Natural Language Processing
自然語言處理

黃瀚萱
Department of Computer Science
National Chengchi University
2020 Fall

Lesson 3

Collocation

Schedule

Date	Topic
9/16	Introduction
9/23	Linguistic Essentials
9/30	**Collocation**
10/7	Language Model
10/14	Word Sense Disambiguation
10/21	NLP and Cybersecurity
10/28	Text Classification
11/4	POS Tagging
11/11	Midterm Exam

Schedule

Date	Topic
11/18	Chinese Word Segmentation
11/25	Word Embeddings
12/2	Neural Networks for NLP
12/9	Parsing
12/16	Discourse Analysis
12/23	Invited Talk
12/30	Final Project Presentation I
1/6	Final Project Presentation II
1/13	Final Exam

Agenda

- Terminology mining
  - TF-IDF
- Collocation mining
  - Frequency based
  - Hypothesis Testing
  - Mutual Information
  - Distant Collocations

Terminology Mining

- What are the important terms in a collection of data?
  - Capturing the semantic concepts of a document or a document set.
- Discovering the new important terminologies.
  - Emmanuel **Macron** (馬克宏) is not an important term in the past.

Hot Term Detection from Newspapers

Feb-17	Mar-17	Apr-17	May-17	Jun-17	Jan-18	Apr-18	Jul-18	Oct-18	Nov-18	Dec-18
Uber	西屋	阿塞德	菲立普	溫畢爾	沃爾夫	陳香梅	納隆薩	哈紹吉	戈恩	歐斯
馬雲	金正	東協	穆勒	羅德曼	歐普拉	奧班	清萊	艾克巴	阿科斯達	福州
潘基文	大馬	特斯拉	Comey	李光耀	李善權	陳納德	岡山縣	Khashoggi	黃背心	長孟
阿馬德	韓松	阿布沙伊夫	英航	Grenfell	馬永火山	芭芭拉	沙曼	Jamal	美黛	幼發
利華	檢察廳	麥馬斯特	博明	格蘭菲塔	趙明均	凱勒	伊姆蘭汗	鮑爾斯	坎普	康明凱
特莉	施特金	艾提	羅森斯坦	李顯揚	草津	EnBW	黑門	USMCA	IN	張首晟
卡爾文森	籠池	梅蘭雄	羅哈尼	費茲	Wolff	度瑪鎮	卡瓦諾	西爾斯	世博	蓋亞
梅蘭妮亞	特幣	沃克特	馬巫德	雲頂	DACA	博鰲	Narongsak	巴路	div	瑞隆茨
巴菲特	任天堂	阿尼斯	李明哲	沙迪克汗	李春興	宋濤	Non	孟宏偉	洞察號	卡塔尼

Term Frequency

- Finding the most common terms
  - Counting the frequency of each term
  - **Term frequency** (TF): How many times a term appears
- The frequent terms are more contributing to the documents?

Most Common Terms

- A set of news articles related to the topic Coffee
- Stopwords dominate the list of most common words

Term	TF
the	1462
to	968
of	819
in	609
and	585
said	576
a	546
coffee	375
for	320
on	272

Document Frequency

- The words that appear in almost all documents are less important because they are not discriminative.
    - We will like to penalize these terms.
- **Document Frequenc**y (DF): How many documents a term appears.

Document Frequency

Term	DF
the	817
to	669
be	639
say	589
of	583
a	492
and	483
in	472
coffee	334
for	285

Term	DF
Average	1
equally	1
reallocate	1
Jan	1
equitable	1
opportunity	1
absolute	1
scale	1
mouth	1
mind	1

TF-IDF

- TD-IDF: a combination of term frequency (TF) and inverse document frequency (IDF)

Total number of documents

$$\text{TF-IDF}_t = tf_t \times \log \frac{N}{df_t}$$

Flatten the drastic curve of IDF

Terms with High TF-IDF

- However, the high frequent terms still dominate

Term	TFIDF
a	607.885
of	595.312
in	571.294
to	570.423
the	569.340
and	535.303
be	520.196
coffee	485.323
for	461.624
have	458.773

Variants of TF-IDF

- Original

$$\text{TF-IDF}_t = tf_t \times \log \frac{N}{df_t}$$

- Further reduce the impact of term frequency

$$\text{TF-IDF}_t = (1 + \log tf_t) \times \log \frac{N}{df_t}$$

$$\text{TF-IDF}_t = (0.5 + 0.5 \times \frac{tf_t}{\max_s tf_s}) \times \log \frac{N}{df_t}$$

Terms with Normalized TF-IDF

$$\text{TF-IDF}_t = (0.5 + 0.5 \times \frac{tf_t}{\max_s tf_s}) \times \log \frac{N}{df_t}$$

Term	TFIDF
Cherry	3.574
Arabica	3.560
Plantation	3.555
Bulk	3.555
BBB	3.555
AVERAGE	3.552
raw	3.552
gentleman	3.552
premium	3.552
tough	3.552

$$\text{TF-IDF}_t = (1 + \log tf_t) \times \log \frac{N}{df_t}$$

Term	TFIDF
Cherry	24.108
NIL	22.075
NA	21.821
Robusta	20.375
SAO	19.895
THE	19.398
MM	19.398
It	19.117
Bank	19.093
Ltd	18.906

Colabs: http://tiny.cc/8pbysz

Collocations (搭配詞)

- A collocation is a phrase consisting of two or more words with **limited compositionality**.

- Compositional
  - The meaning can be figured from the meanings of the parts.
  - nice guy ≈ nice + guy
  - Black cat ≈ black + cat

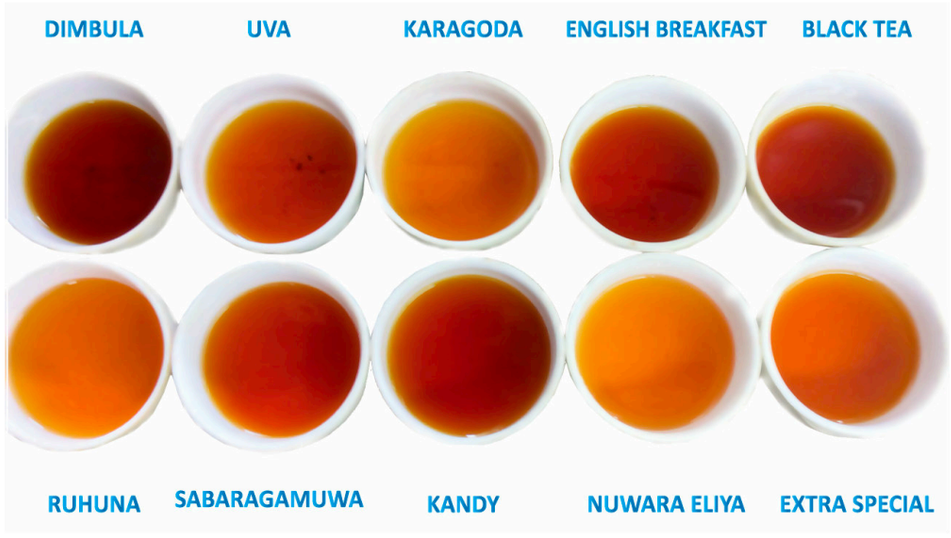
Non-Compositional

- Non-compositional
  - 1 + 1 ≠ 2
- make up ≠ make + up

- strong tea ≠ strong + tea

rich in some active agent

having a great physical strength



https://zh.wikipedia.org/wiki/File:Pure_Ceylon_Single_Estate_Black_Tea_in_Australia.jpg https://www.flickr.com/photos/reanjos/3048554599

