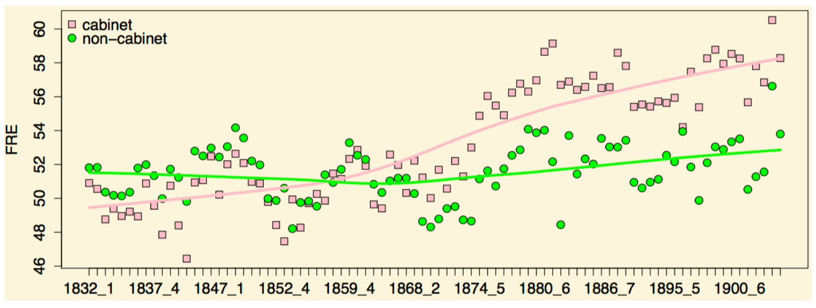
- Most commonly used readability scores focus on a combination of syllables and sentence length
  - Shorter sentences = more readable
  - Fewer syllables = more readable
- Research question: Do Members of Parliament use less complex language when appealing to a more diverse electorate?
- Context: Parliamentary speeches before and after the Great Reform Act (1867) (See Spirling, 2015, Journal of Politics)



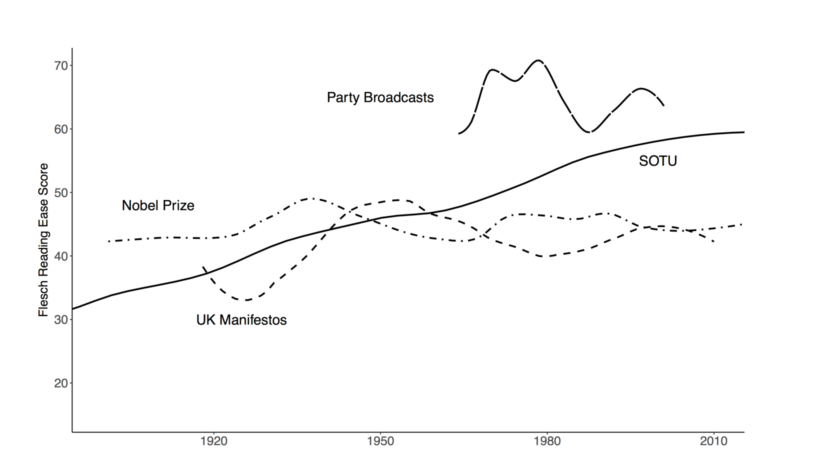
Readability Example

Flesch score:

$$206.835 - 1.105\left(\frac{\text{total number of words}}{\text{total number of sentences}}\right) - 84.6\left(\frac{\text{total number of syllables}}{\text{total number of words}}\right)$$



Readability Example (II)



Text summaries in practice

Thankfully, `quanteda` makes it trivial to calculate many of these statistics...

```
# Number of tokens
debate_tokens <- ntoken(debate_corpus)

# Number of types
debate_types <- ntype(debate_corpus)

# Token-type ratio
debate_ttr <- textstat_lexdiv(debate_corpus, "TTR")

# Readability
debate_read <- textstat_readability(debate_corpus,
                                    measure = "Flesch")
```

Break

Go here: https://jblumenau.shinyapps.io/validate/

Content Analysis

Hand-coding: "Classic" content analysis

- Key feature: use of "human" coders to implement a pre-defined coding scheme, by reading and coding texts
- Human decision-making is the central feature of coding decisions, not a computer or other mechanized tool
- Example: hand-coding sentences into pre-defined categories
- Alternative 1: dictionary-based approaches (somewhat more automated)
  - More on this in about 2 minutes
- Alternative 2: inductive scaling or clustering of texts from the quantitative matrix (entirely automated)
  - More on this tomorrow and Wednesday

Hand-coding: “Classic” content analysis

- Validity is usually the objective, rather than reliability
  - Validity: am I measuring what I am claiming to measure?
  - Reliability: am I able to reliably replicate my coding?
- Another motivating factor could be ease of use, or the difficulty of implementing an automated procedure
- May be *computer-assisted*, especially for unitization
- Many common “CATA” tools exist – e.g. QDAMiner

Components of classical content analysis designs

Unitizing The systematic distinguishing of segments of text that are of interest to the analysis.

