MY360/459 Quantitative Text Analysis: Descriptive statistical methods

Patrick Gildersleve

Slide content courtesy of Dr Blake Miller

January 22, 2024

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

Outline for today

- Defining documents and features
- Strategies for feature weighting
- Strategies for feature selection
- Descriptive statistics for text

Overview of text as data methods

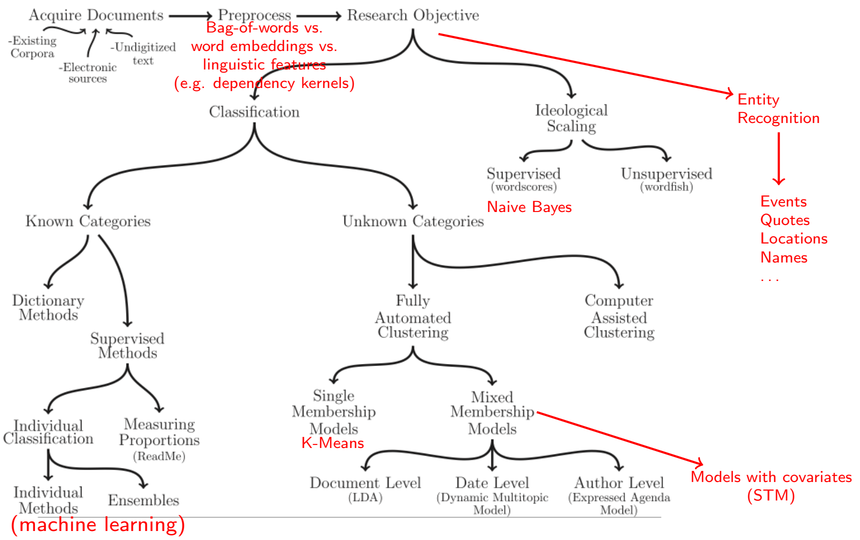


Fig. 1 in Grimmer and Stewart (2013)

Review of 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.)

Review of 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)

word	win	winning	wins	won	winner
stem	win	win	win	won	winner
lemma	win	win	win	win	win

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

Document Term Matrices Review

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- The document term matrix for four words in four Shakespeare plays.
- The red boxes show that each document is represented as a column vector of length four.
- Each document is represented by a vector of words
- Vectors are similar for the two comedies
- Both are different than the two historical plays
- Comedies have more fools and wit and fewer battles.

A potential recipe for preprocessing

1. Remove capitalization, punctuation
2. Segment into words, characters, morphemes
3. Discard Order ("Bag of Words" Assumption)
4. Discard stop words
5. Create Equivalence Class: stem, lemmatize, or synonym
6. Discard less useful features
7. Other reduction, specialization

A Complete Example

"Political power grows out of the barrel of a gun" - Mao

A Complete Example

"Political power grows out of the barrel of a gun" - Mao

Compound Words/Collocations: With substantive justification, words can be combined or split to improve inference.

A Complete Example

"Political power grows out of the barrel of a gun" - Mao

Compound Words/Collocations: An analyst may want to combine words into a single term that can be analyzed.

A Complete Example

"Political power grows out of the barrel of a gun" - Mao

Compound Words/Collocations: An analyst may want to combine words into a single term that can be analyzed.

A Complete Example

[Political], [power], [grows], [out], [of], [the], [barrel of a gun]

Compound Words/Collocations: An analyst may want to combine words into a single term that can be analyzed.

A Complete Example

[Political], [power], [grows], [out], [of], [the], [barrel of a gun]

Stopword Removal: Removing terms that are not related to what the author is studying from the text.

A Complete Example

[Political], [power], [grows], [out], [of], [the], [barrel of a gun]

Stopword Removal: Removing terms that are not related to what the author is studying from the text.

A Complete Example

[Political], [power], [grows], [out], [barrel of a gun]

Stopword Removal: Removing terms that are not related to what the author is studying from the text.

A Complete Example

[Political], [power], [grows], [out], [barrel of a gun]

Stemming: Takes the ends off conjugated verbs or plural nouns, leaving just the "stem."

