Today

1. Project
   - You should all have received feedback on your one-paragraph proposal
   - Next key deadline: October 7th, two-page summary
   - Peer feedback will be due one week later

2. Introduction to text analysis

3. Solutions to challenge 3

4. APIs
Quantitative Text Analysis
Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data
Text as data
Text as data
Text as data
Text as data
Why quantitative text analysis?

Justin Grimmer’s haystack metaphor: QTA improves reading

▶ Analyzing a straw of hay: understanding the meaning of a sentence
  ▶ Humans are great! But computer struggle
▶ Organizing the haystack: describing, classifying, scaling texts
  ▶ Humans struggle. But computers are great!
  ▶ (What this course is about)

Principles of quantitative text analysis (Grimmer & Stewart, 2013)

1. All quantitative models are wrong – but some are useful
2. Quantitative methods for text amplify resources and augment humans
3. There is no globally best method for automated text analysis
4. Validate, validate, validate
Quantitative text analysis requires assumptions

1. Texts represent an observable implication of some underlying characteristic of interest
   - An attribute of the author
   - A sentiment or emotion
   - Salience of a political issue

2. Texts can be represented through extracting their *features*
   - most common is the *bag of words* assumption
   - many other possible definitions of “features” (e.g. word embeddings)

3. A *document-feature matrix* can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest
Outline

- Motivation
- **Workflow**
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data
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 incentivise. It has the
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 documentary unit of 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
Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data
Overview of text as data methods

- Acquire Documents
  - Existing Corpora
  - Undigitized text
- Preprocess
- Research Objective

- Classification
  - Known Categories
    - Dictionary Methods
      - Supervised Methods
        - Individual Classification
          - Measuring Proportions (ReadMe)
        - Individual Methods
          - Ensembles
    - Supervised Models
      - Naive Bayes
      - Models with covariates (STM)
  - Unknown Categories
    - Fully Automated Clustering
      - Document Level (LDA)
    - Mixed Membership Models
      - Date Level (Dynamic Multitopic Model)
      - Author Level (Expressed Agenda Model)

- Ideological Scaling
  - Supervised (wordscores)
  - Unsupervised (wordfish)

Fig. 1 in Grimmer and Stewart (2013)
Figure 2: The probability that a State of the Union address is easier to understand than a fifth grade text baseline.

Benoit, Munger & Spirling (2017)
Document classification into known categories

- Graph 1: # Economy Articles
- Graph 2: Net Tone (Monthly Average)
- Graph 3: Pearson's r = -.22 (p = .16)
Ideological scaling (Lauderdale & Herzog, PA 2016)
Document classification into unknown categories


- Data: General Social Survey (2008) in Germany
- Responses to questions: *Would you please tell me what you associate with the term “left”? and would you please tell me what you associate with the term “right”?*
- Open-ended questions minimize priming and potential interviewer effects
- Automated text analysis to discover unknown categories and classify responses
### Table 1: Top scoring words associated with each topic, and English translations