Sampling Choice (and justification of the choice) of text units to sample, from population of possible text units.

Coding Classifying each coded unit of text from the sample according to the pre-defined category scheme.

Summarizing Reducing the coded data to summary quantities of interest.

Inference and reporting The final steps wherein the analyzed results are used to generalize about social world, and communicating these results to others.

Dictionary Analysis

Motivation

Are female politicians less aggressive than male politicians?

A repeated claim in the qualitative literature on gender and politics is that female politicians have a distinct style from male politicians.

Crucially, many scholars (and observers) of legislative politics argue that women are less aggressive in the context of debate than are their male colleagues.

Most of the evidence for these claims is taken from small-N classical content analysis studies.

We will review this question by applying an existing sentiment dictionary to a large-N corpus of parliamentary texts.

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

Rationale for dictionaries

- Rather than count words that occur, pre-define words associated with specific meanings
- Two components:
  1. key: the label for the equivalence class for the concept or canonical term e.g. "dog"
  2. values: (multiple) terms or patterns that are declared equivalent occurences of the key class e.g. "Dalmatian", "Labrador", "Poodle"
- Frequently involves lemmatization: transformation of all inflected word forms to their "dictionary look-up form" – more powerful than stemming

Counting words

At its simplest, a dictionary is just a list of words ($m = 1, ..., M$) that is related to a common concept.

Aggression
stupid
dishonest
lier
idiot
ignorant
hate
fight
battle

Counting words

Applying a dictionary to a corpus of texts ($i = 1, ..., N$) simply requires counting the number of times each word occurs in each text and summing them.

If W_{im} is a vector measuring 1 if word m appears in text i and 0 otherwise, then the dictionary score for document i is:

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

Or, the proportion of words in document i that appear in the dictionary.

Counting words

"That statement is as barbaric as it is downright stupid; it is nothing more than an ignorant, cruel and deliberate misconception to hide behind."

$$t_i = \frac{\sum_{m=1}^{M} W_{im}}{N_i} = \frac{1+1}{14} = 0.14$$

Counting *weighted* words

A slight development on this would be to assign each word in the dictionary a weight which reflects something about the importance of the word to the concept

Aggression	Weight
stupid	.6
dishonest	.2
lie	.5
idiot	.7
ignorant	.3
brutal	.4
violence	.5

Note that weights are implicit in *all* dictionary approaches. Typically, all words are counted equally which implies a score of 1 for all words. This is not necessarily correct!

Counting *weighted* words

We can adjust the previous formula to incorporate the weights (s_m):

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

Why normalise by N_i? Some texts will be longer than others and we do not want these texts to mechanically be assigned higher scores.

Counting *weighted* words

"That statement is as barbaric as it is downright stupid; it is nothing more than an ignorant, cruel and deliberate misconception to hide behind."

$$t_i = \frac{\sum_{m=1}^{M} W_{im}}{N_i} = \frac{0.6 + 0.3}{14} = 0.06$$

Weights or no weights?

The vast majority of applications of dictionary methods in the social science literature and in industry applications use unweighted (or, equally weighted) dictionary approaches.

Why learn this then?

1. The equal weighting assumption is not necessarily reasonable or effective
2. The idea of assigning weights to words is something that will come up tomorrow in the context of supervised learning and again on Wednesday in the context of topic models

Advantages of dictionaries: Many existing implementations

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 82 language dimensions
- $\approx$ 4,500 words and word stems, each defining one or more word categories
- For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb.
- Hierarchical: so "anger" is 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 and Chung, 2009)

	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

Advantages of dictionaries: 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*

Advantages of dictionaries: Fast and easy to apply

Here, `debates` is a `data.frame` of parliamentary debates, which contains about a million speeches.

