Quantitative text analysis: descriptive statistical methods

Blake Miller

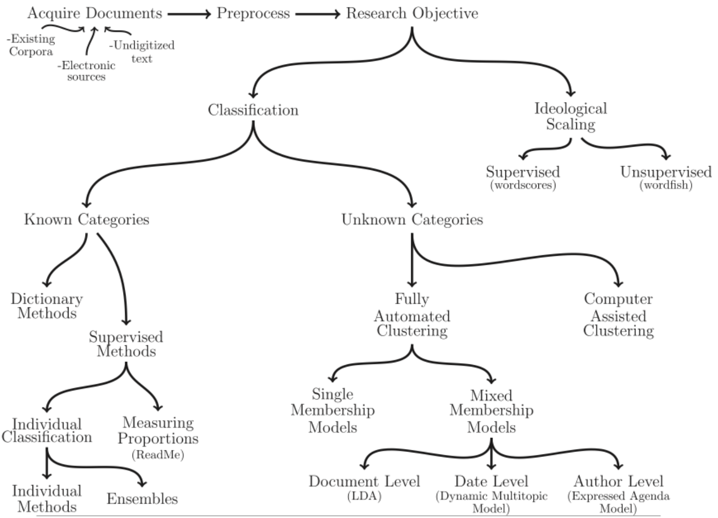
MY 459: Quantitative Text Analysis

January 14, 2019

Course website: lse-my459.github.io

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

Overview of text as data methods



Outline for today

- Where to obtain data
- Defining documents
- Defining features
- Strategies for feature selection
- Defining feature weights
- Descriptive statistics for text

Outline for today

- Where to obtain data
- Defining documents
- Defining features
- Strategies for feature selection
- Defining feature weights
- Descriptive statistics for text

Where to obtain textual data?

Some tips...

- Existing datasets, e.g.

Where to obtain textual data?

Some tips...

- Existing datasets, e.g.
  - UCD's EuroParl project

Where to obtain textual data?

Some tips...

- Existing datasets, e.g.
  - UCD's EuroParl project
  - Hansard Archive of parliamentary debates in UK

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...

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)

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

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:

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, reddit) and blogs

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, reddit) and blogs
  - Scraping other websites

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, reddit) and blogs
  - Scraping other websites
- Digitize your own text data using OCR (optical character recognition) software

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, reddit) 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

Problems you are likely to encounter

- Problems with encoding
- File formats that cannot be read as plain text

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)

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)
- Misspelllings

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)
- Misspelllings
- Different normalizations (e.g. for Japanese)

Outline for today

- Where to obtain data
- Defining documents
- Defining features
- Strategies for feature selection
- Defining feature weights
- Descriptive statistics for text

Strategies for selecting units of textual analysis

What can the document be?

- Words

Strategies for selecting units of textual analysis

What can the document be?

- Words
- *n*-word sequences

Strategies for selecting units of textual analysis

What can the document be?

- Words
- *n*-word sequences
- Sentences

Strategies for selecting units of textual analysis

What can the document be?

- Words
- *n*-word sequences
- Sentences
- Pages

Strategies for selecting units of textual analysis

What can the document be?

- Words
- *n*-word sequences
- Sentences
- Pages
- Paragraphs

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)

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)

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

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

Sampling strategies for selecting texts

- Difference between a sample and a population

Sampling strategies for selecting texts

- Difference between a sample and a population
- *May not be feasible* to perform any sampling

Sampling strategies for selecting texts

- Difference between a sample and a population
- *May not be feasible* to perform any sampling
- *May not be necessary* to perform any sampling

Sampling strategies for selecting texts

- Difference between a sample and a population
- *May not be feasible* to perform any sampling
- *May not be necessary* to perform any sampling
- Be wary of sampling that is a feature of the social system: "social bookkeeping"

Sampling strategies for selecting texts

- Difference between a sample and a population
- *May not be feasible* to perform any sampling
- *May not be necessary* to perform any sampling
- Be wary of sampling that is a feature of the social system: "social bookkeeping"
- Different types of sampling vary from random to purposive

Sampling strategies for selecting texts

- Difference between a sample and a population
- *May not be feasible* to perform any sampling
- *May not be necessary* to perform any sampling
- Be wary of sampling that is a feature of the social system: "social bookkeeping"
- Different types of sampling vary from random to purposive
  - random sampling

