Language Analytics

Session 2: String Processing with Python

Ross Deans Kristensen-McLachlan
rdkm@cas.au.dk

Course outline

- 1. Introductions
- **2. String Processing with Python**
- 3. NLP for linguistic analysis
- 4. Text Classification 1
- 5. Text Classification 2
- 6. Word embeddings
- 7. Language modelling 1
- 8. Language modelling 2
- 9. BERT
- 10. More BERT
- 11. Project pitches
- 12. Generative models
- 13. Social impact

Break

1. What is a corpus?

1. What is a corpus?

- Following Hunston (2012: 2):

> *[…] a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts […]*
>
> *A corpus is planned, though chance may play a part in the text collection, and it is designed for some linguistic purpose. The specific purpose of the design determines the selection of texts […]*

1. What is a corpus?

If a corpus represents, very roughly and partially, a speaker's experience of language, the access software re-orders that experience so that it can be ***examined in ways that are usually impossible***

Hunston (2012: 3)

1. What is a corpus?

- Specialised
  - Texts of a particular type
  - ‘Representative’
- General
  - Texts of many diverse types
  - Unlikely to be ‘representative’
- Comparative
  - Two or more designed along the same lines
  - Enables direct comparison

- Parallel
  - Two or more in different languages
  - Texts from one translated into different languages
- Historical
  - Texts from different periods of time

1. What is a corpus?

- Why use corpora?

1. What is a corpus?

- Why use corpora?
  - Our intuition into how language is actually used is often quite poor
    - Using corpora allow us to make empirical evaluations from real-world data

1. What is a corpus?

- Why use corpora?
  - Our intuition into how language is actually used is often quite poor
    - Using corpora allow us to make empirical evaluations from real-world data
  - Exploring *the great unread* (Cohen 2009: 59)
    - We can manually study a handful of texts in detail
      - What about 10? 100? 10,000? More?

1. What is a corpus?

- Why use corpora?
  - Our intuition into how language is actually used is often quite poor
    - Using corpora allow us to make empirical evaluations from real-world data
  - Exploring *the great unread* (Cohen 2009: 59)
    - We can manually study a handful of texts in detail
      - What about 10? 100? 10,000? More?
- Text corpora allow us to make empirical, quantifiable observations of language in use that can easily scale and move between 'close' and 'distant' readings

Discussion

- Are there possible limitations of our working definition of a text corpus or text corpora?
- How might text corpora be relevant in your discipline(s)?
- Are there any specific corpora that interest you?
  - What type of corpora are they?