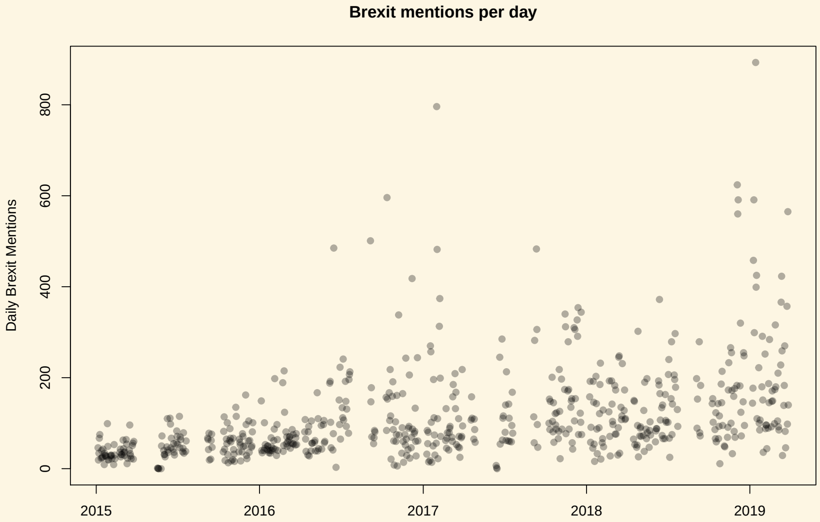
```
library(quanteda)
brexit_dict <- dictionary(list(brexit = c("brexit",
                                          "leave",
                                          "remain",
                                          "sovereignty",
                                          "control")))

dictionary_dfm <- dfm(debates$texts, dictionary = brexit_dict)
```

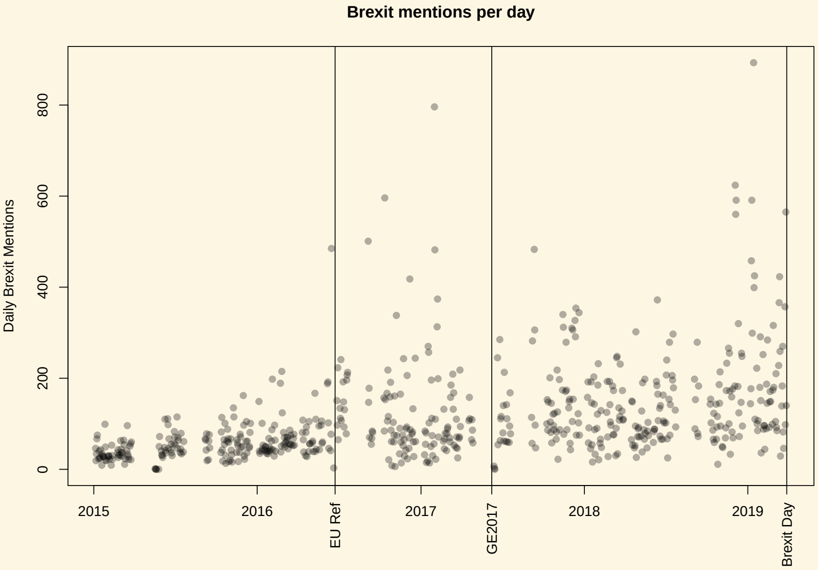
In contrast to some of the other methods we will study, dictionaries can by easily applied to thousands of texts in a matter of seconds.

The code above runs in about a minute.

Advantages of dictionaries: Fast and easy to apply



Advantages of dictionaries: Fast and easy to apply



Disdvantages of dictionaries

- Problem 1: polysemes – words that have multiple meanings
  - 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
  - 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 2: Dictionaries often lack important negative financial words, for example; *felony*, *litigation*, *restated*, *misstatement*, and *unanticipated*
- Problem 3: Some dictionaries might do more to pick up the *topic* of a document than the *tone* of a document

Disdvantages of dictionaries

> *"That statement is as barbaric as it is downright stupid; it is nothing more than an ignorant, cruel and deliberate misconception to hide behind."*
>
> *"Terrible acts of brutality and violence have been carried out against the Rohingya people."*

- Dictionaries may miss words that are important to the concept
  - "barbaric" is probably an aggressive word in this context
- Dictionaries do not typically capture modifiers
  - "downright" is an intensifier (also: negators like "not good")
- Dictionaries often fail to capture all synonyms
  - "deliberate misconception" is parliamentary language for "lie"
- Dictionaries may not capture the relevant concept
  - brutality/violence: descriptions, rather than *expressions*, of aggression

Vaidation

What kind of validation might we use here?

