Corpus Linguistics

Statistical Measures in Information Retrieval

Niko Schenk

Institut für England- und Amerikastudien
Goethe-Universität Frankfurt am Main
Winter Term 2015/2016

January 10, 2017

Motivation

- **N-Gram statistics** involve **frequency measures** over words (*n*-grams) which can be applied to corpus data. (meaning: you can count words in "different ways")
- Useful to automatically find *interesting* linguistic patterns.
  - E.g., "important words" (keywords) in a collection of document, author-specific vocabulary, characteristics of a certain text genre, topics, collocations, etc.
- $\rightarrow$ Hypothesis **generation** method.
  - as opposed to hypothesis **testing** methods (cf. previous lectures).

Motivation

- Usually, *n*-grams are ranked according to their statistical relevance (from highest to lowest values).
- The topmost *n*-grams/words are "most interesting" (according to some measure of "interestingness").
- We will discuss **five** basic statistical corpus measures from the domain of information retrieval.
  - $\rightarrow$ to find *keywords*, *collocations* and to identify the *author* of a specific text.

A Short Reminder—N-Grams

`https://de.wikipedia.org/wiki/N-Gramm`

1. unigram: 1-word, e.g., [*holidays*]
2. bigram: 2-word phrase, e.g., [*this is*], [*New York*]
3. trigram: 3-word phrase, e.g., [*has been recently*], [*Johann Wolfgang von*]
4. quadgram: 4-word phrase, e.g., [*quite recently . But*], ...
5. ...

1. Introduction

2. *N*-Gram Measures
   - Term Frequency
   - Type-Token Ratio
   - Mutual Information
   - Document Frequency
   - Term Frequency–Inverse Document Frequency

Term Frequency

The **term frequency** (tf) of a term (word/n-gram) t is defined as the number of occurrences of t in a corpus.

Term Frequency

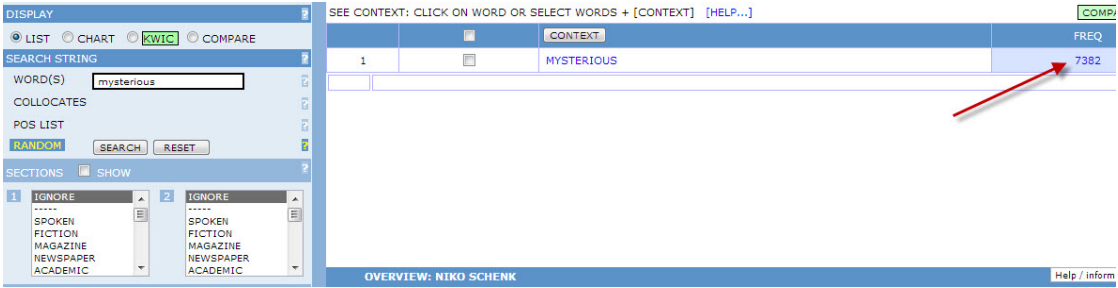

Figure: Term frequency of the unigram "*mysterious*" in the COCA corpus.

Term Frequency

Given an arbitrary English text (corpus),

- what are the most frequent words?
- what is their functionality?
- part-of-speech?

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Article Talk

Most common words in English

From Wikipedia, the free encyclopedia

The list below of **most common words in English** cannot be definitive. It is based on an analysis of the Oxfor Online, associated with the Oxford English Dictionary.[1] This source includes writings of all sorts from "literary Hansard to the language of chatrooms, emails, and weblogs",[2] unlike some sources which use texts from onl

The Reading Teachers Book of Lists claims that the first 25 words make up about one-third of all printed mate

Note that "word" may mean either a word form (essentially, a distinct spelling), or a lexeme (essentially, a "bas occurrences of "are", "is", "were", "was", etc.[5] Note also that these top 100 lemmas listed below account for

Rank	Word	Rank	Word	Rank	Word	Rank	Word	Rank	Word
1	the	21	this	41	so	61	people	81	back
2	be	22	but	42	up	62	into	82	after
3	to	23	his	43	out	63	year	83	use

An experiment from last year...