Definition of Collocation

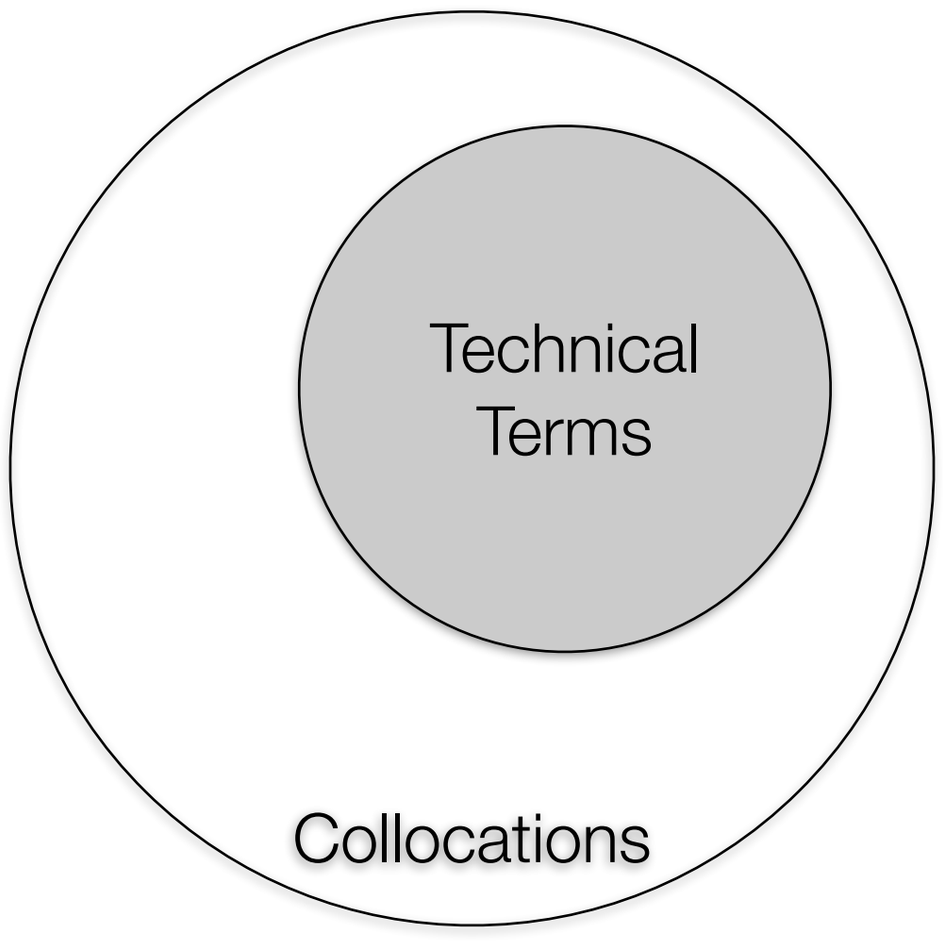
- A collocation is a sequence of two or more **consecutive** words that has the properties of a syntactic and semantic unit and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components. (Chouekra, 1998)

- The components in a collocation are not necessarily consecutive.
  - **make** my face **up**
- Collocations cannot be directly translated.
  - Green house => 溫室　綠色房子

https://pixabay.com/photos/greenhouse-green-house-hothouse-395492/

Collocation vs Technical Terms

- Terms, technical terms, Terminological phrases
  - The collocations extracted from technical domains with the process **terminology extraction**.



Terminology Extraction

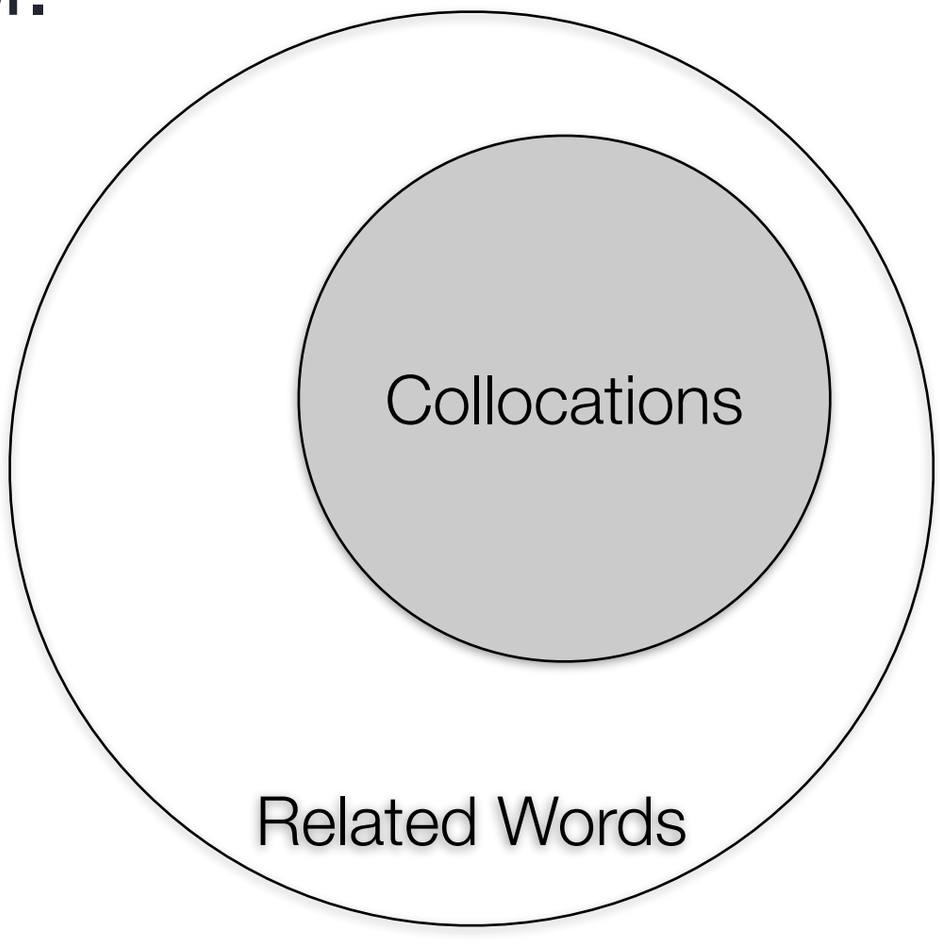
- So, the process of terminological extraction can be seen as the process of collocation mining on a collection of documents from a specific domain such as medical and law.

