Chapter NLP:IX

IX. NLP Applications

- Frequency Extraction
- Keyword Extraction
- Cooccurrence Analysis
- Information Extraction
- Text Clustering
- Text Classification
- Machine Translation
- Text Generation
- Chat Bots
- Natural Language Understanding
- Machine Translation
- Text Generation
- Chat Bots
- Natural Language Understanding

Cooccurrence Analysis

Overview

Structuralist semantics [F. de Saussure]:

- **Syntagmatic relation:** Signifiers which occur conjointly complement w.r.t function and content
- **Paradigmatic relation:** Signifiers which occur in similar contexts have similar function w.r.t. grammar and content $\rightarrow$ cp. distributional hypothesis

Computing Cooccurrences

- **Local context $\mathrm{C}(\mathbf{w})$:** Set of words that occur in the same *window* as w
- **Global context $\mathrm{G}(\mathbf{w})$:** set of words which occur conjointly with w in a statistically significant manner
- **windows:** Sentences, Paragraphs, Documents, Headlines, k left/right neighbor words

	Paradigma 1	**Paradigma 2**	**Paradigma 3**	**Paradigma 4**	**Paradigma 5**	**Paradigma 6**
Syntagma 1	Der	Hund	läuft	die	Straße	hinab
Syntagma 2	Ein	Dackel	rennt	einen	Weg	hinauf
Syntagma 3	Ein	Sittich	läuft	die	Bäume	hinauf
Syntagma 4	Der	Wal	rennt	die	Schienen	hinauf
Syntagma 5	Er	–	rennt	die	Wand	hinab

Cooccurrence Analysis

Example

“The sun is shining” $\rightarrow C_{sentence}(sun) = \{\texttt{The, is, shining}\}$

“The sun is burning” $\rightarrow C_{sentence}(sun) = \{\texttt{The, is, burning}\}$

“The light is shining” $\rightarrow C_{sentence}(light) = \{\texttt{The, is, shining}\}$

$G(sun)$ = $\{\texttt{The, is, shining, burning}\}$

$$\mathbf{G}(\mathbf{sun}) \sim \mathbf{G}(\mathbf{light})$$

Cooccurrence Analysis

Methodology

Counting co-occurrence

- Focus on high frequent events in text data (Zipf's law!)
- Maximal frequency pair: "the – of"

Statistical significancy

- Measure of deviation from random conjoint occurrence

measurements (bag of words within windows)

- n_A – windows w containing **type A**
- n_B – windows w containing **type B**
- n_{AB} – windows w containing **type A and B**
- n – number of all windows w

Determine significance of co-occurrence

- statistical test measuring "surprise"
- Captures semantic characteristics of a text
- **There is not the single measure**

Significance measures

- Frequency *(remember Zipf!)*
- Sørenson Dice Coefficient (Set based similarity of two samples)
- Pointwise Mutual Information (PMI)
- Log Likelihood (LL)
- Poisson Significance

Cooccurrence Analysis

From DTM to TTM

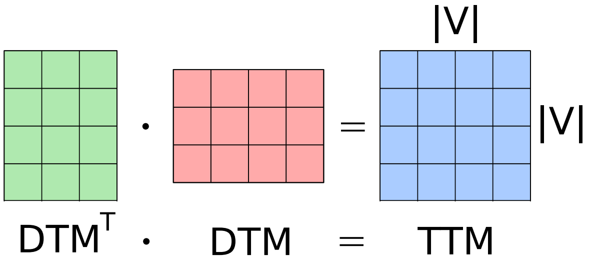
We need to contstrunct a Term-Term Matrix from a DTM in order to represent the counts n_{AB}

- ❑ n_A – In a sentence based DTM (e.g. a Sentence-Term Matrix) this is number of rows containing term A
- ❑ n_B – The number of rows containing term B
- ❑ n – number of all rows in a DTM
- ❑ n_{AB} – This could be easily calculated by the dot product DTM$^T \times$ DTM