- Assume our toy corpus consists of all homework assignments and emails which were submitted by each student in the class.
- Results for the most frequent words are very similar, although the corpus consists of only $\approx$ 22k words.

Words Sorted by Term-Frequency in the Students Toy Corpus

the	(1904)
of	(1012)
to	(926)
in	(784)
a	(759)
be	(744)
and	(669)
is	(658)
I	(632)
...	

Term Frequency Distributions for Individual Students

chr...	j...	l...	m...-l...	m...	p...	ph...
the (193)	the (85)	the (116)	the (108)	the (70)	the (91)	the (517)
of (108)	in (42)	of (73)	to (59)	of (53)	in (70)	to (480)
to (104)	you (37)	a (58)	in (53)	corpus (50)	a (69)	in (371)
in (80)	is (31)	and (54)	of (49)	snippet (36)	be (63)	a (370)
a (69)	to (31)	to (53)	a (39)	corpus snippet (34)	of (60)	of (269)
be (66)	of	corpus	is (29)	to (30)	and (55)	a (140)
and (44)	we	in	be (15)	and (25)	corpus (53)	be (130)
I (42)	and	I	one (14)	data (23)	to (49)	I (111)
corpus (40)	that	be	and	from	snippet (35)	and (102)

r...	s...	t...	v...	vi...	ve...	total
the (22)	the (141)	in (32)	the (159)	the (269)	the (101)	the (1904)
in (14)	of (75)	the (23)	to (99)	of (171)	be (90)	of (1012)
a (12)	a (53)	to (22)	it (97)	it (159)	and (82)	to (926)
used (10)	to (39)	corpus (14)	a (88)	be (132)	is (73)	in (784)
word (9)	I (21)	corpus snippet (11)	is (69)	in (111)	corpus (33)	a (759)
words (8)	in (20)	snippet (11)	I (64)	our (98)	of (11)	be (744)
and (7)	is (19)	I (10)	in (32)	from (41)	used	and (669)
used in	and	a	of (20)	one (14)	around	is (658)
for	it	and	and	my	one	I (632)

Properties & Benefits of Using Frequency Lists

- Top-most words are **function words**.
  - Semantically "valuable" words (nouns, verbs, adjectives) are less frequent.
- Given a collection of documents by a particular author, a frequency list is a **characteristic fingerprint** of that author.
- Frequency lists are **comparable**!
  - cf. cosine similarity.
  - Careful: **normalization necessary (e.g., per million words)**

Simple Definition

token = unigram (or word), usually delimited by spaces
type = **distinct** form of a token

$$\textbf{type-token ratio} = \frac{\#types\ (i.e.\ number\ of\ \textbf{different tokens})}{\#tokens\ (i.e.\ number\ of\ \textbf{all tokens})}$$

Example

`This is a nice car. I love this car. It is really fast. Its color is blue.`

Tokenized text (converted to lower-case):
`this is a nice car . i love this car . it is really fast . its color is blue .`

#*tokens* : 21 `this/is/a/nice/car/./i/love/this/car/./it/is/really/fast/./its/color/is/blue/.`

#*types* : 14 `this/is/a/nice/car/./i/love/it/really/fast/its/color/blue`
$\rightarrow$ **type-token ratio** of document $= \frac{14}{21} \approx 0.67$

Importance of the Type-Token Ratio

- The type-token ratio is usually calculated for each **document** or a set of documents (e.g., essays written by a student).
- It usually measures the **richness of vocabulary**.
- The measure can be used for authorship identification.
  - $\rightarrow$ **Texts written by the same person have similar type-token ratios!**
  - characteristic "fingerprint"/writing-style of a person
  - language-independent
  - independent of size of text or document

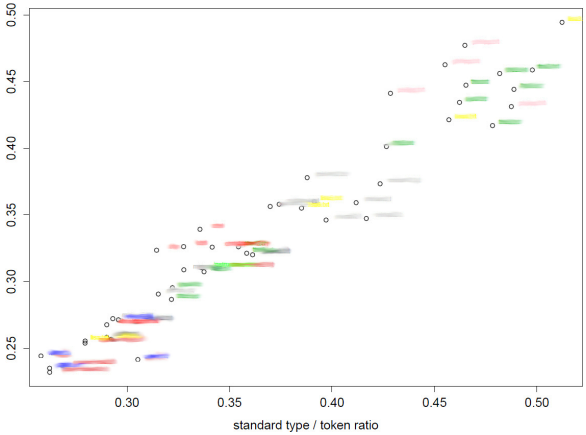

Figure: Type-token ratios for individual student assignments. Documents written by the same student have the same color. Based on the type-token ratio, groupings are visible.