Collocation vs Related Words

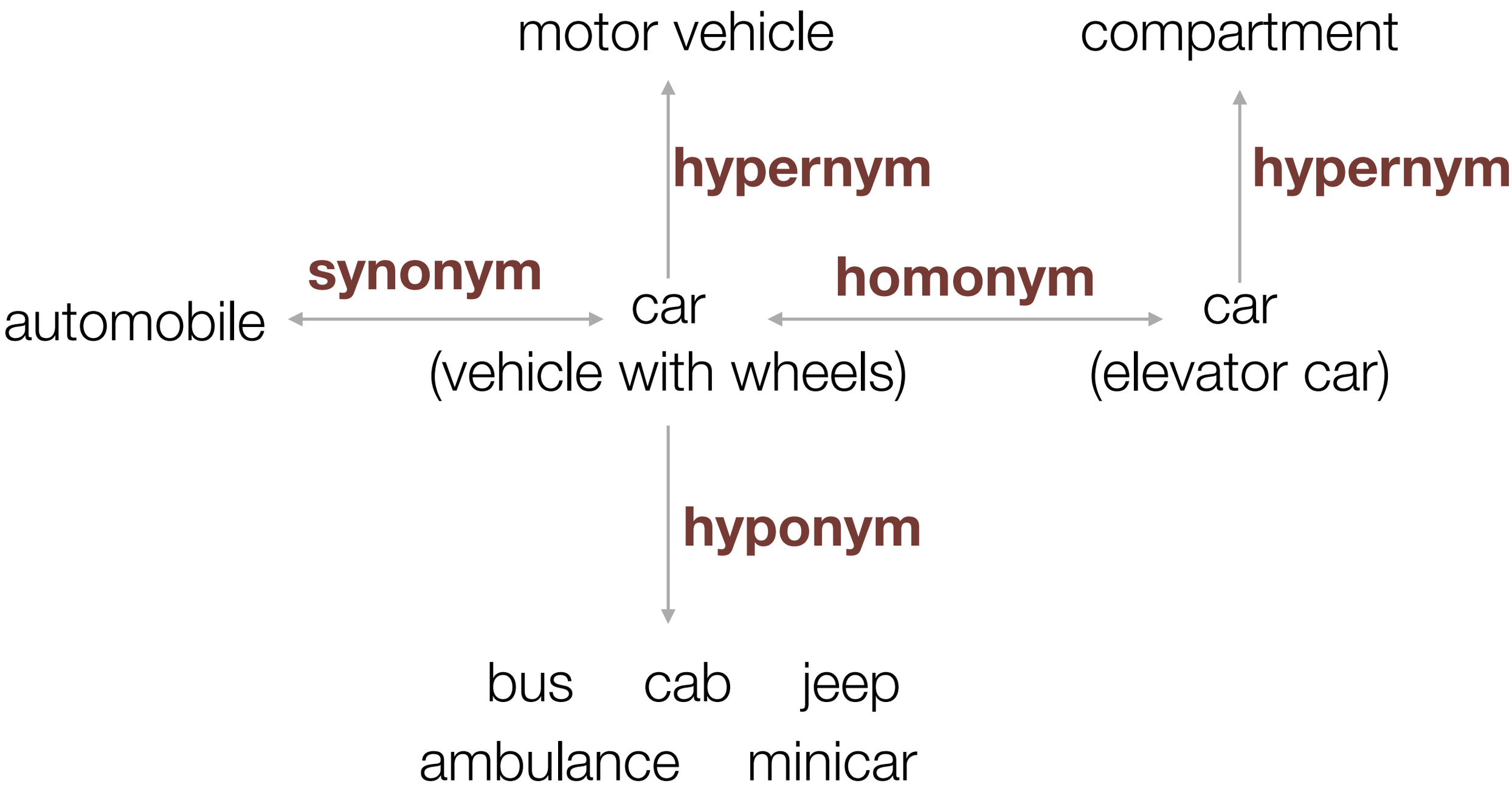
- Related words (aka. associated words) are associated with each other.

- Do not necessarily occur in a common grammatical unit and with a particular order.

  - doctor and nurse
  - tiger and lion
  - plane and airport



Related Words in WordNet

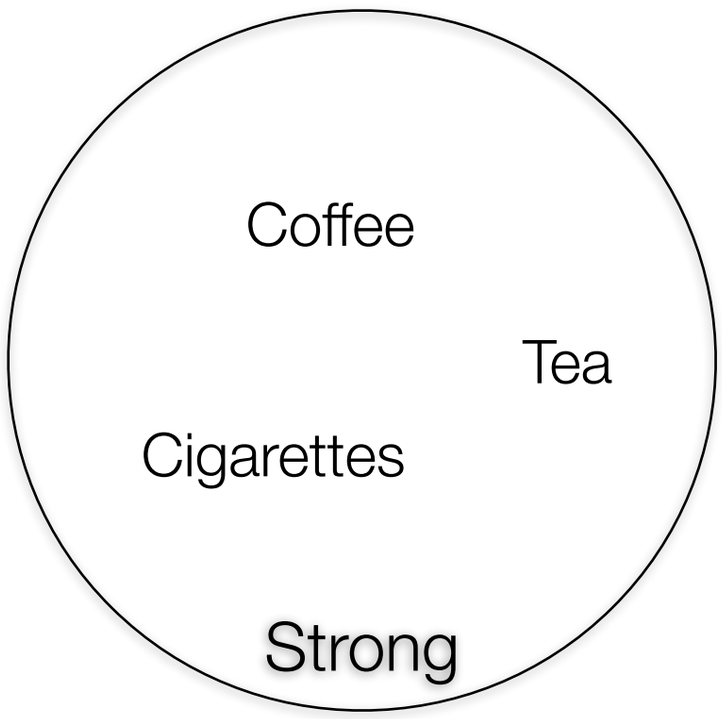


Applications of Collocations

- Computational lexicography
  - Automatically building a dictionary
- Natural language generation
  - powerful tea vs strong tea
  - take a decision vs make a decision
- Parsing
- Linguistic research
- Language education

Contextual Theory of Meaning

- Collocations sometimes reflect interesting attitudes towards different types of substances.
  - Strong tea vs powerful drug



Types of Collocations

- Light verb (輕動詞) + Noun
  - Call for papers
- Verb particle（語助詞）
  - Make up
- Proper Nouns / Names (專有名詞)
  - Donald Trump
- Terminological Expressions
  - Topological sort

Types of Collocations

- Light verbs
  - Verbs with little semantic content
  - The semantic meaning is contributed by the collocated nouns

Verb	Noun
have	a dinner
took	a walk
have	a haircut

Verb Particles

- A phrasal verb is a verb plus a preposition (V + P) that creates a meaning different from the original verb.
    - Make up -> put make-up on one’s face
    - Take in -> decieve