Sampling strategies for selecting texts

- Difference between a sample and a population
- *May not be feasible* to perform any sampling
- *May not be necessary* to perform any sampling
- Be wary of sampling that is a feature of the social system: "social bookkeeping"
- Different types of sampling vary from random to purposive
  - random sampling
  - non-random sampling

Sampling strategies for selecting texts

- Difference between a sample and a population
- *May not be feasible* to perform any sampling
- *May not be necessary* to perform any sampling
- Be wary of sampling that is a feature of the social system: "social bookkeeping"
- Different types of sampling vary from random to purposive
  - random sampling
  - non-random sampling
- Key is to make sure that what is being analyzed is a valid representation of the phenomenon as a whole – a question of research design

Outline for today

- Where to obtain data
- Defining documents
- Defining features
- Strategies for feature selection
- Defining feature weights
- Descriptive statistics for text

Defining Features

- characters

Defining Features

- characters
- words

Defining Features

- characters
- words
- word stems or lemmas: this is a form of defining *equivalence classes* for word features

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.

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.

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*

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)

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 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。

Defining Features (cont.)

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

  莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

- (if qualitative coding is used) coded or annotated text segments

Defining Features (cont.)

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

  莎拉波娃现在居住在美国东南部的佛罗里达。今年４月
  ９日，莎拉波娃在美国第一大城市纽约度过了１８岁生
  日。生日派对上，莎拉波娃露出了甜美的微笑。

- (if qualitative coding is used) coded or annotated text segments
- word embeddings (more on this later in the course)

Defining Features (cont.)

- "word" sequences, especially when inter-word delimiters (usually white space) are not commonly used, as in Chinese
  莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。
- (if qualitative coding is used) coded or annotated text segments
- word embeddings (more on this later in the course)
- linguistic features, such as parts of speech

Parts of speech

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

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol

Parts of speech (cont.)