Applying dictionaries outside the domain for which they were developed can lead to errors.

One way of assessing the seriousness of these errors is to conduct **validation tests**

Main idea: are the texts that are flagged by the dictionary more representative of the relevant concept than other texts?

Applying dictionaries in quanteda

```r
library(quanteda)
aggression_words <- read.csv("aggression_words.csv")[,1]
aggression_texts <- read.csv("aggression_texts.csv")[,1]
```

1. `aggression_words` is a vector of 222 words from the an existing "Aggression" dictionary
2. `aggression_texts` is a vector of 10937 sentences from parliamentary speeches

Our goal is to use `aggression_words` to score the texts in `aggression_texts`.

Applying dictionaries in quanteda

First we convert the texts to a `corpus` object:

```
aggression_corpus <- corpus(aggression_texts)
```

And the words to a `dictionary` object:

```
aggression_dictionary <- dictionary(list(aggression = aggression_words))
```

Finally, we "apply" the dictionary to the corpus using the `dfm` function:

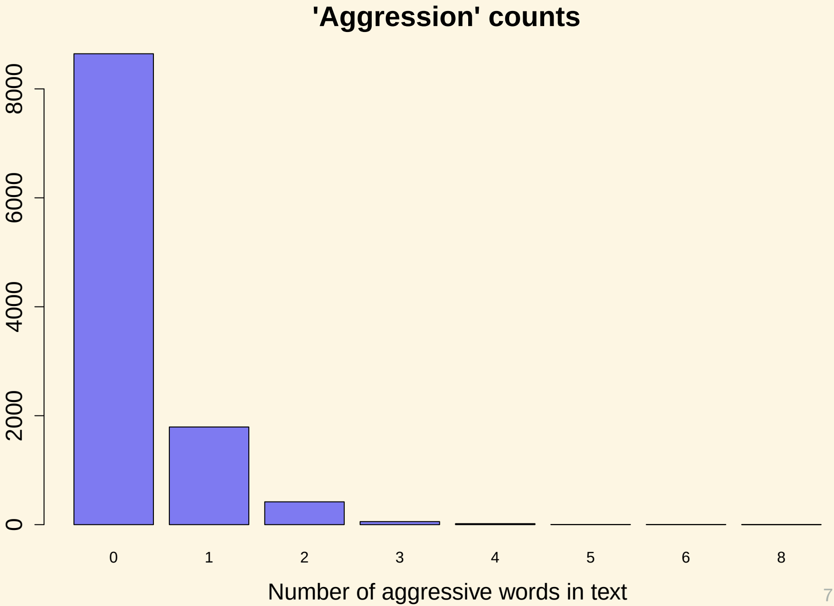
```
aggression_dfm <- dfm(x = aggression_corpus,
                      dictionary = aggression_dictionary)

print(aggression_dfm)
```

```
## Document-feature matrix of: 10,937 documents, 1 feature (79.0% sparse).
```

`aggression_dfm` is a document-feature matrix, where the only "feature" is the dictionary counts

Applying dictionaries in quanteda



Applying dictionaries in quanteda

Finally, we can calculate the score by dividing the dictionary counts by the number of words in each text:

```
aggression_proportions <- as.numeric(aggression_dfm[,1])/
                                     ntoken(aggression_corpus)

summary(aggression_proportions)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.000000 0.000000 0.008111 0.000000 0.190476
```

Face validity (1)

Intuition: Does the measure vary in sensible ways?

In this case, one obvious test is whether MPs speeches are more aggressive during Prime Minister's Questions (PMQs).

Face validity (1)

```r
coef(summary(lm(aggression_proportions ~ pmq_dummy)))
```

```
##                Estimate   Std. Error  t value     Pr(>|t|)
## (Intercept) 0.008110713 0.0001847766 43.89470 0.00000e+00
## pmq_dummy   0.008362690 0.0004699469 17.79497 5.26835e-70
```

There is clear evidence that PMQ debates tend to have higher levels of aggressive language than other debates.