Multiword Expressions

- Composed of two or more words and syntactically and semantically **idiosyncratic** (特殊性)

- Act as a single unit at some level of linguistic analysis
  - Idioms: kick the bucket (嗝屁、翹辮子)
  - Compound nouns: post office
  - Verb particle: **Look** something **up**
  - Proper nouns: New York

Types of Multiword Expressions

- Fixed
    - in short vs in very short
- Semi-fixed
    - non-decomposable idioms
        - They **kick the bucket**
        - He **kicks the bucket**
        - The bucket was kicked **(Non-sense)**
- Compound nominals
    - car park, car parks
- Proper names

Types of Multiword Expressions

- Syntactically-Flexible Expressions
  - Decomposable idioms
    - let the cat out of the bag (洩漏秘密)
    - the cat is out of the bag
  - Verb-particle constructions
  - Light verbs
- Institutionalized phrases (習慣用法)
  - salt and pepper
  - Bread and butter
  - traffic light

Criteria of Collocations

- Non-compositional (不可拆分)
- Non-substitutable (不可替換)
- Non-modifiable (不可變動)

Non-Compositional (不可拆分)

- A phrase is compositional if its meaning is a straightforward composition of the meanings of its parts.
  - new companies
- A phrase is non-compositional if the meaning cannot be predicted from the meaning of the parts.
  - hot dog ← not really hot; not a dog.

- Collocations are not always fully non-compositional.
  - It is usually an element of meaning added to the combination.
  - **strong** tea ← the meaning of tea remains.
- Idioms are the most extreme examples of non-compositionality.
  - to hear it through the grapevine (據小道消息得知)
  - kick the bucket (翹辮子)

cannot be inferred from the literal forms

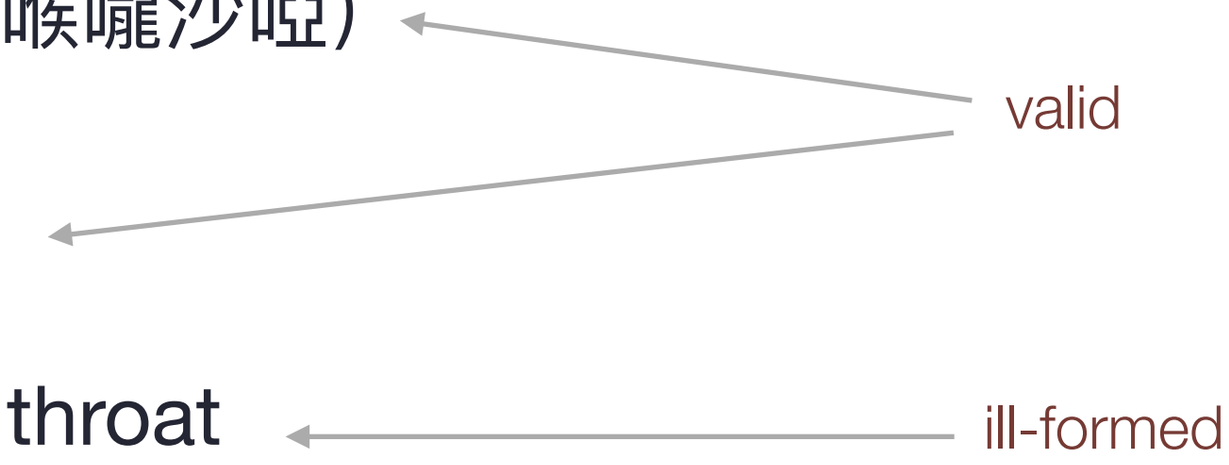
Non-Substitutable (不可替代)

- The component of a collocation cannot be replaced with other words even if they have the same meaning.
  - Yellow wine vs white wine
    - a better description of the color of white wine

Non-modifiable (不可變動)

- Many collocations cannot be freely modified.
  - with additional words.
  - with grammatical transformation.

- Frozen expressions: idioms
  - I have a frog in **my** throat (喉嚨沙啞)
  - He has a frog in **his** throat
  - He has an **ugly** frog in **his** throat



Approaches to Collocation Mining

- Frequency based
- Hypothesis testing
- Mutual information
- Mean and Variance (for distant collocations)
  - The components are not consecutive.

Frequency-based Approach

- Basic idea: Two (or more) words may have a special function if they frequently co-occur together.

- The most frequently occurring **bigrams**

- Filtering the results with part-of-speech information.

- Simple computational method + linguistic knowledge = It works.

The Brown Corpus

- The first “big” digital corpus
  - 1.15M words with part-of-speech tagged.
- Categorized in 15 genres
  - adventure, belles lettres, editorial, fiction, government, hobbies, humor, learned, lore, mystery, news, religion, reviews, romance, and science fiction.
- Created in 1961 at the Brown University

N-grams

N				
1	programming	for	social	scientists
2	programming for	for social	**social scientists**	
3	programming for social	for social scientists		
4	programming for social scienitst			

Collocation

Raw Frequency Counting of the Brown corpus

- The results are not interesting.
- All function word pairs.

Word 1	Word 2	Count
of	the	9625
in	the	5546
to	the	3426
on	the	2297
and	the	2136
for	the	1759
to	be	1697
at	the	1506
with	the	1472
of	a	1461

Filtering with POS Tags

- Very simple heuristic method that filters the results with POS tag patterns. (Justeson & Katz, 1995)

N	POS Tag Pattern	Example
Bigram	Adjective Noun	good place
Bigram	Noun Noun	computer network
Trigram	Adjective Adjective Noun	artificial neural network
Trigram	Adjective Noun Noun	probability distribution
Trigram	Noun Adjective Noun	mean squared error
Trigram	Noun Noun Noun	computer network tutorial
Trigram	Noun Preposition Noun	interest of conflict