- 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"
 [9] "join/VB"         "the/DT"           "board/NN"         "as/IN"
[13] "a/DT"            "nonexecutive/JJ" "director/NN"       "Nov./NNP"
[17] "29/CD"           "./."              "Mr./NNP"          "Vinken/NNP"
[21] "is/VBZ"          "chairman/NN"      "of/IN"            "Elsevier/NNP"
[25] "N.V./NNP"        ",/,"              "the/DT"           "Dutch/JJ"
[29] "publishing/NN"   "group/NN"         "./."
```

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)

	DV = Editorialisation		DV = Popularity	
	Model 1	Model 2	Model 3	Model 4
Type: journalist	5.10***	4.32***	2.70***	2.49***
	(1.13)	(1.26)	(0.22)	(0.30)
Tweets about Europe (%)	–0.03+	–0.03+	0.01***	0.01***
	(0.02)	(0.02)	(0.002)	(0.002)
Editorialisation Index			0.02***	0.02***
			(0.004)	(0.004)
(Intercept)	7.58**	7.94**	–4.03***	–3.92***
	(2.59)	(2.47)	(0.40)	(0.41)
Country fixed effects	YES	YES	YES	YES
Outlet fixed effects	YES	YES	YES	YES
R^2	0.12	0.12	0.71	0.71
Adj. R^2	0.08	0.08	0.70	0.70
Num. obs.	2662	2662	2662	2662
RMSE	7.63	7.63	1.08	1.08

Barberá, Vaccari, Valeriani (2016) [control variables ommitted]

Outline for today

- Where to obtain data
- Defining documents
- Defining features
- Strategies for feature selection
- Defining feature weights
- Descriptive statistics for text

Strategies for feature selection

How to choose which features to include?

- All? Computationally inefficient, and rare words are generally uninformative

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):

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

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

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

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

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

Common English stop words

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

- But no list should be considered universal

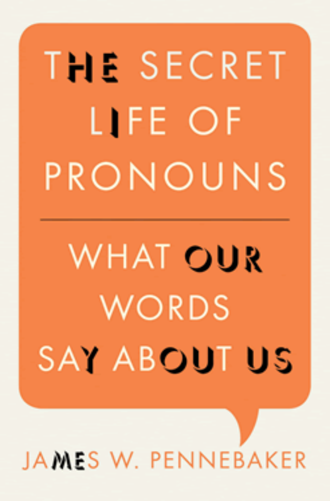
A more comprehensive list of stop words

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, ever, every, everybody, everyone, everything, 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, it's, 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, must, 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, 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, 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, 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, yes, yet, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves, zero

Stopwords

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 words

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

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.

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

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`

Stemming words

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

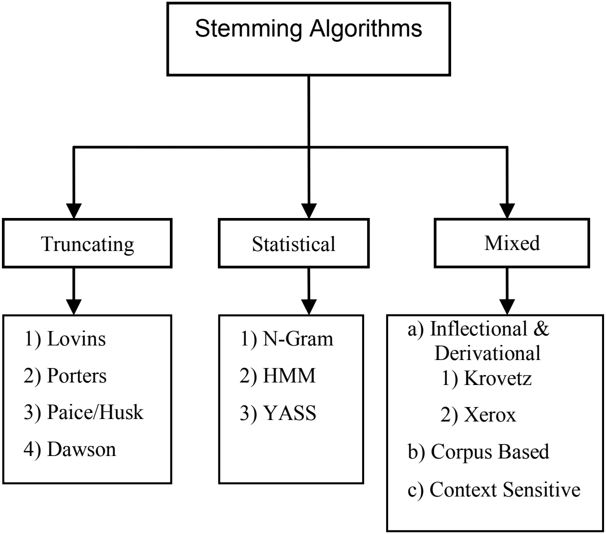
stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

both convert the morphological variants into stem or root terms

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

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

Varieties of stemming algorithms



Issues with stemming approaches

- The most common is probably the Porter stemmer

Issues with stemming approaches

- The most common is probably the Porter stemmer
- But this set of rules gets many stems wrong, e.g.

Issues with stemming approaches

- The most common is probably the Porter stemmer
- But this set of rules gets many stems wrong, e.g.
  - `policy` and `police` considered (wrongly) equivalent

Issues with stemming approaches

- The most common is probably the Porter stemmer
- But this set of rules gets many stems wrong, e.g.
  - `policy` and `police` considered (wrongly) equivalent
  - `general` becomes `gener`, `iteration` becomes `iter`

Issues with stemming approaches

- The most common is probably the Porter stemmer
- But this set of rules gets many stems wrong, e.g.
  - `policy` and `police` considered (wrongly) equivalent
  - `general` becomes `gener`, `iteration` becomes `iter`
- Other corpus-based, statistical, and mixed approaches designed to overcome these limitations

Issues with stemming approaches

- The most common is probably the Porter stemmer
- But this set of rules gets many stems wrong, e.g.
  - `policy` and `police` considered (wrongly) equivalent
  - `general` becomes `gener`, `iteration` becomes `iter`
- Other corpus-based, statistical, and mixed approaches designed to overcome these limitations
- Key for you is to be careful through inspection of morphological variants and their stemmed versions

Issues with stemming approaches

- The most common is probably the Porter stemmer
- But this set of rules gets many stems wrong, e.g.
  - `policy` and `police` considered (wrongly) equivalent
  - `general` becomes `gener`, `iteration` becomes `iter`
- Other corpus-based, statistical, and mixed approaches designed to overcome these limitations
