

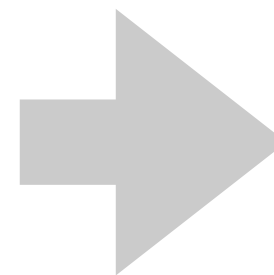
Programming for Social Scientists: Text Mining II

黃瀚萱

Last Week

- From unstructured sequence to meaningful symbols.

Hindu refers to any person who regards themselves as culturally, ethnically, or religiously adhering to aspects of Hinduism. [1][2] It has historically been used as a geographical, cultural, and later religious identifier for people indigenous to the Indian subcontinent.[3][4] The historical meaning of the term Hindu has evolved with time. Starting with the Persian and Greek references to the land of the Indus in the 1st millennium BCE through the texts of the medieval era,[5] the term Hindu implied a geographic, ethnic or cultural identifier for people living in the Indian subcontinent around or beyond the Sindhu (Indus) river.[6]

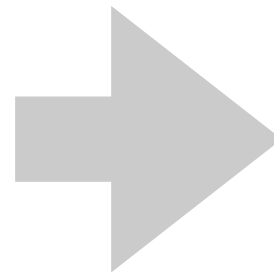


Word	Count
Hindu	11
term	6
century	6
Hindus	6
use	4
Indian	4
text	3
subcontinent	3
India	3
cultural	3
world	3
religious	3

Goal: Collocation Mining

- Mining more complex concepts from big data.

...Freeman and slave, patrician and plebeian, lord and serf, guildmaster(c) and journeyman, in a word, oppressor and oppressed, stood in constant opposition to one another, carried on an uninterrupted, now hidden, now open fight, that each time ended, either in the revolutionary reconstitution of society at large, or in the common ruin of the contending classes...

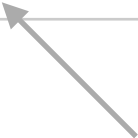


Word 1	Word 2
third	estate
constitution	adapted
corporate	guilds
laid	bare
eternal	truths
distinctive	feature
torn	asunder
radical	rupture
eighteenth	century
immense	majority
buying	disappears
absolute	monarchy

Background Knowledge

- Fundamental text processing
- Counting
- Simple statistics
 - Mean
 - Deviation
 - Chi-square
 - Probability

N-grams

<i>N</i>				
1	programming	for	social	scientists
2	programming for	for social	social scientists	
3	programming for social	for social scientists	 Collocation	
4	programming for social scienitst			

Approaches to Collocation Mining

- Frequency-based
- Mean and Variance
- Hypothesis Testing
- Mutual Information

Frequency-based Method

- Counting and sorting

Word 1	Word 2	Count
of	the	244
in	the	91
the	bourgeoisie	66
the	proletariat	50
to	the	43
by	the	40
for	the	38
of	production	38
with	the	34
the	bourgeois	33
conditions	of	29
means	of	25

Frequency-based Method with Stopword Removing

- Counting and sorting

Word 1	Word 2	Count
working	class	23
bourgeois	society	15
class	antagonisms	11
ruling	class	11
modern	industry	11
productive	forces	9
modern	bourgeois	8
middle	ages	7
private	property	7
bourgeois	property	7
class	struggle	6
old	society	6

Distant Collocations

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

Counting the Collocations with Distances

```
window_size = 9
```

```
word_pair_counts = Counter()
```

```
word_pair_distance_counts = Counter()
```

```
for i in range(len(tokens) - 1):
```

```
    for distance in range(1, window_size):
```

```
        if i + distance < len(tokens):
```

```
            w1 = tokens[i]
```