Top Bigram Collocations

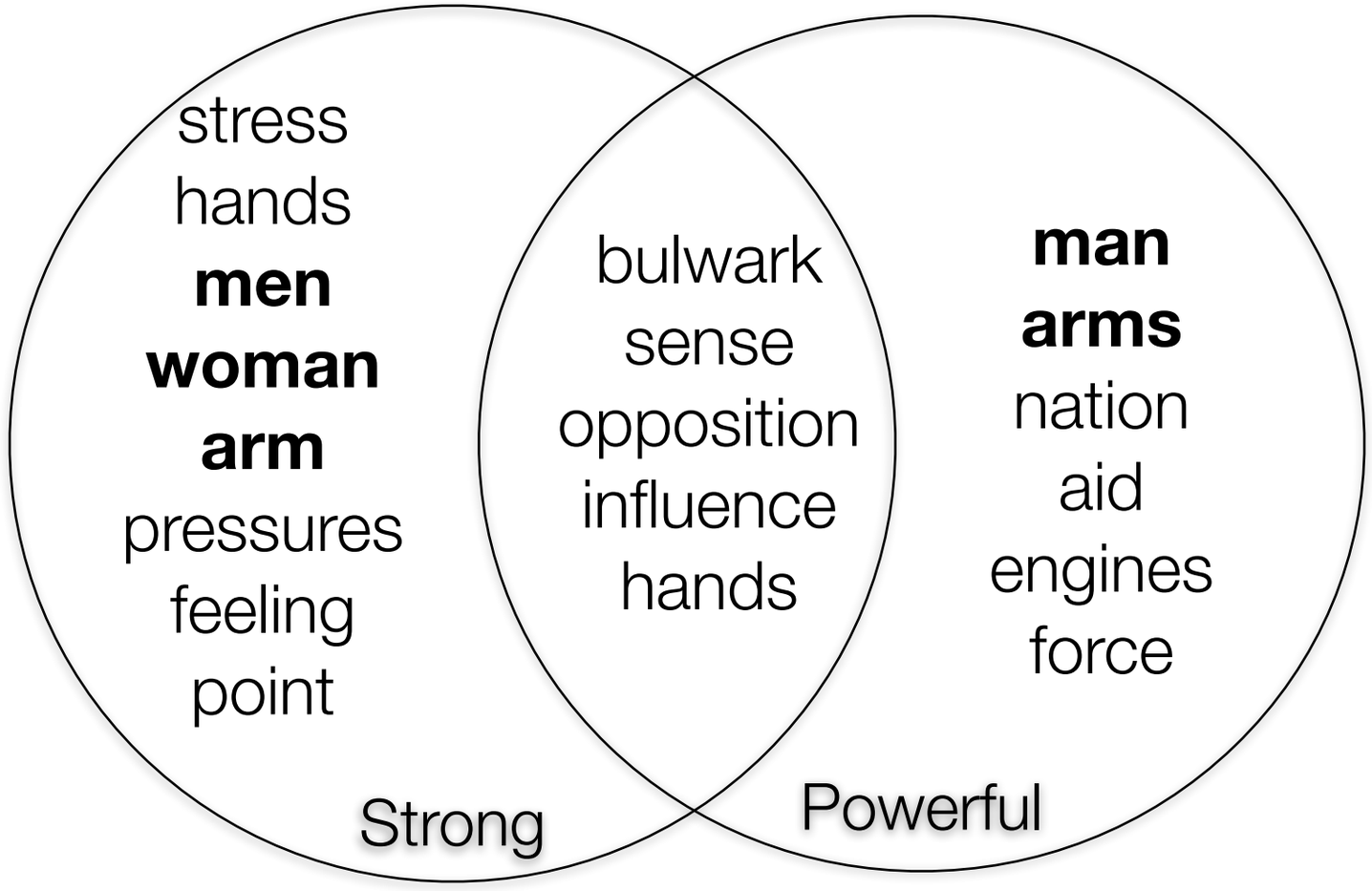
Word 1	Word 2	Count	Pattern
Rhode	Island	90	N N
World	War	60	N N
U.	S.	57	N N
fiscal	year	56	A N
high	school	54	A N
old	man	52	A N
Peace	Corps	52	N N
Los	Angeles	47	N N
young	man	47	A N
great	deal	43	A N
President	Kennedy	40	N N
General	Motors	40	N N
long	time	39	A N
San	Francisco	39	N N
Du	Pont	34	N N

Top Trigram Collocations

Word 1	Word 2	Word 3	Count	Pattern
way	of	life	28	N P N
point	of	view	26	N P N
time	to	time	24	N P N
period	of	time	20	N P N
matter	of	fact	17	N P N
basic	wage	rate	16	A N N
Drug's	chemical	name	15	A A N
John	A.	Notte	15	N N N
number	of	people	13	N P N
small	business	concerns	12	A N N
years	of	age	12	N P N
number	of	years	12	N P N
couple	of	weeks	11	N P N
uniform	fiscal	year	11	A A N
General	Motors	stock	10	N N N

Strong vs Powerful

- Now we can compare the collocations of strong with those of powerful.
- Note that the data is pretty sparse so that the reliability can be an issue.



Drawback of Frequency Based Approach

- They are frequent, but the combination is not significant.

Word 1	Word 2	Count
of	the	9625
in	the	5546
to	the	3426
on	the	2297
and	the	2136
for	the	1759
to	be	1697
at	the	1506
with	the	1472
of	a	1461

Hypothesis Test Method

- With an assumption that probabilities are approximately normally distributed.

- Collocations are something that occur abnormally.

    - The collocations can be verified with a significant test.

- The *t* test

- Pearson’s chi-square test

Pearson's Chi-Square Test for Collocation Mining

	w_1	Rest of w_1
w_2	O_{11}	O_{12}
Rest of w_2	O_{21}	O_{22}

$$X^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

$$N = O_{11} + O_{12} + O_{21} + O_{22}$$

$$X^2 = 3.841 \text{ at a probability level of } \alpha = 0.05$$

Pearson's Chi-Square Test for Collocation Mining

	w_1	Rest of w_1
w_2	O_{11}	O_{12}
Rest of w_2	O_{21}	O_{22}

	w_1	Rest of w_1
w_2	$C(w_1\ w_2)$	$C(w_1\ -w_2)$
Rest of w_2	$C(-w_1\ w_2)$	$C(-w_1\ -w_2)$

Pearson's Chi-Square Test for Collocation Mining

	w_1	Rest of w_1
w_2	O_{11}	O_{12}
Rest of w_2	O_{21}	O_{22}

	w_1	Rest of w_1
w_2	*C(computer science)*	*C(computer AND NOT science)*
Rest of w_2	C(NOT computer AND science)	C(NOT computer AND NOT science)

Check Collocation with Google Search

	w_1	Rest of w_1
w_2		
Rest of w_2		

N = 1,060,000,000

Check Collocation with Google Search

	w_1	Rest of w_1
w_2	47,200	
Rest of w_2		

$$N = 1{,}060{,}000{,}000$$

Check Collocation with Google Search

	w_1	Rest of w_1
w_2	47,200	304,000 - 47,200
Rest of w_2		

$$N = 1{,}060{,}000{,}000$$

Check Collocation with Google Search

	w_1	Rest of w_1
w_2	47,200	304,000 - 47,200
Rest of w_2	29,000,000 - 47,200	

$$N = 1{,}060{,}000{,}000$$

Check Collocation with Google Search

	w_1	Rest of w_1
w_2	47,200	256,800
Rest of w_2	28,952,800	1,060,000,000-47,200 -256,800-28,952,800

$$N = 1{,}060{,}000{,}000$$

Check Collocation with Google Search

	w_1	Rest of w_1
w_2	47,200	256,800
Rest of w_2	28,952,800	1,030,743,200

$$X^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

$$N = O_{11} + O_{12} + O_{21} + O_{22}$$

$$X^2 = 3.841 \text{ at a probability level of } \alpha = 0.05$$

Check Collocation with Google Search

	w_1	Rest of w_1
w_2	47,200	256,800
Rest of w_2	28,952,800	1,030,743,200

$$
\begin{aligned}
&X^2 \\
&= \frac{1060000000 \times (47200 \times 1030743200 - 256800 \times 28952800)^2}{(47200 + 256800) \times (47200 + 28952800) \times (28952800 + 1030743200) \times (256800 + 1030743200)} \\
&= 186950.231 \\
&> 3.841
\end{aligned}
$$