2. Exploring corpora

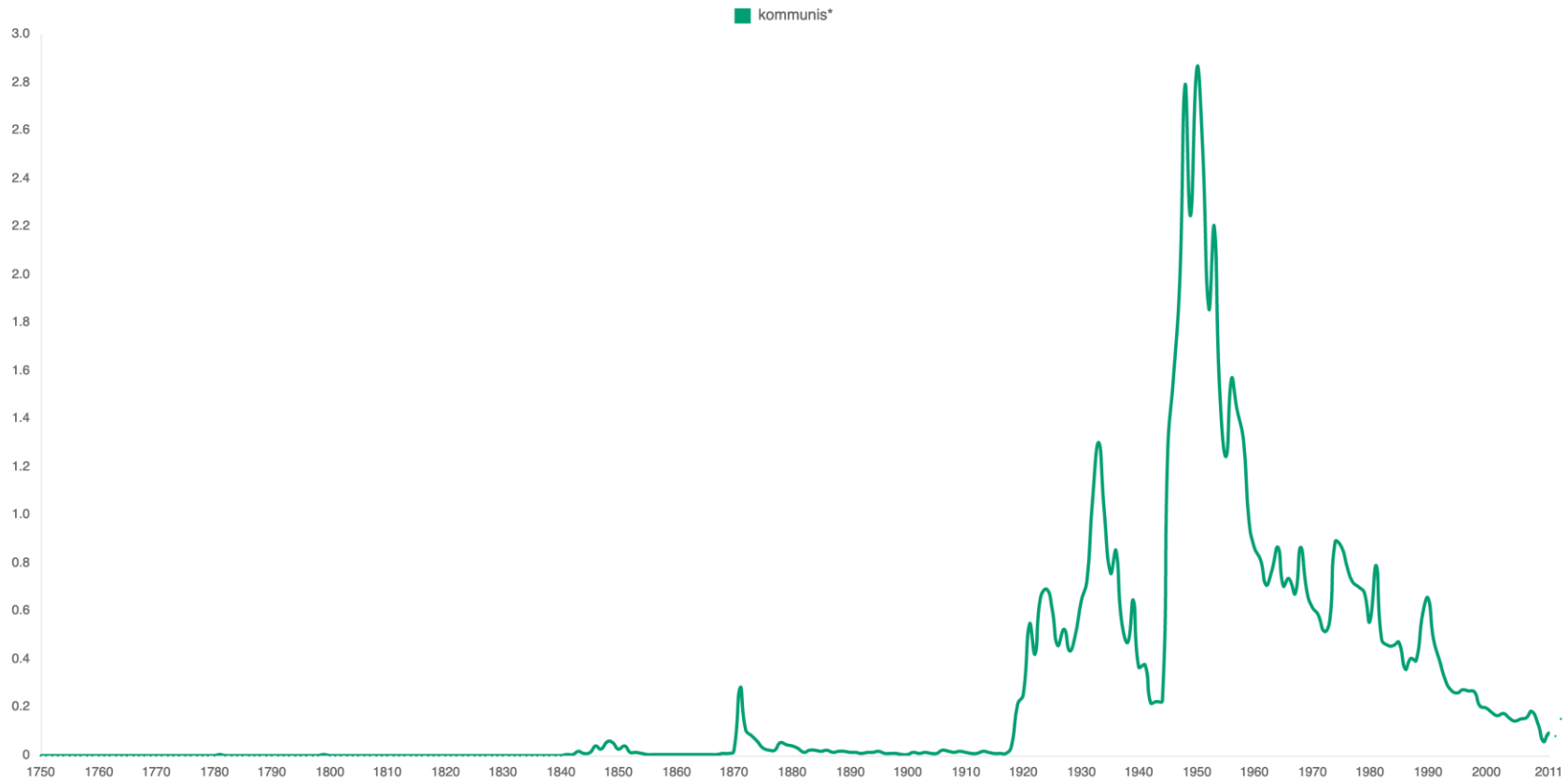
- Word frequencies
- Concordances, KWIC
- Collocation

2. Exploring corpora – word frequency

- Word frequency tells us…
  - how often an individual word appears
  - how often it appears relative to corpus size
  - if there are different distributions in different corpora
  - if distributions change over time

- Online examples
  - Google n-gram viewer
    - https://books.google.com/ngrams
  - Smurf
    - http://labs.statsbiblioteket.dk/smurf/
  - English corpora online
    https://www.english-corpora.org/

2. Exploring corpora – word frequency



*kommunistisk** in Danish newspapers over time

2. Exploring corpora – word frequency

SECTION	ALL	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
FREQ	41637	0	0	4	2	20	51	41	158	149	37	21	40	2727	2716	6326	12265	7251	4528	2813	2040
WORDS (M)	1589	5.0	7.1	11.6	28.1	30.4	33.0	34.2	37.1	60.0	51.2	64.7	79.8	71.7	95.2	94.8	121.0	152.0	163.3	183.7	177.1
PER MIL	26.19	0.00	0.00	0.34	0.07	0.66	1.55	1.20	4.26	2.48	0.72	0.32	0.50	38.04	28.53	66.70	101.39	47.69	27.72	15.31	11.52
SEE ALL SUB-SECTIONS AT ONCE																					

*communis** in British parliament over time

2. Exploring corpora - concordances

- Concordances (key words in context, KWIC)…

  […] bring together many instances of use of a word or phrase, allowing the user to observe regularities in use that tend to remain unobserved when the same words or phrases are met in their normal contexts. (Hunston 2012: 9)

2. Exploring corpora - concordances

- Concordances (key words in context, KWIC)…

  […] bring together many instances of use of a word or phrase, allowing the user to observe regularities in use that tend to remain unobserved when the same words or phrases are met in their normal contexts. (Hunston 2012: 9)