- ❑ **The DTM needs to be weighted binary**
- ❑ In general each component of the Term-Term-Matrix (Context Matrix) can be calculated by:

$$n_{AB} = \sum_{k=1}^{|V|} \text{DTM}^T_{Ak}\text{DTM}_{kB}$$

- ❑ Example 3×4 DTM:



$$n_{AB} = \text{DTM}^T_{A1}\text{DTM}_{1B} + \text{DTM}^T_{A2}\text{DTM}_{2B} + \cdots + \text{DTM}^T_{A4}\text{DTM}_{4B}$$

Cooccurrence Analysis

Significance measures

- Frequency

$$sig_{baseline}(A, B) = n_{AB}$$

- Dice (Set based similarity of two samples)

$$sig_{dice}(A, B) = \frac{2 \cdot n_{AB}}{n_A + n_B}$$

- Pointwise Mutual Information (PMI)

$$sig_{MI}(A, B) = log\left(\frac{p(A, B)}{p(A) \cdot p(B)}\right) = log\left(\frac{n \cdot n_{AB}}{n_A \cdot n_B}\right)$$

Cooccurrence Analysis

Significance measures II

- Log Likelihood

$$sig_{LL} = \begin{cases} -\ \ 2log\lambda & \text{if } n_{AB} < \frac{n_A \cdot n_B}{n} \\ \ \ \ 2log\lambda & \text{else} \end{cases} \quad \text{with,}$$

$$\lambda = \left[\begin{array}{l} n \cdot log(n) - n_A \cdot log(n_A) - n_B \cdot log(n_B) + n_{AB} \cdot log(n_{AB}) \\ +(n - n_A - n_B + n_{AB}) \cdot log(n - n_A - n_B + n_{AB}) \\ +(n_A + n_{AB}) \cdot log(n_A - n_{AB}) + (n_B - n_{AB}) \cdot log(n_B - n_{AB}) \\ -(n - n_A) \cdot log(n - n_A) - (n - n_B) \cdot log(n - n_B) \end{array} \right]$$

- Poisson Significance

$$sig_{Poisson} = \frac{log(n_{AB}!) - n_{AB} \cdot log\left(\frac{n_A \cdot n_B}{n}\right) + \frac{n_A \cdot n_B}{n}}{log(n)}$$

Cooccurrence Analysis

Application Examples

(Change of) meaning may be inferred from cooccurrence results

Cooccurrence analysis → comparison of different result sets

- Change of context units (neighbours, sentence, document, . . .)
- Filter terms by POS-/NE-types
- Tracking change of global contexts by comparing time ranges

Visual Analytics

- Tables
- Graphs
- KIWC - Lists (Keyword in Context, Concordances)

Cooccurrence Analysis

Table Drawing

Significant Sentences based Cooccurrences for the term **“Coronavirus”** in Guardian Corpus 2020

	Freq-terms	Freq	MI-terms	MI	Dice-Terms	Dice	LL-Terms	LL	P-Terms	P
1	case	15708.00	people	23.83	case	0.18	case	23164.21	case	675.18
2	pandemic	11594.00	case	23.67	pandemic	0.16	pandemic	18680.36	pandemic	553.30
3	update	10613.00	health	23.10	relate	0.15	relate	18440.16	relate	546.74
4	people	10442.00	government	22.99	update	0.14	outbreak	18114.52	outbreak	533.99
5	relate	9948.00	pandemic	22.97	death	0.13	update	14312.98	update	427.85
6	health	9740.00	update	22.94	outbreak	0.13	death	13533.02	death	406.15
7	report	9549.00	report	22.89	report	0.12	spread	10693.76	coronavirus	403.80
8	death	9201.00	test	22.85	test	0.12	report	10188.16	spread	322.44
9	test	9142.00	death	22.58	health	0.12	crisis	9948.84	report	307.47
10	outbreak	7984.00	relate	22.53	crisis	0.10	test	9008.86	crisis	301.69
11	government	7242.00	country	22.44	spread	0.10	confirm	8551.58	test	272.77
12	uk	6799.00	week	22.44	uk	0.10	health	7920.38	confirm	259.85
13	crisis	6592.00	day	22.41	people	0.09	infection	6725.85	health	239.48
14	country	6493.00	uk	22.34	country	0.09	positive	6426.47	infection	206.35
15	number	6138.00	time	22.29	confirm	0.09	uk	5842.84	positive	196.93
16	spread	5934.00	number	22.23	number	0.09	toll	5631.01	uk	179.53
17	trump	5589.00	trump	22.20	infection	0.08	amid	5340.27	toll	170.80
18	week	5545.00	state	22.18	government	0.08	daily	5056.09	amid	161.84
19	lockdown	5475.00	work	22.01	lockdown	0.08	record	4959.31	daily	155.70
20	state	5264.00	lockdown	21.98	trump	0.08	number	4366.44	record	153.41