- Key for you is to be careful through inspection of morphological variants and their stemmed versions
- Sometimes not appropriate! e.g. Schofield and Minmo (2016) find that "stemmers produce no meaningful improvement in likelihood and coherence (of topic models) and in fact can degrade topic stability"

Selecting more than words: collocations

collocations bigrams, or trigrams e.g. *capital gains tax*

how to detect: pairs occuring more than by chance, by measures of χ^2 or *mutual information* measures

example:

Summary Judgment	Silver Rudolph	Sheila Foster
prima facie	COLLECTED WORKS	Strict Scrutiny
Jim Crow	waiting lists	Trail Transp
stare decisis	Academic Freedom	Van Alstyne
Church Missouri	General Bldg	Writings Fehrenbacher
Gerhard Casper	Goodwin Liu	boot camp
Juan Williams	Kurland Gerhard	dated April
LANDMARK BRIEFS	Lee Appearance	extracurricular activities
Lutheran Church	Missouri Synod	financial aid
Narrowly Tailored	Planned Parenthood	scored sections

Table 5: Bigrams detected using the mutual information measure.

Identifying collocations

- Does a given word occur next to another given word with a higher relative frequency than other words?

Identifying collocations

- Does a given word occur next to another given word with a higher relative frequency than other words?
- If so, then it is a candidate for a collocation

Identifying collocations

- Does a given word occur next to another given word with a higher relative frequency than other words?
- If so, then it is a candidate for a collocation
- We can detect these using measures of association, such as a likelihood ratio, to detect word pairs that occur with greater than chance frequency, compared to an independence model

Identifying collocations

- Does a given word occur next to another given word with a higher relative frequency than other words?
- If so, then it is a candidate for a collocation
- We can detect these using measures of association, such as a likelihood ratio, to detect word pairs that occur with greater than chance frequency, compared to an independence model
- The key is to distinguish "true collocations" from uninteresting word pairs/triplets/etc, such as "of the"

Example

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Table 5.1 Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

(from Manning and Schütze, *FSNLP*, Ch 5)

Example

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Table 5.1 Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

(from Manning and Schütze, *FSNLP*, Ch 5)

Contingency tables for bigrams

Tabulate every token against every other token as pairs, and compute for each token:

	token2	¬token2	Totals
token1	n_{11}	n_{12}	n_{1p}
¬token1	n_{21}	n_{22}	n_{2p}
Totals	n_{p1}	n_{p2}	n_{pp}

Then compute the "independence" model:

$$Pr(\text{token1}, \text{token2}) = Pr(\text{token1})Pr(\text{token2})$$

statistical association measures

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

G^2 likelihood ratio statistic, computed as:

$$2 * \sum_i \sum_j (n_{ij} * log \frac{n_{ij}}{m_{ij}}) \tag{1}$$

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

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

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

Outline for today

- Where to obtain data
- Defining documents
- Defining features
- Strategies for feature selection
- Defining feature weights
- Descriptive statistics for text

Weighting strategies for feature counting

term frequency Some approaches trim very low-frequency words. Rationale: get rid of rare words that expand the feature matrix but matter little to substantive analysis

Weighting strategies for feature counting

term frequency Some approaches trim very low-frequency words. Rationale: get rid of rare words that expand the feature matrix but matter little to substantive analysis

document frequency Could eliminate words appearing in few documents

Weighting strategies for feature counting

term frequency
: Some approaches trim very low-frequency words. Rationale: get rid of rare words that expand the feature matrix but matter little to substantive analysis

document frequency
: Could eliminate words appearing in few documents

inverse document frequency
: Conversely, could weight words more that appear in the most documents

Weighting strategies for feature counting

term frequency
: Some approaches trim very low-frequency words. Rationale: get rid of rare words that expand the feature matrix but matter little to substantive analysis

document frequency
: Could eliminate words appearing in few documents

inverse document frequency
: Conversely, could weight words more that appear in the most documents

tf-idf
: a combination of term frequency and inverse document frequency, common method for feature weighting

Strategies for feature *weighting*: tf-idf

- $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ where $n_{i,j}$ is number of occurences of term t_i in document d_j, k is total number of terms in document d_j