A Complete Example

[Political], [power], [grows], [out], [barrel of a gun]

Stemming: Takes the ends off conjugated verbs or plural nouns, leaving just the "stem."

A Complete Example

[Polit], [power], [grow], [out], [barrel of a gun]

Stemming: Takes the ends off conjugated verbs or plural nouns, leaving just the "stem."

A Complete Example

Finally, we can turn tokens and documents into a "document-term matrix." Imagine we have a second document in addition to the Mao quote, "the political science students study politics", which tokenizes as follows.

Document #1: [polit], [power], [grow], [out], [barrel of a gun]
Document #2: [polit], [scien], [student], [studi], [polit]

Output: Document Term Matrix

$$
\begin{array}{c}
 \\
\text{power} \\
\text{grow} \\
\text{out} \\
\text{barrel of a gun} \\
\text{student} \\
\text{studi} \\
\text{polit} \\
\text{scien}
\end{array}
\begin{array}{cc}
\text{Doc 1} & \text{Doc 2} \\
\left(\begin{array}{c} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{array}\right. &
\left.\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 2 \\ 1 \end{array}\right)
\end{array}
$$

But raw frequency is a bad representation

- Frequency is clearly useful; if sugar appears a lot near apricot, that's useful information.
- But overly frequent words like "the", "it", or "they" are not very informative about content
- Some terms carry more information about contents
- Need a function that resolves this frequency paradox!

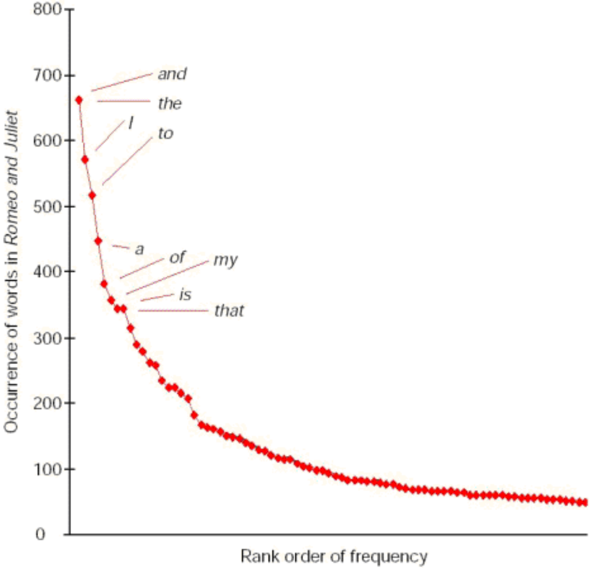
Outline for today

- Defining documents and features
- Strategies for feature weighting
- Strategies for feature selection
- Descriptive statistics for text

Power-law/Zipf Distribution of Word Frequency

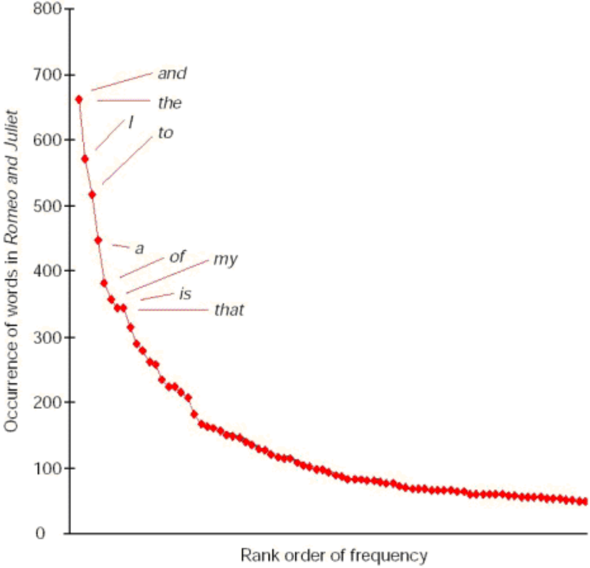
- Many words with a small frequency of occurrence
- A few words with a very large frequency
- High skew (asymmetry)
- Comparing to a normal distribution:
  - Many people with a medium height
  - Almost nobody with a very high or very low height
  - Height is symmetric

Power-law/Zipf Distribution of Word Frequency



Question: Where are the most informative words on this plot?