Cooccurrence Analysis

Table Drawing

Significant Sentences based Cooccurrences for the term **“Mask”** in Guardian Corpus 2020

	Freq-terms	Freq	MI-terms	MI	Dice-Terms	Dice	LL-Terms	LL	P-Terms	P
1	wear	8177.00	people	24.19	wear	0.53	wear	61997.19	wear	1775.86
2	face	5501.00	coronavirus	23.51	face	0.24	face	24752.99	face	765.19
3	people	2705.00	face	23.41	distance	0.10	glove	5088.22	distance	159.58
4	public	1899.00	wear	23.05	protective	0.08	distance	4846.58	glove	158.11
5	coronavirus	1685.00	health	22.80	glove	0.08	mandatory	4137.23	mandatory	133.85
6	distance	1655.00	public	22.70	mandatory	0.08	protective	4077.31	protective	133.76
7	health	1298.00	government	22.56	public	0.07	surgical	3342.40	surgical	100.85
8	social	1111.00	trump	22.27	hand	0.06	public	3036.51	public	99.44
9	trump	1081.00	time	22.24	social	0.05	compulsory	2392.82	compulsory	76.88
10	hand	947.00	case	22.24	medical	0.05	hand	2295.17	hand	76.78
11	protective	915.00	make	22.13	surgical	0.05	wash	2020.56	wash	66.96
12	make	900.00	week	22.10	wash	0.05	alcohol-based	1807.81	recommend	56.96
13	government	848.00	update	22.01	recommend	0.05	gown	1730.87	gown	55.89
14	include	829.00	state	21.97	require	0.05	recommend	1700.35	mandate	55.83
15	medical	810.00	day	21.95	compulsory	0.04	mandate	1700.25	social	52.95
16	mandatory	779.00	work	21.91	transport	0.04	flex	1598.22	alcohol-based	52.65
17	state	769.00	include	21.90	shop	0.04	social	1583.79	nose	50.64
18	update	756.00	report	21.90	cover	0.04	mouth	1540.75	mouth	50.63
19	glove	754.00	country	21.79	people	0.04	nose	1535.40	flex	48.76
20	spread	715.00	test	21.77	advice	0.04	rub	1470.91	covering	48.34