Google "Donald Trump" site:en.wikipedia.org

全部 圖片 新聞 影片 地圖 更多 設定 工具

約有 29,100 項結果 (搜尋時間：0.42 秒)

「"Donald Trump" site:en.wikipedia.org」的圖片搜尋結果

→ 更多符合「"Donald Trump" site:en.w

Donald Trump - Wikipedia

https://en.wikipedia.org › wiki › Donald_Tr

Donald John Trump (born June 14, 1946) is the
Before entering politics, he was a businessman

Google "Donald" site:en.wikipedia.org

全部 圖片 新聞 影片 地圖 更多 設定 工具

約有 232,000 項結果 (搜尋時間：0.59 秒)

Donald - Wikipedia

https://en.wikipedia.org › wiki › Donald ▾ 翻譯這個網頁

given name derived from the Gaelic name Dòmhnall. This comes from the
ualos The final -d in Donald is partly ...

", "world wielder" **Language(s)**: English, Scottish Gaelic
Donny, Dolly **Derivation**: Proto-Celtic *dumno-ualos

· Given name · Don · Donald

Google "Trump" site:en.wikipedia.org

全部 新聞 圖片 影片 地圖 更多 設定 工具

約有 60,800 項結果 (搜尋時間：0.58 秒)

Donald Trump - Wikipedia

https://en.wikipedia.org › wiki › Donald_Trump ▾

Donald John Trump (born June 14, 1946) is the 45th and current president of the United States. Before entering politics, he was a businessman and television ...

Vice President: Mike Pence

Political party: Republican (1987–1999, 2009–...

Education: The Wharton School (BS in Econ.)

Born: Donald John Trump; June 14, 1946 (age ...

Family and personal life · Business career

Check “Donald Trump” with Google Search

	w_1	Rest of w_1
w_2	29,100	232,000
Rest of w_2	60,800	1,060,000,000-29,100 -232,000-60,800

$$
\begin{aligned}
X^2 &= \frac{1060000000 \times (29100 \times 1059678100 - 232000 \times 60800)^2}{(29100 + 232000) \times (29100 + 60800) \times (60800 + 1059678100) \times (256800 + 1059678100)} \\
&= 38195120.21 \\
&> 3.841
\end{aligned}
$$

From Formula to Python Code

```
def chisquare(o11, o12, o21, o22):

    n = o11 + o12 + o21 + o22

    x_2 = (n * ((o11 * o22 - o12 * o21)**2)) / ((o11 + o12) *
(o11 + o21) * (o12 + o22) * (o21 + o22))

    return x_2
```

$$X^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

$$N = O_{11} + O_{12} + O_{21} + O_{22}$$

Collocations with Chi-Square Test

Word 1	Word 2	Count	X^2
danish	languages	1	11780
fanatics	hole	1	11780
battles	lies	1	11780
egotistical	calculation	1	11780
greatest	pleasure	1	11780
duodecimo	editions	1	11780
decisive	hour	1	11780
nursery	tale	1	11780
czar	metternich	1	11780
instinctive	yearnings	1	11780
numberless	indefeasible	1	11780
trades	unions	1	11780

More Frequent Collocations with Chi-Square Test

Word 1	Word 2	Count	X^2
productive	forces	9	9636.544203
middle	ages	7	4928.714781
no	longer	14	4150.496033
working	class	23	2477.732678
modern	industry	11	1128.037662
class	antagonisms	11	1042.736309
private	property	7	1022.522314
ruling	class	11	966.767323
can	not	9	775.745125
their	own	11	759.449519
proportion	as	8	720.619853
have	been	7	702.43862

Less Significant Collocations

Word 1	Word 2	Count	X^2
and	of	20	0.906966
of	class	9	0.569228
bourgeoisie	the	7	0.548588
that	the	15	0.476118
of	its	7	0.436822
and	in	11	0.401790
society	the	6	0.307430
all	the	11	0.192300
the	property	6	0.041125
and	to	9	0.027804
the	class	10	0.009971
class	the	10	0.009971

Collocations with Chi-Square Test and Stopword Removing

Word 1	Word 2	Count	X^2
third	estate	2	11780.000000
constitution	adapted	2	11780.000000
productive	forces	9	9636.544203
eternal	truths	3	8834.249809
corporate	guilds	2	7852.666553
absolute	monarchy	4	7537.599406
eighteenth	century	3	7066.799694
immense	majority	3	6624.749575
laid	bare	2	5888.999830
distinctive	feature	2	5234.221968
torn	asunder	2	5234.221968
middle	ages	7	4928.714781

Less Significant Collocations with Chi-Square Test and Stopword Removing

Word 1	Word 2	Count	X^2
old	property	2	28.685615
bourgeois	revolution	2	26.136797
modern	bourgeoisie	3	21.380029
whole	bourgeoisie	2	20.752016
revolutionary	class	2	20.224783
bourgeois	state	2	17.063629
every	class	2	16.075081
bourgeois	conditions	3	16.040705
bourgeois	form	2	14.680146
one	class	2	10.562861
bourgeois	production	2	5.662454
bourgeois	class	3	5.261506

> 3.841

Mutual Information

- Pointwise mutual information

$$I(x', y') = log_2 \frac{P(x', y')}{P(x')P(y')}$$

$P(x', y')$: joint probability of events x' and y'.

$P(x')$: probability of the event x'.

$P(y')$: probability of the event y'.

Mutual Information for Collocation Mining

$$I(w_1, w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

$$P(w_1, w_2) = \frac{\text{Occurrences of } w_1 w_2}{\text{Number of all bigrams}}$$

$$P(w) = \frac{\text{Occurrences of } w}{\text{Number of all unigrams.}}$$

Computing Collocation with Google

$$I(computer, science) = log_2 \frac{P(computer, science)}{P(computer)P(science)}$$

$$= log_2 \left(\frac{\frac{47200}{1060000000}}{\frac{256800}{1060000000} \frac{28952800}{1060000000}} \right)$$

$$= 2.750$$

Check MI of “Donald Trump”

$$
\begin{aligned}
I(Donald, Trump) &= log_2 \frac{P(Donald, Trump)}{P(Donald)P(Trump)} \\
&= log_2 \left(\frac{\frac{29100}{1060000000}}{\frac{232000}{1060000000} \frac{60800}{1060000000}} \right) \\
&= 11.055
\end{aligned}
$$