- $idf_i = \ln \frac{|D|}{|\{d_j : t_i \in d_j\}|}$
  where
  - $|D|$ is the total number of documents in the set
  - $|\{d_j : t_i \in d_j\}|$ is the number of documents where the term t_i appears (i.e. $n_{i,j} \neq 0$)
- $tf\text{-}idf_i = tf_{i,j} \cdot idf_i$

Computation of tf-idf: Example

Example: We have 100 political party manifestos, each with 1000 words. The first document contains 16 instances of the word "environment"; 40 of the manifestos contain the word "environment".

- The *term frequency* is $16/1000 = 0.016$

Computation of tf-idf: Example

Example: We have 100 political party manifestos, each with 1000 words. The first document contains 16 instances of the word "environment"; 40 of the manifestos contain the word "environment".

- The *term frequency* is $16/1000 = 0.016$
- The *document frequency* is $100/40 = 2.5$, or $\ln(2.5) = 0.916$

Computation of tf-idf: Example

Example: We have 100 political party manifestos, each with 1000 words. The first document contains 16 instances of the word "environment"; 40 of the manifestos contain the word "environment".

- The *term frequency* is $16/1000 = 0.016$
- The *document frequency* is $100/40 = 2.5$, or $\ln(2.5) = 0.916$
- The *tf-idf* will then be $0.016 * 0.916 = 0.0147$

Computation of tf-idf: Example

Example: We have 100 political party manifestos, each with 1000 words. The first document contains 16 instances of the word "environment"; 40 of the manifestos contain the word "environment".

- The *term frequency* is $16/1000 = 0.016$
- The *document frequency* is $100/40 = 2.5$, or $\ln(2.5) = 0.916$
- The *tf-idf* will then be $0.016 * 0.916 = 0.0147$

Computation of tf-idf: Example

Example: We have 100 political party manifestos, each with 1000 words. The first document contains 16 instances of the word "environment"; 40 of the manifestos contain the word "environment".

- The *term frequency* is $16/1000 = 0.016$
- The *document frequency* is $100/40 = 2.5$, or $\ln(2.5) = 0.916$
- The *tf-idf* will then be $0.016 * 0.916 = 0.0147$

Computation of tf-idf: Example

Example: We have 100 political party manifestos, each with 1000 words. The first document contains 16 instances of the word "environment"; 40 of the manifestos contain the word "environment".

- The *term frequency* is $16/1000 = 0.016$
- The *document frequency* is $100/40 = 2.5$, or $\ln(2.5) = 0.916$
- The *tf-idf* will then be $0.016 * 0.916 = 0.0147$
- If the word had only appeared in 15 of the 100 manifestos, then the *tf-idf* would be 0.0304 (three times higher).

Computation of tf-idf: Example

Example: We have 100 political party manifestos, each with 1000 words. The first document contains 16 instances of the word “environment”; 40 of the manifestos contain the word “environment”.

- The *term frequency* is $16/1000 = 0.016$
- The *document frequency* is $100/40 = 2.5$, or $\ln(2.5) = 0.916$
- The *tf-idf* will then be $0.016 * 0.916 = 0.0147$
- If the word had only appeared in 15 of the 100 manifestos, then the *tf-idf* would be 0.0304 (three times higher).

Computation of tf-idf: Example

Example: We have 100 political party manifestos, each with 1000 words. The first document contains 16 instances of the word "environment"; 40 of the manifestos contain the word "environment".

- The *term frequency* is $16/1000 = 0.016$
- The *document frequency* is $100/40 = 2.5$, or $\ln(2.5) = 0.916$
- The *tf-idf* will then be $0.016 * 0.916 = 0.0147$
- If the word had only appeared in 15 of the 100 manifestos, then the *tf-idf* would be 0.0304 (three times higher).
- A high weight in tf-idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; hence the weights hence tend to filter out common terms

Other weighting schemes

- the SMART weighting scheme (Salton 1991, Salton et al): The first letter in each triplet specifies the term frequency component of the weighting, the second the document frequency component, and the third the form of normalization used (not shown). Example: *lnn* means log-weighted term frequency, no idf, no normalization

Term frequency		Document frequency	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1
l (logarithm)	$1+\log(\mathrm{tf}_{t,d})$	t (idf)	$\log \frac{N}{\mathrm{df}_t}$
a (augmented)	$0.5+\frac{0.5\times\mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0,\log \frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$
b (boolean)	$\begin{cases}1 & \text{if } \mathrm{tf}_{t,d}>0\\ 0 & \text{otherwise}\end{cases}$		
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$		

- Note: Mostly used in information retrieval, although some use in machine learning

Outline for today

- Where to obtain data
- Defining documents
- Defining features
- Strategies for feature selection
- Defining feature weights
- Descriptive statistics for text

Simple descriptive table about texts: Describe your 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

Quantities for describing texts

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

Quantities for describing texts