Face validity (2)

How does this approach perform? Let's look at the top sentences:

	score	text
text3998	0.19	I fully appreciate that it is the Opposition's job to oppose, but there are times when opposition is destructive.
text7416	0.18	We unequivocally condemn Hamas's dreadful and murderous rocket attacks and defend Israel's right to defend itself.
text2941	0.14	They were asking ridiculous prices, because they had the sole remedy for a complaint, so could exploit that situation.
text106	0.13	Terrible acts of brutality and violence have been carried out against the Rohingya people.
text144	0.13	The motion condemns the early release scheme for those who have assaulted police officers.

While some seem reasonable, others indicate that we are picking up topic rather than tone.

Human validation as a gold standard

What is the "gold standard" for judging whether our dictionary works?

Typically, we compare the performance of our method to human judgements of our concept of interest.

In essence, we can ask people to rate sentences according to their "aggressiveness" and see whether this correlates with our measure.

Key assumption: Human coders can accurately and reliably recognise instances of aggression in text.

Which of these questions is easier?

1. On a scale from 0 to 100, how aggressive is this sentence?
   - "I regard it as an essential weapon in the armoury of the fight against terrorism."
2. Which of these sentences is more aggressive?
   - "I regard it as an essential weapon in the armoury of the fight against terrorism."
   - "I also welcome the fact that the Bill will encourage more young people to take advantage of the programme."

Paired comparisons tend to give more useful and reliable information than single ratings.

Set-up

1. Apply 8 basic QTA measures (including 6 dictionaries) to 8 million sentences
   - Aggression
   - Positive Emotion
   - Negative Emotion
   - Fact
   - Anecdote
   - Complexity
   - Repetition
2. Score each sentence using uniform word weights
3. Present pairs of sentences to human coders and ask them to select which sentence is most representative of a certain concept

Validation app

Go here: https://jblumenau.shinyapps.io/validate/

Validation measure

Does the difference in sentence-level dictionary scores predict human judgements?

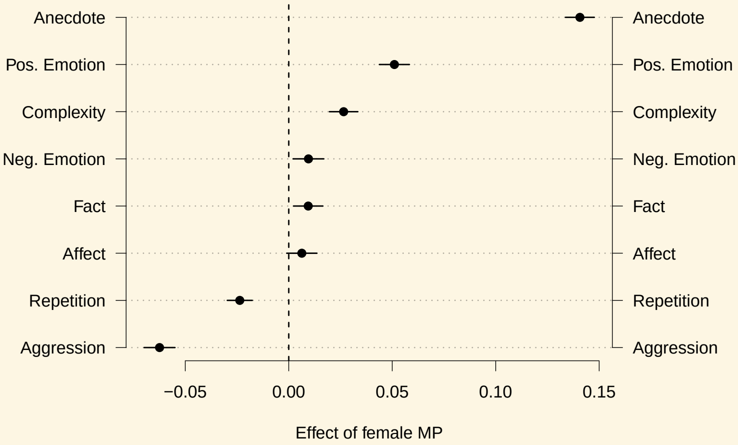
- Sample pairs of sentences from the corpus
  - Score each pair as $\text{Diff}_i = \text{Style Score}_{1i} - \text{Style Score}_{2i}$
- Randomly present to human coders, code (Y_i) whether:
  - Sentence one is more <style> (1)
  - About the same (0)
  - Sentence two is more <style> (-1)
- Calculate the relationship between human coding and dictionaries by:
  - $Y_i = \alpha + \beta \text{Diff}_i$
  - $Cor(Y_i, \text{Diff}_i)$
- Repeat for each dictionary

Live results

The excitement is palpable.

Are women less aggressive?

Let's believe for a second that our validation strategy worked.



Conclusion

Conclusion

- QTA allows us to draw inferences from very large collections of text without (too much) human interpretation
- All quantitative models of text are wrong, but some are useful
- Simple quantitative metrics of text can be very revealing
- Supervised text models, such as Naive Bayes, are easy to apply and can be very helpful in dealing with huge corpora
- `quanteda` is awesome
- Validation is very important!

Road map

For the rest of the week we will build upon the tools we covered today

- Tomorrow: Supervised learning with text, and text scaling models
- Wednesday: Unsupervised text models (topic models)
- Thursday: Data from the web