<table>
<thead>
<tr>
<th>Left topic 1: <strong>Parties</strong> (proportion = .26, average lr-scale value = 5.38)</th>
</tr>
</thead>
<tbody>
<tr>
<td>linke, spd, partei, linken, pds, politik, kommunisten, parteien, grünen, punks</td>
</tr>
<tr>
<td><em>the left, spd, party, the left, pds, politics, communists, parties, greens, punks</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Left topic 2: <strong>Ideologies</strong> (proportion = .26, average lr-scale value = 5.36)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kommunismus, links, sozialismus, lafontaine, rechts, aber, gysi, linkspartei, richtung, gleichmacherei</td>
</tr>
<tr>
<td><em>communism, left, socialism, lafontaine, right, but, gysi, left party, direction, levelling</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Left topic 3: <strong>Values</strong> (proportion = .24, average lr-scale value = 4.06)</th>
</tr>
</thead>
<tbody>
<tr>
<td>soziale, gerechtigkeit, demokratie, soziales, bürger, gleichheit, gleiche, freiheit, rechte, gleichberechtigung</td>
</tr>
<tr>
<td><em>social, justice, democracy, social, citizen, equality, equal, freedom, rights, equal rights</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Left topic 4: <strong>Policies</strong> (proportion = .24, average lr-scale value = 4.89)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sozial, menschen, leute, ddr, verbinde, kleinen, einstellung, umverteilung, sozialen, vertreten</td>
</tr>
<tr>
<td><em>social, humans, people, ddr, associate, the little, attitude, redistribution, social, represent</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Right topic 1: <strong>Ideologies</strong> (proportion = .27, average lr-scale value = 5.00)</th>
</tr>
</thead>
<tbody>
<tr>
<td>konservativ, nationalsozialismus, rechtsradikal, radikal, ordnung, politik, nazi, recht, menschen, konservative</td>
</tr>
<tr>
<td><em>conservative, national socialism, right-wing radicalism, radical, order, politics, nazi, right, people, conservatives</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Right topic 2: <strong>Parties</strong> (proportion = .25, average lr-scale value = 5.26)</th>
</tr>
</thead>
<tbody>
<tr>
<td>npd, rechts, cdu, csu, rechten, parteien, leute, aber, verbinde, rechtsradikal</td>
</tr>
<tr>
<td><em>npd, right, cdu, csu, the right, parties, people, but, associate, right-wing radicals</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Right topic 3: <strong>Xenophobia</strong> (proportion = .25, average lr-scale value = 4.55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ausländerfeindlichkeit, gewalt, ausländer, demokratie, nationalismus, rechtsradikalismus, diktatur, national, intoleranz, faschismus</td>
</tr>
<tr>
<td><em>xenophobia, violence, foreigners, democracy, nationalism, right-wing radicalism, dictatorship, national, intolerance, fascism</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Right topic 4: <strong>Right-wing extremists</strong> (proportion = .23, average lr-scale value = 4.90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nazis, neonazis, rechtsradikale, rechte, radikale, radikalismus, partei, ausländerfeindlich, reich, nationale</td>
</tr>
<tr>
<td><em>nazis, neonazis, right-wing radicals, rightists, radicals, radicalism, party, xenophobia, rich, national</em></td>
</tr>
</tbody>
</table>

**Note:** “proportion” indicates the average estimated probability that any given response is assigned to a topic. “average lr-scale value” is the mean position on the left-right scale (from 0 to 10) of individuals whose highest probability belongs to that particular topic.
Document classification into unknown categories

Fig. 6: Left-right scale means for different subsamples of associations with left (dashed = sample mean, bars = 95% CIs)

Fig. 7: Left-right scale means for different subsamples of associations with right (dashed = sample mean, bars = 95% CIs)

Fig. 9: Systematic relationship between associations with “left” and “right” and characteristics of respondents

Note: Each line indicates a 95% confidence interval (and 66% confidence interval in darker color) for the coefficient of eight different regressions of topic usage (in a scale from 0 to 100) at the respondent level on seven individual-level characteristics. The line on the bottom right corner (second row, second plot), for example, shows that individual a one-category change in age is associated with around one percentage point increase in the probability that the individual associated “right” with political parties.

What political issues do U.S. legislators emphasize on Twitter?

- Data: 651,116 tweets sent by US legislators from January 2013 to December 2014.
- Unit of analysis: tweets aggregated by day, party, and chamber
- 2,920 documents = 730 days $\times$ 2 chambers $\times$ 2 parties
- Automated text analysis to discover unknown categories and classify responses
- Validation: [http://j.mp/lda-congress-demo](http://j.mp/lda-congress-demo)
Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data
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 incentivise. It has the...
Some key basic concepts

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

document  each of the units of the corpus

types  for our purposes, a unique word
	
tokens  any word – so token count is total words

e.g.  A corpus is a set of documents. This is the second document in the corpus. is a corpus with 2 documents, where each document is a sentence. The first document has 6 types and 7 tokens. The second has 7 types and 8 tokens. (We ignore punctuation for now.)
Some more key basic concepts

