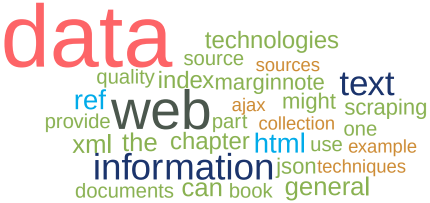
Introduction to Text Mining with R

Introduction



Simon Munzert | IPSDS

Why quantitative analysis of social media text?

Why quantitative analysis of social media text?

Justin Grimmer's haystack metaphor: **automated text analysis improves reading**

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

Why quantitative analysis of social media text?

Justin Grimmer's haystack metaphor: **automated text analysis improves reading**

- Analyzing a straw of hay: understanding meaning
  - 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 automated 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 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 of the post
    - A sentiment or emotion
    - Salience of a political issue

Quantitative text analysis requires assumptions

1. Texts represent an observable implication of some underlying characteristic of interest
   - An attribute of the author of the post
   - 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. n-grams)

Quantitative text analysis requires assumptions

1. Texts represent an observable implication of some underlying characteristic of interest
    - An attribute of the author of the post
    - 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. n-grams)
3. A **document-feature matrix** can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

Key concepts

Some key basic concepts

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

Example:

`A corpus is a set of documents. This is the 2nd document in the corpus.`

→ 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 key basic concepts

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

document: each of the units of the corpus (e.g. a Facebook post)

Example:

`A corpus is a set of documents. This is the 2nd document in the corpus.`

→ 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 key basic concepts

(text) corpus	a large and structured set of texts for analysis
document	each of the units of the corpus (e.g. a Facebook post)
types	for our purposes, a unique word

Example:

`A corpus is a set of documents. This is the 2nd document in the corpus.`

→ 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 key basic concepts

(text) corpus	a large and structured set of texts for analysis
document	each of the units of the corpus (e.g. a Facebook post)
types	for our purposes, a unique word
tokens	any word – so token count is total words

Example:

`A corpus is a set of documents. This is the 2nd document in the corpus.`

→ 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)

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

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

From words to numbers

Bag-of-words approach: workflow

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

docs \ words	made	because	had	into	get	some	through	next	where	many	irish
t06_kenny_fg	12	11	5	4	8	4	3	4	5	7	10
t05_cowen_ff	9	4	8	5	5	5	14	13	4	9	8
t14_ocaolain_sf	3	3	3	4	7	3	7	2	3	5	6
t01_lenihan_ff	12	1	5	4	2	11	9	16	14	6	9
t11_gormley_green	0	0	0	3	0	2	0	3	1	1	2
t04_morgan_sf	11	8	7	15	8	19	6	5	3	6	6
t12_ryan_green	2	2	3	7	0	3	0	1	6	0	0
t10_quinn_lab	1	4	4	2	8	4	1	0	1	2	0
t07_odonnell_fg	5	4	2	1	5	0	1	1	0	3	0
t09_higgins_lab	2	2	5	4	0	1	0	0	2	0	0
t03_burton_lab	4	8	12	10	5	5	4	5	8	15	8
t13_cuffe_green	1	2	0	0	11	0	16	3	0	3	1
t08_gilmore_lab	4	8	7	4	3	6	4	5	1	2	11
t02_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

Bag-of-words approach: from words to numbers

1. **Preprocess text:**

   "A corpus is a set of documents."

   "This is the second document in the corpus."

Bag-of-words approach: from words to numbers

1. **Preprocess text:** lowercase,

“**a** corpus is a set of documents.”

“**t**his is the second document in the corpus.”

Bag-of-words approach: from words to numbers

1. **Preprocess text:** lowercase, remove stop words and punctuation,

“~~a~~ corpus ~~is~~ ~~a~~ set ~~of~~ documents~~.~~”

“~~this~~ ~~is~~ ~~the~~ second document ~~in~~ ~~the~~ corpus~~.~~”

Bag-of-words approach: from words to numbers

1. **Preprocess text:** lowercase, remove stop words and punctuation, stem,

   "corpus set document~~s~~"

   "second document corpus"

Bag-of-words approach: from words to numbers

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

   [corpus, set, document, corpus set, set document]

   [second, document, corpus, second document, document corpus]

Bag-of-words approach: from words to numbers

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

   [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.

	corpus	set	document	corpus set	...	M n-grams
Document 1	1	1	1	1	...	
Document 2	1	0	1	0	...	
...						
Document n	0	1	1	0	...	

Word frequencies and their properties

Bag-of-words approach disregards grammar and word order and uses word frequencies as features. **Why?**

Word frequencies and their properties

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:

Word frequencies and their properties

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

Word frequencies and their properties

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*-grams) are rare

Word frequencies and their properties

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*-grams) are rare
- Some approaches focus on occurrence of a word as a binary variable, irrespective of frequency: a binary outcome

Word frequencies and their properties

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-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, multinomial, and related distributions