Day 9: Text Analysis

ME314: Introduction to Data Science and Machine Learning

Jack Blumenau

12th August 2019
Day 10 Outline

Key Features of QTA

Documents and Features

Descriptive Text Analysis

Content Analysis

Dictionary Analysis

Validation

Conclusion
Key Features of QTA
An economic miracle is taking place in the United States, and the only thing that can stop it are foolish wars, politics, or ridiculous partisan investigations.

The United States of America right now has the strongest, most durable economy in the world. We're in the middle of the longest streak of private sector job creation in history.

To build a prosperous future, we must trust people with their own money and empower them to grow our economy.

We reinvented Government, transforming it into a catalyst for new ideas that stress both opportunity and responsibility and give our people the tools they need to solve their own problems.

An economic miracle is taking place in the United States, and the only thing that can stop it are foolish wars, politics, or ridiculous partisan investigations.

To build a prosperous future, we must trust people with their own money and empower them to grow our economy.

Source texts

Processed text as a document-feature matrix

<table>
<thead>
<tr>
<th>documents</th>
<th>features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinton-2000</td>
<td>economy united wall crime climate</td>
</tr>
<tr>
<td>Bush-2008</td>
<td>6 4 0 0 1</td>
</tr>
<tr>
<td>Obama-2016</td>
<td>16 4 1 0 4</td>
</tr>
<tr>
<td>Trump-2019</td>
<td>5 19 6 2 0</td>
</tr>
</tbody>
</table>

Quantitative analysis and inference

- Describing texts quantitatively or stylistically
- Identifying keywords
- Measuring ideology or sentiment in documents
- Mapping semantic networks
- Identifying topics and estimating their prevalence
- Measuring document or term similarities
- Classifying documents
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 qualitative 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 principles 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

![Graph showing posterior probability of document classification]

- Predicted Petitioner
- Predicted Respondent

Log wordscores mean for document

Posterior $P(\text{class}=\text{Petitioner}|\text{document})$
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:

<table>
<thead>
<tr>
<th></th>
<th>time</th>
<th>flies</th>
<th>fruit</th>
<th>like</th>
<th>an</th>
<th>a</th>
<th>banana</th>
<th>arrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sentence 2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- 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 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
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:

[Image of ancient text]

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.

*Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz*
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.
  
  Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

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

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

\[
\begin{align*}
N = 1 & : \textcolor{blue}{\text{This is a sentence}} \quad \text{unigrams:} \quad \text{this, is, a, sentence} \\
N = 2 & : \textcolor{blue}{\text{This is a sentence}} \quad \text{bigrams:} \quad \text{this is, is a, a sentence} \\
N = 3 & : \textcolor{blue}{\text{This is a sentence}} \quad \text{trigrams:} \quad \text{this is a, is a sentence}
\end{align*}
\]
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

<table>
<thead>
<tr>
<th>Number</th>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>CC</td>
<td>Coordinating conjunction</td>
</tr>
<tr>
<td>2.</td>
<td>CD</td>
<td>Cardinal number</td>
</tr>
<tr>
<td>3.</td>
<td>DT</td>
<td>Determiner</td>
</tr>
<tr>
<td>4.</td>
<td>EX</td>
<td>Existential <em>there</em></td>
</tr>
<tr>
<td>5.</td>
<td>FW</td>
<td>Foreign word</td>
</tr>
<tr>
<td>6.</td>
<td>IN</td>
<td>Preposition or subordinating conjunction</td>
</tr>
<tr>
<td>7.</td>
<td>JJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>8.</td>
<td>JJR</td>
<td>Adjective, comparative</td>
</tr>
<tr>
<td>9.</td>
<td>JJS</td>
<td>Adjective, superlative</td>
</tr>
<tr>
<td>10.</td>
<td>LS</td>
<td>List item marker</td>
</tr>
<tr>
<td>11.</td>
<td>MD</td>
<td>Modal</td>
</tr>
<tr>
<td>12.</td>
<td>NN</td>
<td>Noun, singular or mass</td>
</tr>
<tr>
<td>13.</td>
<td>NNS</td>
<td>Noun, plural</td>
</tr>
<tr>
<td>14.</td>
<td>NNP</td>
<td>Proper noun, singular</td>
</tr>
<tr>
<td>15.</td>
<td>NNPS</td>
<td>Proper noun, plural</td>
</tr>
<tr>
<td>16.</td>
<td>PDT</td>
<td>Predeterminer</td>
</tr>
<tr>
<td>17.</td>
<td>POS</td>
<td>Possessive ending</td>
</tr>
<tr>
<td>18.</td>
<td>PRP</td>
<td>Personal pronoun</td>
</tr>
<tr>
<td>19.</td>
<td>PRPS</td>
<td>Possessive pronoun</td>
</tr>
<tr>
<td>20.</td>
<td>RB</td>
<td>Adverb</td>
</tr>
</tbody>
</table>
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 current current 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  ORG_I
## 10  Quantitative quantitative PROPN ORG_I
## 11  Methods  methods PROPN ORG_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
## 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