From Formula to Python Code

```
import math

def mutual_information(w1_w2_prob, w1_prob, w2_prob):

    return math.log2(w1_w2_prob / (w1_prob * w2_prob))
```

$$I(w_1, w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

$$P(w_1, w_2) = \frac{\text{Occurrences of } w_1 w_2}{\text{Number of all bigrams}}$$

$$P(w) = \frac{\text{Occurrences of } w}{\text{Number of all unigrams.}}$$

Results of Mutual Information

Word 1	Word 2	Count	MI
danish	languages	1	13.524297
fanatics	hole	1	13.524297
battles	lies	1	13.524297
egotistical	calculation	1	13.524297
greatest	pleasure	1	13.524297
duodecimo	editions	1	13.524297
decisive	hour	1	13.524297
nursery	tale	1	13.524297
czar	metternich	1	13.524297
instinctive	yearnings	1	13.524297
numberless	indefeasible	1	13.524297
trades	unions	1	13.524297

Results of Mutual Information with Count > 5

Word 1	Word 2	Count	MI
productive	forces	9	10.064865
middle	ages	7	9.461287
no	longer	14	8.215308
private	property	7	7.202369
working	class	23	6.762457
modern	industry	11	6.699483
have	been	7	6.669874
class	antagonisms	11	6.582849
proportion	as	8	6.513070
ruling	class	11	6.475934
can	not	9	6.453431
just	as	6	6.223563

New Term Extraction with MI

- Pairwise mutual information can be easily extended to more terms

- Bigram

$$\mathrm{MI}(a,b) = \log \frac{P(a,b)}{P(a)P(b)}$$

- Trigram

$$\mathrm{MI}(a,b,c) = \log \frac{P(a,b,c)}{P(a)P(b)P(c)}$$

Results from the PTT Gossiping Board

- 台灣 沒有 新聞 問卦 中國 八卦 什麼 有沒 就是 可以 一個 總統 的八 不是 自己 馬習 兩岸 習會 現在 因為 馬英 英九 我們 完整 還是 知道 整新 大家 如果 真的 不會 所以 這樣 美國 表示 記者 怎麼

- 有沒有 的八卦 馬習會 馬英九 完整新 國民黨 新加坡 習近平 朱立倫 中華民 民進黨 華民國 蔡英文 領導人 體來源 聞連結 聞標題 的時候 或短網 短網址 聞內文 最經典 立法院 九二共 二共識 綜合報 年輕人 參選人 柯文哲 果日報 俄羅斯

- 中華民國 或短網址 九二共識 民共和國 歐陽娜娜 空軍一號 香格里拉 聖約翰科 岸同屬一 國台辦主 合外電報 北地檢署 格里拉飯 搓圓仔湯 馬來西亞 每日郵報

- 香格里拉飯 打斷骨頭連 斷骨頭連著 夢由藝文工 骨頭連著筋 次丟鞋會白 丟鞋會白嗎 頭連著筋的 長澤茉里奈 係與時俱進 象山或貓空 殺毛孩自食 涉殺毛孩自 淫與賣臺的 捕野狗繁殖 女涉殺毛孩 生数学竞赛

Filtering

- Removing the overlaps
- Prefer the longer terms

Results from the PTT Gossiping Board

- 台灣 新聞 問卦 沒有 什麼 總統 可以 中國 自己 兩岸 一個 知道 因為 表示 如果 就是 記者 媒體 現在 我們 怎麼 覺得 應該 所以 今天 這樣 問題 報導 大家 美國

- 有沒有 的八卦 馬習會 馬英九 國民黨 新加坡 習近平 朱立倫 民進黨 蔡英文 領導人 的時候 最經典 立法院 年輕人 參選人 柯文哲 俄羅斯 越來越 王金平

- 中華民國 或短網址 九二共識 歐陽娜娜 空軍一號 香格里拉 北地檢署 搓圓仔湯 馬來西亞 每日郵報 從頭到尾 血濃於水 伊斯蘭國 西奈半島 翁山蘇姬 有期徒刑 維基百科 莫名其妙 水到渠成 廣電三法 便利商店 逢中必反 保外就醫 斷章取義 沸沸揚揚 她賣老嗆 三環三線

- 長澤茉里奈 捕野狗繁殖 金箔片皮豬 竹葉東星斑 睜眼說瞎話 荷葉邊系服 窺探者憲章 女流本因坊 摩爾曼斯克 真紅眼黑龍 萬般皆下品 沙姆沙伊赫 溫良恭儉讓 冰與火之歌 休灰了志氣

Shorter Period (One Week)

- 新聞 台灣 沒有 問卦 什麼 可以 自己 知道 因為 總統 表示 一個 如果 就是 怎麼 現在 應該 中國 覺得 所以 這樣 報導 媒體

- 有沒有 的八卦 國民黨 朱立倫 民進黨 蔡英文 馬英九 的時候 洪秀柱 馬習會 柯文哲 新加坡 黨主席 年輕人 發信站 立法院 公務員 習近平 越來越 宋楚瑜

- 媒體來源 或短網址 中華民國 莫名其妙 便利商店 柱下朱上 馬來西亞 停班停課 伊斯蘭國 有期徒刑 三商美邦 從頭到尾 無風無雨 亂七八糟 三不五時 搓圓仔湯

- 波多野結衣 生鮮或冷藏 得沸沸揚揚 亞歷塞維奇 佛羅里達州 睜眼說瞎話 帕金森氏症 溫良恭儉讓 阿囧津真矢 馬立連夢陸 面欲復擊之 耳熟能詳的

- 批踢踢實業坊 為面洋銃所擊 四谷赤坂麹町 安潔莉娜裘莉 哩來貢跨麥咧 唐伯虎點秋香 挖東牆補西牆 冠芽及匍匐莖 靠滋靠滋靠滋 防破片護目鏡 有枷具冥床斬 良禽擇木而棲 挖嘛嗯細條肛 喔咿呀喔咿呀