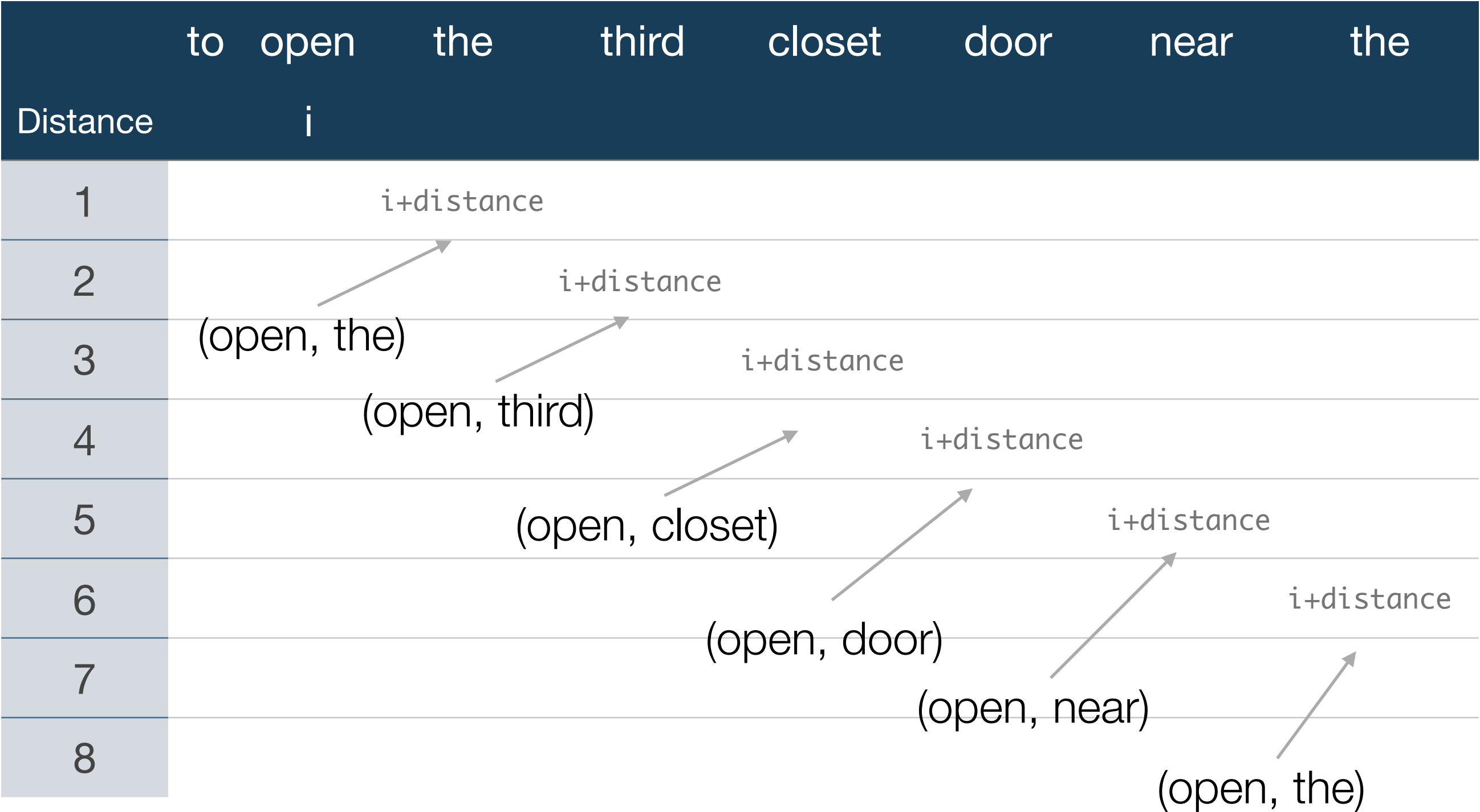
```
            w2 = tokens[i + distance]
```

```
            word_pair_distance_counts[(w1, w2, distance)] += 1
```

```
            word_pair_counts[(w1, w2)] += 1
```

```
for (w1, w2, distance), c in word_pair_distance_counts.most_common(20):
```

```
    print("%s\t%s\t%d\t%d" % (w1, w2, distance, c))
```



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 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	3.8333333	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

Hypothesis Test Method

- The t test
 - With an assumption that probabilities are approximately normally distributed.
- Pearson's chi-square test

Pearson's Chi-Square Test

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

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

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

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

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	χ^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	χ^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	χ^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	χ^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	χ^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

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

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

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

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

Assignments

- 與Mutual Information相比，Frequency based 的缺點是什麼？
- 處理 metamorphosis_franz_kafka.txt，找出三種 collocations
 - Frequency-based
 - Chi-square test
 - Mutual information

Bonus

- metamorphosis_franz_kafka.txt 裡有很多對話或自言自語，用雙引號區別。請只用括號裡的文字，建立 collocations

"Oh, God", he thought, "what a strenuous career it is that I've chosen! Travelling day in and day out. Doing business like this takes much more effort than doing your own business at home, and on top of that there's the curse of travelling, worries about making train connections, bad and irregular food, contact with different people all the time so that you can never get to know anyone or become friendly with them. It can all go to Hell!" He felt a slight itch up on his belly; pushed himself slowly up on his back towards the headboard so that he could lift his head better;