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; though; thoroughly; those; 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
debates18 includes speeches made in the 2018 in the House of Commons

```r
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 total of $2^7 = 128$ possible feature matrices
  - These selection decisions can have substantive implications for the inferences we draw from QTA
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!

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

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

```r
##
## [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
```

The idea of a local “context” is central to more advanced QTA analyses such as word-embeddings.
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 introduction 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 (Spirling, 2015)

- 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)
Readability Example (Spirling, 2015)

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 (Benoit, Spirling, and Munger (2019))

Are these simple measures really sufficient? What might be missing?

1. **Other features of complexity/readability** (word rarity; Syntactic and grammatical structure)
   - Use relative frequency of terms compared to “the” in google books (dynamic over time)
   - Use number of clauses; proportion of nouns/verbs/adjectives/adverbs

2. **In-domain validation** (are the predictors of “complexity” the same in politics and education?)
   - Crowdsourcsource comparison task of pairs of political sentences (SOTU addresses)

3. **Uncertainty estimates** (is a text with FRE = 50 really more readable than one with FRE = 55?)
   - Bradley-Terry model for paired comparisons to provide probabilistic statements of relative complexity
Findings:

1. **Most important predictors** are sentence length, the proportion of nouns, word rarity, word length
   - Sound familiar?
2. **Modest improvement** over FRE score (3 percentage point improvement over 70% baseline)
3. **Very high correlation** with basic Flesch measure
Figure 2  Probability That a State of the Union Address Is Easier to Understand Than a Fifth Grade Text Baseline, Compared to FRE
Thankfully, quantedA makes it trivial to calculate many of these statistics...

```r
# 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. QDA Miner
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
- “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:
  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 occurrences 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 \((m = 1, \ldots, M)\) that is related to a common concept.

<table>
<thead>
<tr>
<th>Aggression</th>
</tr>
</thead>
<tbody>
<tr>
<td>stupid</td>
</tr>
<tr>
<td>dishonest</td>
</tr>
<tr>
<td>liar</td>
</tr>
<tr>
<td>idiot</td>
</tr>
<tr>
<td>ignorant</td>
</tr>
<tr>
<td>hate</td>
</tr>
<tr>
<td>fight</td>
</tr>
<tr>
<td>battle</td>
</tr>
</tbody>
</table>
Counting words

Applying a dictionary to a corpus of texts \((i = 1, \ldots, N)\) 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.
“That statement is as barbaric as it is downright stupid; it is nothing more than an ignorant, cruel and deliberate misconception to hide behind.”