Motivation


blau
00:28
82 %
@goo...
Details
Das
Senden
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das
Senden
ist
war
wird
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das ist
Senden
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das ist
Senden
ein
eine
nicht
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


blau
00:28
82 %
@goo...
Details
Das ist ein sehr guter
Senden
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


●●●●○ blau
00:28
82 %
@goo...
Details
Das ist ein sehr guter
Senden
Tag
Zustand
Zeitvertreib
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


●●●○○ blau
00:28
82 %
@goo...
Details
Das ist ein sehr guter Tag
Senden
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation


●●●○○ blau
00:28
82 %
@goo...
Details
Das ist ein sehr guter Tag
Senden
für
zum
um
q w e r t z u i o p ü
a s d f g h j k l ö ä
y x c v b n m
123
Leerzeichen
Return

Motivation

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

		CONTEXT	FREQ
1	☐	HARD TO	39204
2	☐	HARD .	9972
3	☐	HARD ,	9363
4	☐	HARD FOR	6968
5	☐	HARD WORK	4765
6	☐	HARD AND	4193
7	☐	HARD TIME	4056
8	☐	HARD ON	3172
9	☐	HARD AS	2677
10	☐	HARD AT	2017
11	☐	HARD ENOUGH	1826
12	☐	HARD DRIVE	1661
13	☐	HARD TIMES	1555
14	☐	HARD NOT	1151

Figure: COCA corpus results for the query `hard *` sorted by **frequency**.

Motivation

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

COMPARE ? SIDE

		CONTEXT	FREQ		ALL	%	MI
1		HARD CLAMMING	13		70	18.57	9.31
2		HARD SLOG	38		349	10.89	8.54
3		HARD DISK	526		6325	8.32	8.15
4		HARD KNOCKS	150		2127	7.05	7.91
5		HARD CURRENCY	529		7673	6.89	7.88
6		HARD CANDIES	50		806	6.20	7.73
7		HARD FROSTS	12		199	6.03	7.69
8		HARD BOP	19		370	5.14	7.45
9		HARD CORNERING	15		319	4.70	7.33
10		HARD CLAMS	58		1243	4.67	7.32
11		HARD DISKS	91		2070	4.40	7.23
12		HARD DRIVES	473		11436	4.14	7.14
13		HARD LIQUOR	151		4315	3.50	6.90
14		HARD CIDER	62		1895	3.27	6.80

Figure: Results for the query `hard *` sorted by **relevance**.

Motivation

The **Mutual Information** (*mi*) is an association metric between words.

- It can be calculated for *n*-grams with length ≥ 2.
- *N*-grams whose individual parts combine more frequently than what would be expected by chance have a high *mi* score.

N-grams and Mutual Information—An Example from a Facebook Corpus

Bigram	MI-value
ramba zamba	14.39831
warminski leitheußer	14.21598
rosamunde pilcher	14.21598
nepper schlepper	14.21598
cum laude	14.21598
ante portas	14.21598
ancien régime	14.21598
aldous huxley	14.21598
solem occidisse	14.06183
popolo viola	14.06183
io muoio	14.06183
idi amin	14.06183
fata morgana	14.06183
edwin dutler	14.03366
osmoderma eremita	13.92830
hong kong	13.92830
henrich tiessen	13.92830
goenneram mappumque	13.92830
faux pax	13.92830
ping pong	13.81052