**stems**  words with suffixes removed (using set of rules)

**lemmas**  canonical word form (the base form of a word that has the same meaning even when different suffixes or prefixes are attached)

<table>
<thead>
<tr>
<th>word</th>
<th>win</th>
<th>winning</th>
<th>wins</th>
<th>won</th>
<th>winner</th>
</tr>
</thead>
<tbody>
<tr>
<td>stem</td>
<td>win</td>
<td>win</td>
<td>win</td>
<td>won</td>
<td>winner</td>
</tr>
<tr>
<td>lemma</td>
<td>win</td>
<td>win</td>
<td>win</td>
<td>win</td>
<td>win</td>
</tr>
</tbody>
</table>

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

**“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
Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data
Strategies for selecting units of textual analysis

What can the document be?

- Words
- $n$-word sequences
- Sentences
- Pages
- Paragraphs
- Natural units (a speech, a poem, a manifesto)
- Aggregation of units (e.g. all speeches by party and year)
- Key: depends on the research design
- Frequent trade-off between cost and accuracy
Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data
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 incentivise. It has the...
Bag-of-words approach

From words to numbers:

1. **Preprocess text**: lowercase, remove stopwords and punctuation, stem, tokenize into unigrams and bigrams (bag-of-words assumption)

   “A corpus is a set of documents.”
   “This is the second document in the corpus.” “a corpus is a set of documents.”
   “this is the second document in the corpus.” “a corpus is a set of documents.”
   “this is the second document in the corpus.” “corpus set documents”
   “second document corpus” [corpus, set, document, corpus set, set document]
   [second, document, corpus, second document, document corpus]

2. **Document-feature matrix**:
   - \( W \): matrix of \( N \) documents by \( M \) unique n-grams
   - \( w_{im} \): number of times \( m \)-th n-gram appears in \( i \)-th document.

\[
\begin{array}{cccccc}
\text{corpus} & \text{set} & \text{document} & \text{corpus set} & \text{set document} & \text{M n-grams} \\
\text{Document 1} & 1 & 1 & 1 & 1 & . . . \\
\text{Document 2} & 1 & 0 & 1 & 0 & . . . \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
\text{Document n} & 0 & 1 & 1 & 0 & . . . \\
\end{array}
\]
Bag-of-words approach disregards grammar and word order and uses word frequencies as features. Why?

- Context is often uninformative, conditional on presence of words:
  - Individual word usage tends to be associated with a particular degree of affect, position, etc. without regard to context of word usage

- Single words tend to be the most informative, as co-occurrences of multiple words \((n\text{-grams})\) are rare

- Some approaches focus on occurrence of a word as a binary variable, irrespective of frequency: a binary outcome

- Other approaches use frequencies: Poisson, multinominal, and related distributions
Defining Features

- characters
- 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)

  _Saunauntensitzer_
Defining Features (cont.)

▶ “word” sequences, especially when inter-word delimiters (usually white space) are not commonly used, as in

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。

Chinese

▶ linguistic features, such as parts of speech
▶ (if qualitative coding is used) coded or annotated text segments
▶ word embeddings (more on this later in the course)
 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>PRP$</td>
<td>Possessive pronoun</td>
</tr>
<tr>
<td>20.</td>
<td>RB</td>
<td>Adverb</td>
</tr>
<tr>
<td>21.</td>
<td>RBR</td>
<td>Adverb, comparative</td>
</tr>
<tr>
<td>22.</td>
<td>RBS</td>
<td>Adverb, superlative</td>
</tr>
<tr>
<td>23.</td>
<td>RP</td>
<td>Particle</td>
</tr>
</tbody>
</table>
several open-source projects make it possible to tag POS in text, such as Apache’s OpenNLP (and R package openNLP wrapper) or TreeTagger.