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

<table>
<thead>
<tr>
<th>Aggression</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>stupid</td>
<td>.6</td>
</tr>
<tr>
<td>dishonest</td>
<td>.2</td>
</tr>
<tr>
<td>lie</td>
<td>.5</td>
</tr>
<tr>
<td>idiot</td>
<td>.7</td>
</tr>
<tr>
<td>ignorant</td>
<td>.3</td>
</tr>
<tr>
<td>brutal</td>
<td>.4</td>
</tr>
<tr>
<td>violence</td>
<td>.5</td>
</tr>
</tbody>
</table>

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 \textit{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.
“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

Linguistic 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)*

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

*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.*
**Advantages of dictionaries: Multi-lingual**

### APPENDIX B

**DICTIONARY OF THE COMPUTER-BASED CONTENT ANALYSIS**

<table>
<thead>
<tr>
<th>NL</th>
<th>UK</th>
<th>GE</th>
<th>IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>elit*</td>
<td>elit*</td>
<td>elit*</td>
<td>elit*</td>
</tr>
<tr>
<td>consensus*</td>
<td>consensus*</td>
<td>konsens*</td>
<td>consens*</td>
</tr>
<tr>
<td>ondemocratisch*</td>
<td>undemocratic*</td>
<td>ondemokratisch*</td>
<td>ondemokratisch*</td>
</tr>
<tr>
<td>referend*</td>
<td>referend*</td>
<td>referend*</td>
<td>referend*</td>
</tr>
<tr>
<td>corrupt*</td>
<td>corrupt*</td>
<td>korrupt*</td>
<td>corrot*</td>
</tr>
<tr>
<td>propagand*</td>
<td>propagand*</td>
<td>propagand*</td>
<td>propagand*</td>
</tr>
<tr>
<td>politici*</td>
<td>politici*</td>
<td>politiker*</td>
<td>politici*</td>
</tr>
<tr>
<td><em>bedrog</em></td>
<td><em>deceit</em></td>
<td>betrüg*</td>
<td>betrug*</td>
</tr>
<tr>
<td><em>bedrieg</em></td>
<td><em>deceiv</em></td>
<td>betrüg*</td>
<td>betrug*</td>
</tr>
<tr>
<td><em>verraa</em></td>
<td><em>betray</em></td>
<td><em>verrat</em></td>
<td>tradi*</td>
</tr>
<tr>
<td><em>verrad</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>schaam*</td>
<td>shame*</td>
<td>scham*</td>
<td>vergogn*</td>
</tr>
<tr>
<td>schand*</td>
<td>scandal*</td>
<td>skandal*</td>
<td>scandal*</td>
</tr>
<tr>
<td>waarheid*</td>
<td>truth*</td>
<td>wahrheit*</td>
<td>verità</td>
</tr>
<tr>
<td>oneerlijk*</td>
<td>dishonest*</td>
<td>unfair*</td>
<td>disonest*</td>
</tr>
<tr>
<td>Context</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>establishm*</td>
<td>establishm*</td>
<td><em>herrs</em></td>
<td>partitocrazia</td>
</tr>
<tr>
<td>heersend*</td>
<td>ruling*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>capitul*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kapitul*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kaste*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>leugen*</td>
<td></td>
<td>lüge*</td>
<td>menzogn*</td>
</tr>
<tr>
<td>lieg*</td>
<td></td>
<td></td>
<td>mentir*</td>
</tr>
</tbody>
</table>

_(from Rooduijn and Pauwels 2011)_
Advantages of dictionaries: Fast and easy to apply

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

```r
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 be 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

Brexit mentions per day

Daily Brexit Mentions

EU Ref
GE2017
Disadvantages 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
Disadvantages 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
Validation
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 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:

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

And the words to a dictionary object:

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

Finally, we “apply” the dictionary to the corpus using the dfm function:

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

print(aggression_dfm)
```

```r
## 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

'Aggression' counts

Number of aggressive words in text

0 2000 4000 6000 8000

0 1 2 3 4 5 6 8

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

```r
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).
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:

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

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.”
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.”

1. 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.”
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.”

1. 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 7 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:
  - $Y_i = \alpha + \beta \text{Diff}_i$
  - $\text{Cor}(Y_i, \text{Diff}_i)$
- Repeat for each dictionary
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