Power-law/Zipf Distribution of Word Frequency



Question: How might we address the problem of highly weighted non-informative tokens?

Power-law/Zipf Distribution of Word Frequency

1. What do we mean by feature selection?
2. What do we mean by feature weighting?

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

Term Frequency (tf)

- A term is more important if it occurs more frequently in a document
- tf: term frequency. frequency count (usually log-transformed):

$$\mathrm{tf}_{t,d} = \begin{cases} 1 + log(\mathrm{count}(t,d)) & \text{if } \mathrm{count}(t,d) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Inverse Document Frequency (idf)

- A term is more discriminative if it occurs only in fewer documents. Why this is true?

  $idf_i = log\left(\frac{N}{df_i}\right)$
- N is the total number of documents in the collection
- df_i is the number of documents in the collection that contain the word i
- Note that IDF is document independent while TF is document dependent!
- Words like "the" or "and" have a very low idf

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 $16/1000 = 0.016$
- The *inverse 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

Tf-idf Weighted Document Term Matrix

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

- A tf-idf weighted document term matrix for four words in four Shakespeare plays, using the counts from the earlier document term matrix
- The idf weighting has eliminated the importance of the ubiquitous word good and vastly reduced the impact of the almost-ubiquitous word fool.

Other weighting schemes

- Okapi BM25 (based on tf-idf)
- 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

Other weighting schemes: Okapi BM25

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)},$$

- Used in information retrieval (e.g. ranking searches in a search engine given a query q)
- BM25 and other similar term-weighting methods were core components of most early search engines
- $f(q_i, D)$ is the document frequency of the query term
- q_i is the keyword (each word in the search)
- k_1, b are free parameters

Outline for today

- Defining documents and features
- Strategies for feature weighting
- Strategies for feature selection
- Descriptive statistics for text

Review of Feature Selection Approaches

Equivalence classes: stems, or lemmas

Parts of speech: part of speech tags

Purposive selection: dictionary, or keywords

Purposive removal: stopwords

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?
- 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)

Example (filtered)

Some parts of speech indicate a collocation is less interesting.

Frequency	Word 1	Word 2	Part-of-speech pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Table 1.5 Frequent bigrams after filtering. The most frequent bigrams in the *New York Times* after applying a part-of-speech filter.

(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

- Defining documents and features
- Strategies for feature weighting
- Strategies for feature selection
- 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.

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

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

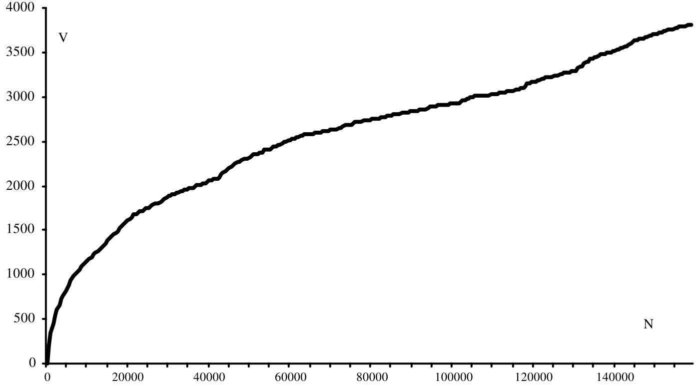


Fig. 1. Chart of vocabulary growth in the tragedies of Racine (chronological order, 500 token intervals).

Complexity and Readability

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

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

```
> 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/)

Discussion Questions

- What does tf-idf have to do with Zipf's law?
- Should we always weight our features? Why/Why not?
- What is a test of independence? How does this help us detect true collocations?
- Why would we want to detect true collocations rather than simply include all bigrams in our dfm?
- How might we apply lexical diversity measures to a social science problem?