Other Application of Mutual Information

- Computing the associations between words and polarities

$$I(w, polarity) = log_2 \frac{P(w, polarity)}{P(w)P(polarity)}$$

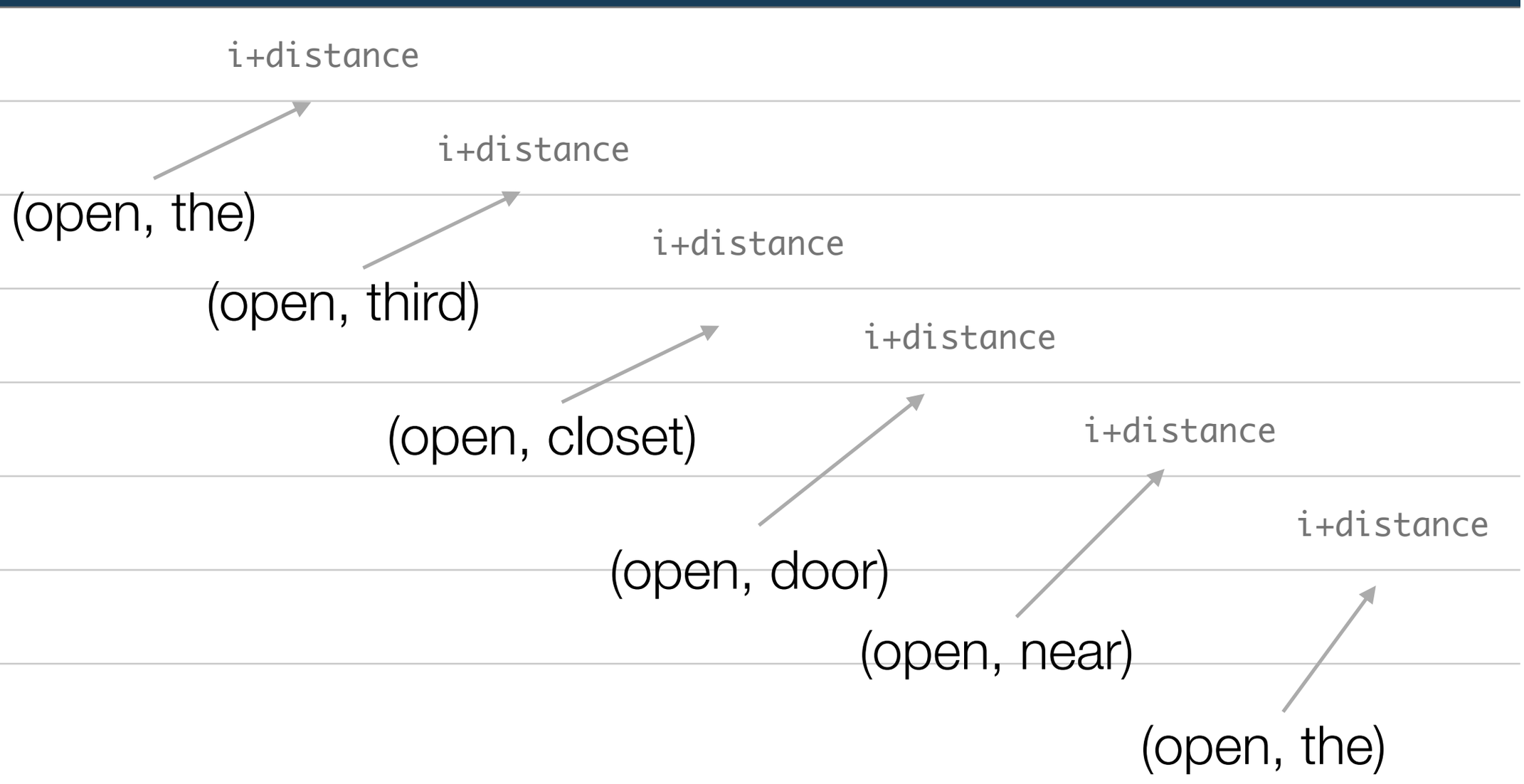
$$I(excellent, Positive) = log_2 \frac{P(excellent, Positive)}{P(excellent)P(Positive)}$$

$$I(excellent, Negative) = log_2 \frac{P(excellent, Negative)}{P(excellent)P(Negative)}$$

Distant Collocations

- open … door
  - open the door
  - open the black door
  - open the third closet door
  - open a bottle of wine and put on the table near the door
- The reasonable distance seems between 2 and 4.

Distance	to	open i	the	third	closet	door	near	the
1								
2								
3								
4								
5								
6								
7								
8								



Raw Results

Word 1	Word 2	Distance	Count
the	of	2	302
of	the	1	244
the	the	3	186
the	the	8	134
the	the	6	129
the	the	7	126
the	of	3	125
the	the	4	117
the	the	5	114
of	the	4	92
of	the	8	91
in	the	1	91

With Stopword Removing

Word 1	Word 2	Distance	Count
proletariat	increase	8	1
and	chagrin	8	1
can	despotic	8	1
hand	other	8	1
free	it	8	1
concentrated	consequen	8	1
we	nations	8	1
of	reproduce	8	1
appropriation	increase	8	1
coming	half	8	1
requiring	lands	8	1
german	opposition	8	1

Removing One-Time Instances: Longest Collocations

Word 1	Word 2	Distance	Count
this	seriously	8	2
only	is	8	2
with	what	8	2
germany	immediatel	8	2
to	petty	8	3
lose	to	8	2
ideas	ideas	8	2
at	property	8	2
and	movements	8	2
an	each	8	2
and	phrases	8	2
disposal	the	8	2

Removing One-Time Instances: Nearest Collocations

Word 1	Word 2	Distance	Count
in	times	1	2
have	already	1	2
they	wrote	1	3
breaks	out	1	3
which	it	1	5
struggle	between	1	2
contact	with	1	2
political	supremacy	1	3
result	from	1	2
bare	existence	1	2
need	to	1	2
family	relations	1	2

Removing One-Time Instances: Collocations with Middle Distance

Word 1	Word 2	Distance	Count
became	of	4.5	2
bourgeois	but	4.5	6
become	and	4.5	4
can	itself	4.5	2
taking	the	4.5	2
preaching	to	4.5	2
and	away	4.5	4
develops	they	4.5	2
has	exploitation	4.5	2
cultivation	of	4.5	4
superseded	of	4.5	2
pauper	and	4.5	2

Filtering with Offset Deviation

- To measure how often the individual offsets **deviate** (偏差) from the **mean** (平均).
  - mean: the average distance
  - deviation: measures how much the individual distance deviate from the mean.

$$s^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}$$

n: number of all occurrences of a word pair.
d_i: the distance of a single word pair instance.
$\bar{d}$: the mean of all d_i

Distant Collocations with Highest Deviations

Word 1	Word 2	Mean Distance	Deviation	Count
what	its	4.5	4.949747	2
subjection	of	4.5	4.949747	2
yearnings	of	4.5	4.949747	2
ones	that	4.5	4.949747	2
in	land	4.5	4.949747	2
way	been	4.5	4.949747	2
dangerous	class	4.5	4.949747	2
of	chemistry	4.5	4.949747	2
political	conditions	4.5	4.949747	2
consciousness	of	4.5	4.949747	2
epochs	of	4.5	4.949747	2
fight	for	4.5	4.949747	2

Distant Collocations with Lowest Deviations

Word 1	Word 2	Mean Distance	Deviation	Count
but	will	2	0	2
result	from	1	0	2
bare	existence	1	0	2
can	capital	6	0	2
our	relations	8	0	2
the	centralizati	7	0	2
need	to	1	0	2
family	relations	1	0	2
revolutionary	against	2	0	2
chiefly	to	1	0	2
be	effected	1	0	2
is	yet	2	0	2

More Frequent Distant Collocations with Lowest Deviations

Word 1	Word 2	Mean Distance	Deviation	Count
be	and	3.833333	1.466804	12
have	of	5.869565	1.455533	23
to	be	1.416667	1.442120	24
the	communism	4.454545	1.439697	11
of	can	5.000000	1.414214	11
in	but	5.000000	1.414214	11
for	a	2.076923	1.382120	13
for	class	5.363636	1.361817	11
as	and	5.153846	1.344504	13
has	of	5.730769	1.343360	26
it	has	1.409091	1.333063	22
working	class	1.360000	1.319091	25