Trigram	MI-value
dierum solem occidisse	10.43753
nondum omnium dierum	10.43753
fehlgeplant und falschgemünzt	10.43753
gedeih und verderb	10.43753
clap clap clap	10.43753
ach und krach	10.43753
divide et impera	10.43753
censeo goenneram mappumque	10.28338
geteert und gefedert	10.28338
jade weser port	10.25521
rechtschreib und grammatikfehler	10.25133
ku klux klan	10.15405
cowboy und indianer	10.14984
gift und galle	10.11780
il popolo viola	10.11780
holla die waldfee	10.10105
hegen und pflegen	10.10105
pacta sunt servanda	10.03206
erstunken und erlogen	10.03206
populi vox rindvieh	10.03206

Computation of Mutual Information

Mutual information for a bigram *bigr* is defined as

$$mi_{bigr} = log(\frac{(tf_{t1,t2} * N_t)}{(tf_{t1} * tf_{t2})})$$

where

- $tf_{t1,t2}$ = term frequency of the bigram
- tf_{t1} = term frequency of the first token in the bigram
- tf_{t2} = term frequency of the second token in the bigram
- N_t = total number of words (tokens) in the corpus

Computation of Mutual Information

Mutual information in plain English: *The measure compares the frequency of the whole expression to the frequency of its parts.*

Mutual Information — Example

- Compute the mutual information for *liebt Farben* based on the following information:
  - frequency of the bigram = 200
  - frequency of *liebt* = 10,500
  - frequency of *Farben* = 2,500
  - number of tokens in the corpus = 2,000,000

Result:

- $$mi_{liebt\ Farben} = log(\frac{(200 * 2,000,000)}{(10,500 * 2,500)}) \approx 1.18$$

Mutual Information — Example

- Compute the mutual information for *IG Farben* based on the following information:
  - frequency of the bigram = 50
  - frequency of *IG* = 45
  - frequency of *Farben* = 2,500
  - number of tokens in the corpus = 2,000,000

Result:

- $$mi_{IG\ Farben} = log(\frac{(50 * 2,000,000)}{(45 * 2,500)}) \approx 2.95$$

Mutual Information — Example

$$mi_{liebt\ \ Farben} \approx 1.18$$

$$mi_{IG\ \ Farben} \approx 2.95$$

Explanation: the contextual variation in which *liebt* occurs in a corpus is much greater compared to *IG*. $\rightarrow$ *IG Farben* serves better as a collocation.

Motivation

- Usually, corpora are split up into smaller units called **documents**.
  - A document is a *subcorpus* whose contents share the same properties.
  - E.g., each student essay in a learner corpus could be represented by a document.
  - Moreover, the BNC/COCA has different genres which could be considered documents. (Each genre again consists of individual documents).

Document Frequency

The **document frequency** measures the number of documents in a corpus in which a particular term (word) appears.

a	9
academic	9
academic writing	9
also	9
and	9
are	9
as	9
assignment	9
be	9
can	9
corpus	9
different	9
for	9
i	9
in	9
in a	9
in the	9
and the	8
at	8
at the	8
best	8
by	8

Table: Document frequencies for a (subset) of *n*-grams for the students corpus consisting of 9 documents. The word *academic*, e.g., occurs in all nine documents, *at* occurs in only eight.

Intuition

- Suppose we have a corpus consisting of individual documents.
- A term which occurs **frequently in only a small subset of the documents** and not so often in all the other documents is more important for this subset of documents (**keyword !**) than...
  - ...a term which occurs **only infrequently** in the whole corpus.
  - ...a term which occurs **frequently in all** documents.

Keywords—An Example

Imagine, you have three text documents – two about politics/the Iraq war and one about a recent soccer/sports event.
Assume further, that the words *Obama* and *offside* occur in the documents.
Informally:

- $\rightarrow$ *Obama* is probably a good keyword describing the first two documents.
- $\rightarrow$ The word *offside* would be a suitable keyword for the third document.
- The word *the* is not a good keyword for any of the documents (because it occurs in all documents equally frequently).

Keywords—Extraction

In what follows, we describe a statistical measure to extract **keywords** automatically from a specific text document.

Keywords—Extraction

We need two factors:

1. A local one: the term frequency of the word in a specific document.
2. A global one: showing how many documents contain the word elsewhere.