- The so-called *distributional hypothesis*
  - ‘You shall know a word by the company it keeps’ (Firth, J. R. 1957:11)
  - In short, the context in which a word is used determines meaning
  - Words used in similar contexts tend to have similar meanings

2. Exploring corpora – concordances

126	L-1966	Ormsby_Gore (L)	the solution I have outlined need not be alarming to	the	Communist	countries	If we had contrived to bring about a Europe in which
127	L-1980	----- (C)	Africans against the English-speaking Africans : In	the	Communist	countries one often finds	workers , employers and Governments
128	L-1968	----- (L)	that this is an internal matter to be settled within	the	Communist	empire : A crime	is no less a crime because it is
129	L-1964	Carington (L)	: It is possible , I suppose , that if	the	Communists	gained power in	Indonesia we should be worse off than we are
130	C-1927	Lansbury (C)	of this country I deny his right to talk about	the	Communist	Government of Russia in	particular and Communism in general as
131	C-1961	----- (C)	objectivity about him knows perfectly well that not one of	the	Communist	Governments and Eastern European	satellites would survive for
132	C-1948	Mott_R (L)	Germany were gradually , or suddenly , to fall into	the	Communist	grip . I have	always believed it to be a mistake to
133	C-1936	Price (N)	situation there , as I view it , is that	the	Communists	have always been	a small number in Spain , and are now
134	C-1926	Jones (C)	, except that she had been announced to speak for	the	Communists	in Caerphilly : If	that be true , what right bad the
135	L-1949	----- (L)	House that that is exactly what is being done by	the	Communists	in Czechoslovakia to the	historic Provinces of Bohemia and
136	L-1962	----- (C)	vicinity so many houses and families have been removed by	the	Communists	in order to make	it less easy for anybody to get through
137	C-1925	Hogg (C)	of the Communist International , and the resolutions of	the	Communist	International expressed enactment on	every member of every
138	C-1932	Eden (N)	the connection between the Soviet Government and	the	Communist	International : As	a result , my right hon: Friend requested the
139	L-1948	----- (L)	representatives of these two groups but , in fact ,	the	Communist	leaders retained control of	the whole resistance movement and
140	C-1976	Renton (L)	abandoning the country to the likes of Kenneth Gill ,	the	Communist	member of the TUC	General Council , who has already rejected the
141	C-1950	Smithers (L)	matter and when it is for the Government to crush	the	Communist	menace in the country	?
142	C-1959	Hinching... (C)	not Communist unity : We must so unite Germany that	the	Communist	order is liquidated and	the power of the working class is set
143	C-1923	Newbold (C)	offices of the General Federation of Trade Unions and of	the	Communist	party 27 were closed	after the removal of all papers and
144	C-1978	----- (L)	Party does not and has never accepted the ideas of	the	Communist	Party about so-called democratic	centralism , which has meant
145	C-1928	----- (C)	party in this country : There was no divergence between	the	Communist	party and the Labour	party on that:subject in those days : There
146	C-1988	Corbyn (C)	of their political beliefs , and the general secretaries of	the	Communist	party and the Workers '	party , Mr: Kutlu and Mr: Sargin
147	C-1951	Blackburn (L)	according to the evidence , was a secret member of	the	Communist	Party at the time	: I feel sure , knowing the generous
148	C-1986	Hamilton (C)	candidate was at the bottom of the poll , with	the	Communist	party candidate defeating	the Tory candidate in one case ? Will
149	C-1926	Jones (C)	to relate to the Communist party : I say that	the	Communist	party have every right	to place its point of view before the
150	C-1928	Saklatvala (C)	did he not see a publication in March last by	the	Communist	party in this country	, that the Communist organisation is an
151	C-1936	Guy (N)	his party on the point , but the attitude of	the	Communist	party 's certainly clear	: They would have no means test and
152	C-1939	McGovern (N)	party in fighting to win the masses for support of	the	Communist	party of Great Britain	and the Communist International-- ;
153	C-1987	Kaufman (C)	President of the United States and the General Secretary of	the	Communist	party of the Soviet	Union are this week in Washington -- ;
154	C-1938	Johnston (N)	in any shape or form representing the Communist party or	the	Communist	party 's views	, but these two postmen are citizens , they
155	C-1940	Gallacher (N)	See how it works downwards : Once you start with	the	Communist	party , the attack	is made in order to get at the
156	L-1966	----- (L)	should prefer it to be into the industrial operations of	the	Communist	Party , but it	may have to go wider than that :
157	C-1949	Nally (L)	catch Mr: Speaker 's eye together with a Member of	the	Communist	Party , and that	being so we-- ; -- ;
158	C-1944	----- (N)	he looks around at us , as if we were	the	Communist	Party :	
159	C-1947	Hynd (L)	other politicians and technical people but also members of	the	Communist	party-- ; I	understand that a unanimous vote of confidence was
160	L-1961	----- (C)	field is left vacant there is an open field for	the	Communist	philosophy of life	: 313 In considering the question of
161	C-1954	----- (C)	hanging on at Geneva , I do not believe that	the	Communist	Powers of either	Russia or China have shown a very great or