> s
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29. 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"  "as/IN"  
[9] "join/VB"  "the/DTD"  "board/NN"  "Nov./NNP"  
[13] "a/DT"  "nonexecutive/JJ"  "director/NN"  "Vinken/NNP"  
[17] "29/CD"  
[21] "is/VBZ"  "chairman/NN"  "of/IN"  "Elsevier/NNP"  
[25] "N.V./NNP"  
[29] "publishing/NN"  "group/NN"  "the/DTD"  "Dutch/JJ"  
"/"
Parts of speech (cont.)

Example: Creating an index of editorialization of journalists’ and media outlets’ political news coverage.

Proportion of tweets that: (1) mention a major party or candidate, (2) include at least one adjective.

Table 2.4 Determinants of editorialisation and popularity of news accounts on twitter (OLS regressions)

<table>
<thead>
<tr>
<th></th>
<th>DV = Editorialisation</th>
<th>DV = Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
</tr>
<tr>
<td>Type: journalist</td>
<td>5.10***</td>
<td>4.32***</td>
</tr>
<tr>
<td></td>
<td>(1.13)</td>
<td>(1.26)</td>
</tr>
<tr>
<td>Tweets about Europe (%)</td>
<td>−0.03*</td>
<td>−0.03*</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Editorialisation Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>7.58**</td>
<td>7.94**</td>
</tr>
<tr>
<td></td>
<td>(2.59)</td>
<td>(2.47)</td>
</tr>
<tr>
<td>Country fixed effects</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Outlet fixed effects</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>R²</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>Num. obs.</td>
<td>2662</td>
<td>2662</td>
</tr>
<tr>
<td>RMSE</td>
<td>7.63</td>
<td>7.63</td>
</tr>
</tbody>
</table>

Barberá, Vaccari, Valeriani (2016) [control variables ommitted]
Strategies for feature selection

How to choose which features to include?

▶ **All?** Computationally inefficient, and rare words are generally uninformative

Potential criteria to select features (“trim” the DFM):

▶ **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

▶ **declared equivalency classes** Non-exclusive synonyms, also known as *thesaurus* (more on this later)
Common English stop words

a, able, about, across, after, all, almost, also, am, 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, 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, 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, would, yet, you, your

- But no list should be considered universal
A more comprehensive list of stop words

as, 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, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, 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, definitely, described, despite, did, didn’t, different, do, does, doesn’t, doing, don’t, done, down, downwards, during, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, every, everybody, everyone, everywhere, ex, exactly, example, except, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, 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’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, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, let’s, like, liked, likely, little, look, looking, looks, ltd, mainly, many, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, much, my, myself, 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, 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, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, que, quite, qv, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, 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’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, there's, 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, un, under, unfortunately, unless, unlikely, until, unto, up, upon, us, use, used, useful, uses, using, usually, value, various, very, via, viz, vs, 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, whereabouts, 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, yes, yet, you, you’d, you’ll, you’re, you’ve, your, yours, yourself, yourselves, zero
Are there cases in which we would want to keep stopwords? Or should we always exclude them from our analysis?
Stopwords sometimes can be informative!

But sometimes we want to add/remove our own new stopwords (e.g. female pronouns, legislative terms, directional terms)
Stemming words

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

**stemming** is 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
- production, producer, produce, produces, produced

**Why?** Reduce feature space by collapsing different words into a stem (e.g. “happier” and “happily” convey same meaning as “happy”)

Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data
Where to obtain textual data?

Some tips...

▶ Existing datasets, e.g.
  ▶ UCD’s EuroParl project
  ▶ Hansard Archive of parliamentary debates in UK
  ▶ Media archives (newspaper articles, TV transcripts...) at LexisNexis, ProQuest, Factiva...
  ▶ Academic articles (JSTOR Data for Research)
  ▶ Open-ended responses to survey questions

▶ Collect your own data:
  ▶ From social media (Twitter, FB) and blogs
  ▶ Scraping other websites

▶ Digitize your own text data using OCR (optical character recognition) software
  ▶ Options: Tesseract (open-source), Abbyy FineReader
Problems you are likely to encounter

- Problems with encoding
- File formats that cannot be read as plain text
- Extraneous junk (page footers, numbers, titles, etc)
- Misspellings
- Different normalizations (e.g. for Japanese)