Cooccurrence Analysis

Graph Drawing

Cooccurrences = network structure → visualization as graph

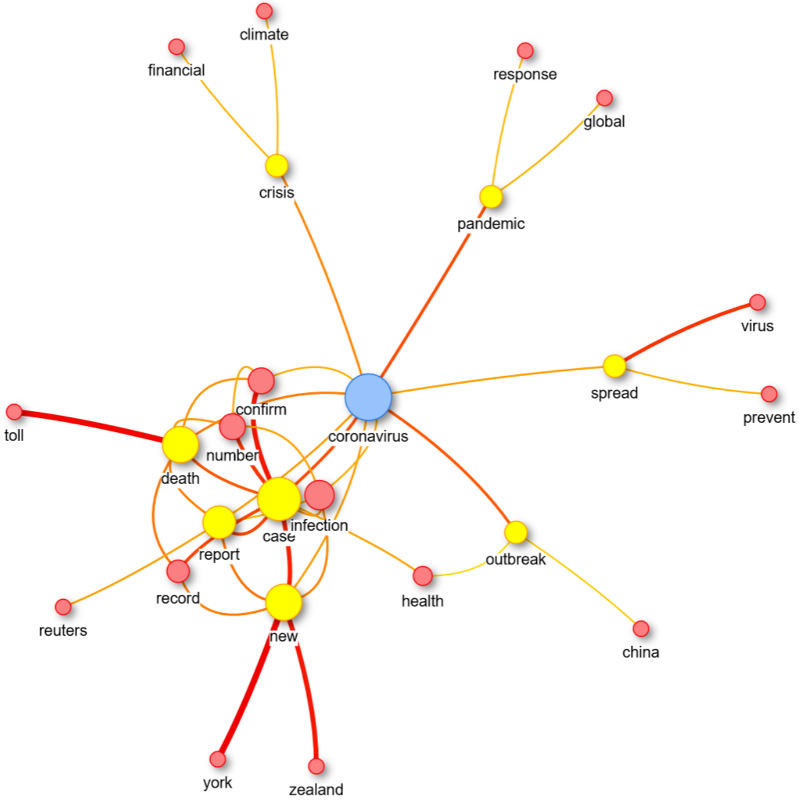
- ❑ Nodes: Terms
- ❑ Edges: Cooccurrence relation

e.g. additional information:

- ❑ Edge width: significancy value
- ❑ Node color: order of cooccurrence

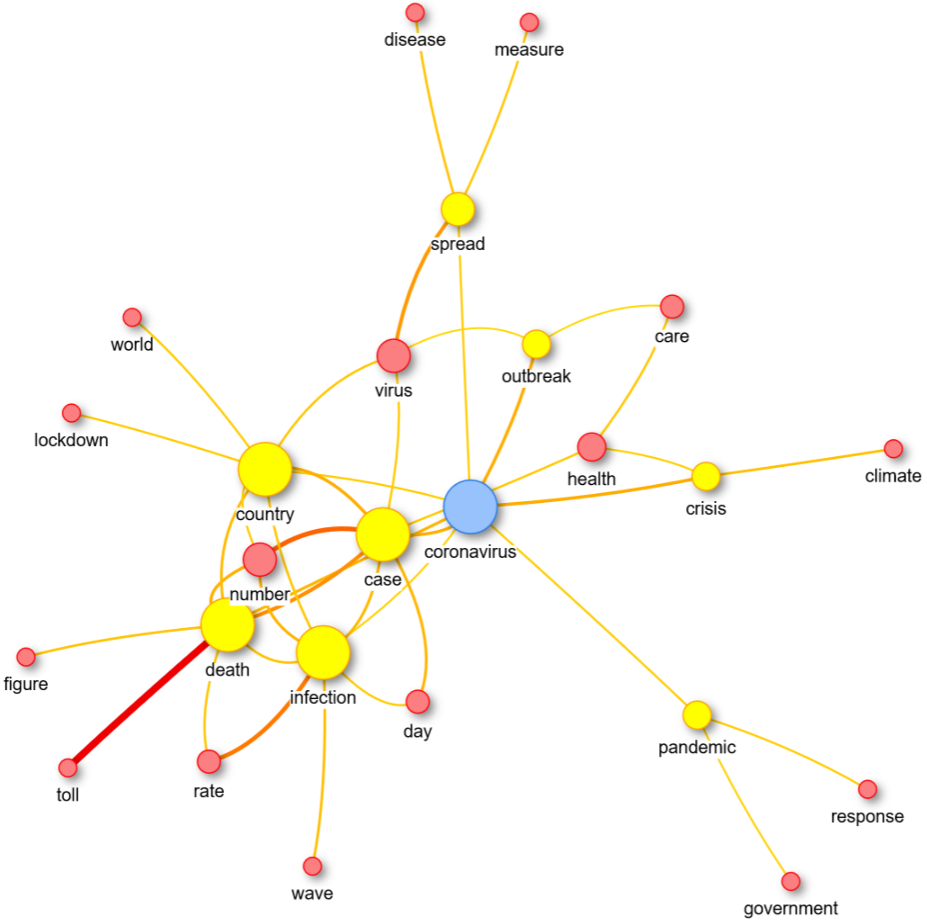
Caution:

- ❑ Algorithms for graph drawing (Force Directed Graphs) produce outcomes which are not necessarily semantically interpretable!