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

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

Quantities for describing texts

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

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

Quantities for describing texts

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

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 counts or proportions of words

Lexical Diversity

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

Lexical Diversity

- Basic measure is the TTR: Type-to-Token ratio
- Problem: This is very sensitive to overall document length, as shorter texts may exhibit fewer word repetitions

Lexical Diversity

- Basic measure is the TTR: Type-to-Token ratio
- Problem: This is very sensitive to overall document length, as shorter texts may exhibit fewer word repetitions
- Special problem: length may relate to the introduction of additional subjects, which will also increase richness

Lexical Diversity: Alternatives to TTRs

TTR $\frac{\text{total types}}{\text{total tokens}}$

Lexical Diversity: Alternatives to TTRs

TTR $\frac{\text{total types}}{\text{total tokens}}$

Guiraud $\frac{\text{total types}}{\sqrt{\text{total tokens}}}$

Lexical Diversity: Alternatives to TTRs

TTR $\frac{\text{total types}}{\text{total tokens}}$

Guiraud $\frac{\text{total types}}{\sqrt{\text{total tokens}}}$

D (Malvern et al 2004) Randomly sample a fixed number of tokens and count those

Lexical Diversity: Alternatives to TTRs

TTR $\frac{\text{total types}}{\text{total tokens}}$

Guiraud $\frac{\text{total types}}{\sqrt{\text{total tokens}}}$

D (Malvern et al 2004) Randomly sample a fixed number of tokens and count those

MTLD "the mean length of sequential word strings in a text that maintain a given TTR value" (McCarthy and Jarvis, 2010) – fixes the TTR at 0.72 and counts the length of the text required to achieve it

Vocabulary 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

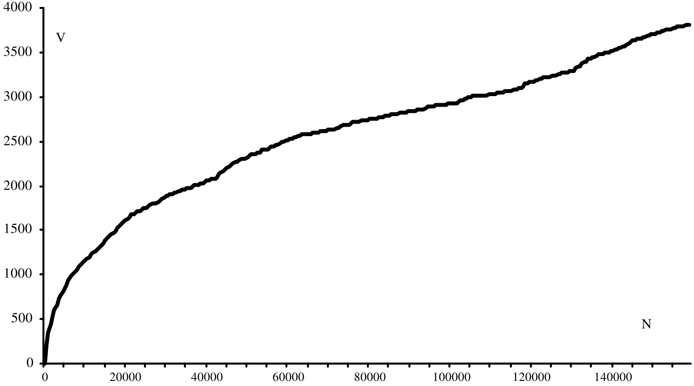


Fig. 1. Chart of vocabulary growth in the tragedies of Racine (chronological order, 500 token intervals).

Complexity and Readability

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

Complexity and Readability

- Use a combination of syllables and sentence length to indicate "readability" in terms of complexity
- Common in educational research, but could also be used to describe textual complexity

Complexity and Readability

- Use a combination of syllables and sentence length to indicate "readability" in terms of complexity
- Common in educational research, but could also be used to describe textual complexity
- Most use some sort of sample

Complexity and Readability

- Use a combination of syllables and sentence length to indicate "readability" in terms of complexity
- Common in educational research, but could also be used to describe textual complexity
- Most use some sort of sample
- No natural scale, so most are calibrated in terms of some interpretable metric

Flesch-Kincaid readability index

- F-K is a modification of the original Flesch Reading Ease Index:

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Interpretation: 0-30: university level; 60-70: understandable by 13-15 year olds; and 90-100 easily understood by an 11-year old student.

Flesch-Kincaid readability index

- F-K is a modification of the original Flesch Reading Ease Index:

  $$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

  Interpretation: 0-30: university level; 60-70: understandable by 13-15 year olds; and 90-100 easily understood by an 11-year old student.

Flesch-Kincaid readability index

- F-K is a modification of the original Flesch Reading Ease Index:

  $$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

  Interpretation: 0-30: university level; 60-70: understandable by 13-15 year olds; and 90-100 easily understood by an 11-year old student.

Flesch-Kincaid readability index

- F-K is a modification of the original Flesch Reading Ease Index:

  $$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

  Interpretation: 0-30: university level; 60-70: understandable by 13-15 year olds; and 90-100 easily understood by an 11-year old student.

- Flesch-Kincaid rescales to the US educational grade levels (1–12):

  $$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

Exploring Texts: Key Words in Context

Exploring Texts: Key Words in Context

KWIC *Key words in context* Refers to the most common format for concordance lines. A KWIC index is formed by sorting and aligning the matching words within a 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

Irish Budget Speeches KIWC in quanteda

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

Wrapping up...

Before this week's seminar:

- Bring a laptop!
- Create a GitHub account
- Install R (from https://www.r-project.org/)
- Install RStudio Desktop (from https://www.rstudio.com/products/rstudio-desktop/)
- Install GitHub Desktop (from https://desktop.github.com/)