Ideally, a keyword occurs **often within a specific document** (local), but does **not show up** in all the **other** documents (global).

Local Factor

We are already familiar with the term frequency of a word.
The local factor considers **only a specific document**:

tf_{t_d} = term frequency of word t in document d.

Global Factor

The global factor **inverse document frequency** is defined as

$idf_t = \log(1 + \frac{N_D}{df_t})$, where

- idf_t = inverse document frequency of term t (t can be a word or a general n-gram)
- N_D = total number of documents in the corpus
- df_t = document frequency of term t in the corpus

Term Frequency–Inverse Document Frequency

The **term frequency–inverse document frequency** of a word t in document d is defined as:

$$tf_{t_d} * idf_t$$

where

- tf_{t_d} = term frequency of word t in document d
- idf_t = inverse document frequency of word t

and serves as an indicator of how good word t serves as a keyword in document d.

Homework Assignment

Three Text Documents

1. This is a nice car. I love this car. It is really fast. Its color is blue. I've bought it recently and it was a bargain.
2. Tübingen is a beautiful town. I've been there a couple of times.
3. Lorem ipsum dolor love sit amet, consectetuer adipiscing elit. Aenean this is commodo lorem its ligula eget dolor is. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus Tübingensis mus.

Homework Assignment

Given a corpus consisting of the three text documents (1-3) and based on the formulae from the previous slides

1. Provide the three **normalized** texts and write them into a single text file.
   - You should **normalize** the texts first (convert to lower-case), i.e. we do not want to treat "This" and "this" as two different occurrences of the same word. Moreover, you can assume a simple whitespace-based tokenization of the words. **Note that punctuation (periods, apostrophes, hyphens) should be removed** in this application before computing statistics! Do **NOT** lemmatize the words.

Term Frequency

2. Compute the global term frequency for all unigrams in the corpus and rank them from highest to lowest. Do the same for all bigrams with tf ≥ 2. What is the most frequent trigram which occurs more than only once in the corpus?
3. What's the proportion of unigram types which occur only once in the corpus?
4. Compute the type-token ratio for documents 1 and 2. Which document exhibits a larger vocabulary? Explain why!

Term Frequency

Assume the text come with English meta information about parts of speech and lemmata,[1]

5. Extract the most frequent English **noun** and its term frequency.
6. What is the most frequent English **verb** (singular, present tense)?
7. Extract the most frequent English **lemma** and its term frequency.

[1] You can assume that unknown words (non-English) get a part of speech label UNKNOWN. For unknown words, the lemma is the same as the observed word.

Mutual Information

8. Compute the mutual information for *stuttgart 21* based on the following information from real corpus data:
   - frequency of *stuttgart* = 3,001
   - frequency of *21* = 10,500
   - frequency of the bigram = 4,012
   - number of tokens in the corpus = 2,100,227

Mutual Information

9. Compute the mutual information for *stuttgart hat* based on the following information:
   - frequency of *stuttgart* = 3,001
   - frequency of *hat* = 60,500
   - frequency of the bigram = 5,013
   - number of tokens in the corpus = 2,100,227

What is your conclusion from comparing the mutual information of the previous two phrases? Which one severs better as a collocation? Explain why!

Mutual Information

10. Finally, calculate the mutual information of the bigrams *this is* and *been there* given the three previous text documents. Rank them according to their relevance.

Document Frequency

11. Compute the document frequency of the unigrams *tübingen*, *lorem* and *car*.
12. In our toy corpus consisting of three documents, what is the word with the highest df?

Term Frequency–Inverse Document Frequency

13. Compute the term frequency for the word *is* **only** in document number 3.
14. Compute the inverse document frequency of the word *is*.
15. The *tf*idf* for *is* in document 3 is defined as the product the term frequency of *is* in document 3 (local factor) and the inverse document frequency of the word *is* (global factor). Compute the value.
16. Similarly, compute the *tf*idf* for the word *lorem* in document 3.
17. Based on the numerical results from the previous two exercises explain why one of them serves as a better keyword (in document 3). *lorem* or *is*? Why?