Cooccurrence Analysis

Cooccurrence study of the term “Coronavirus” based on nouns only



Cooccurrence Analysis

Remarks

Remarks:

- ❑ Those algorithms are called Spring Embedders and Force Directed Graph Drawing [Kobourov S.G. 2012]
  - "Force-directed algorithms are among the most flexible methods for calculating layouts of simple undirected graphs. Also known as spring embedders, such algorithms calculate the layout of a graph using only information contained within the structure of the graph itself, rather than relying on domain-specific knowledge. Graphs drawn with these algorithms tend to be aesthetically pleasing, exhibit symmetries, and tend to produce crossing-free layouts for planar graphs. In this survey we consider several classical algorithms, starting from Tutte's 1963 barycentric method, and including recent scalable multiscale methods for large and dynamic graphs."
- ❑ Cooccurrence Graphs tend to build small world networks. A small-world network is a type of mathematical graph in which most nodes are not neighbors of one another, but the neighbors of any given node are likely to be neighbors of each other and most nodes can be reached from every other node by a small number of hops or steps. [Wikipedia]
- ❑ Those algorithms are initialized randomly. In order to fixate the layout we must use seed values.

Cooccurrence Analysis

KWIC Drawing

KWIC-Lists: “Keyword in context” (Manning, Schütze,: "Foundations of Statistical Natural Language Processing", p. 35.)

- Selection of snippets by single keyword
- Centering display around key word

	pre	*keyword*	*post*
1	be caused by a	**coronavirus**	, the family of
2	Chinese of a novel	**coronavirus**	emerging once more from
3	with new strain of	**coronavirus**	. Broadcasters look likely
4	new strain of the	**coronavirus**	, marking the first
5	to Sydney because of	**coronavirus**	outbreak " I saw
6	be about the unfolding	**coronavirus**	but he begins by
7	, similar to the	**coronavirus**	. Top medics ,
8	viruses . Although the	**coronavirus**	recently discovered in Wuhan
9	don’t know where the	**coronavirus**	has come from -
10	Middle East plan and	**coronavirus**	. Climate emergency is
11	out of Wuhan in	**coronavirus**	evacuation A plane ,
12	the centre of the	**coronavirus**	outbreak , as officials
13	human-to-human transmission of the	**coronavirus**	in Europe , where
14	the spread of the	**coronavirus**	. Dean Smith lost
15	decision to frame the	**coronavirus**	on its front page
16	" inappropriately labelled the	**coronavirus**	by race " and
17	the source of the	**coronavirus**	being China . Mondays
18	. No backing for	**coronavirus**	claim The Daily Telegraph
19	fears that he had	**coronavirus**	" . The Daily
20	were afraid of the	**coronavirus**	is unsourced in the

Cooccurrence Analysis

What else?

Cooccurrence analysis:

- ❑ Global contexts → meaning of terms (“discourse level”)
- ❑ Significancy of cooccurrence relation is crucial

DH perspective: “visual hermeneutics” / distant reading of collections through graphical representations

Informational enrichment by creative filtering:

- ❑ Different sub collections
- ❑ Time ranges
- ❑ Person names / NE
- ❑ Certain POS-types
- ❑ . . .

What differences in results do you expect from different windows?

- ❑ Sentences – semantics, logic
- ❑ Paragraphs – discourse semantics

- ❑ Documents – topic semantics
- ❑ Headlines – framing?
- ❑ k left/right neighbour words – concordances, **collocation**