Concordance for *communis** in British parliament

3. Exploring corpora - collocations

- Collocation is…

  […] the statistical tendency of words to co-occur. A list of the collocates of a given word can yield similar information to that provided by concordance lines, with the difference that more information can be processed more accurately by the statistical operations of the computer than can be dealt with by the human observer. (Hunston 2012: 12)

3. Exploring corpora - collocations

- Collocation is…

  […] the statistical tendency of words to co-occur. A list of the collocates of a given word can yield similar information to that provided by concordance lines, with the difference that more information can be processed more accurately by the statistical operations of the computer than can be dealt with by the human observer. (Hunston 2012: 12)

- How likely are words to co-occur within a given window size?
  - E.g. ±5 words from our target word

3. Exploring corpora - collocations

- Collocation essentially models the following question:
  - Given the occurrence of a word *u*, how likely is it to also see word *v* within a given distance from *u?*
- One way to do this is to calculate the *strength of association* between two words
- We do this by comparing the *expected frequencies* with the *observed frequencies*
- For a more comprehensive overview of available metrics, see Evert (2010)
  - http://collocations.de

3. Exploring corpora - collocations

- One useful and easy to understand association measure is (*pointwise) mutual information* (Church & Hanks 1990)

- *Do two words co-occur more than expected if they were independent?*

$$PMI\left(word_1, word_2\right) = \log_2\left(\frac{p\left(word_1 \,\&\, word_2\right)}{p\left(word_1\right)p\left(word_2\right)}\right)$$

- In ordinary language…
  - P($word_1$) = probability of $word_1$ appearing in corpus
  - P($word_2$) = probability of $word_2$ appearing in corpus
  - P($word_1$ & $word_2$) = probability of $word_1$,$word_2$ appearing together in given window size

An example: w_1= hat; w_2=rolled; window =±5

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat.

A man must not be precipitate, or he runs over it ; he must not rush into the opposite extreme, or he loses it altogether. [. . .]

There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it.

The wind puffed, and Mr. Pickwick puffed, and the hat rolled over and over as merrily as a lively porpoise in a strong tide ; and on it might have rolled, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

(adapted from Evert & Krenn (2003))

An example: w_1= hat; w_2=rolled; window =±5

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a **hat**.

A man must not be precipitate, or he runs over it ; he must not rush into the opposite extreme, or he loses it altogether [. . .]

There was a fine gentle wind, and Mr. Pickwick's **hat** **rolled** sportively before it.

The wind puffed, and Mr. Pickwick puffed, and the **hat** **rolled** over and over as merrily as a lively porpoise in a strong tide ; and on it might have **rolled**, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

(adapted from Evert & Krenn (2003))

3. Exploring corpora - collocations

HELP	?		FREQ		ALL	%	MI	
1	☐	PARTY	4185		702610	0.60	4.83	
2	☐	COUNTRIES	1789		441806	0.40	4.27	
3	☐	WORLD	1311		542904	0.24	3.53	
4	☐	CHINA	1214		65215	1.86	6.47	
5	☐	CHINESE	834		44795	1.86	6.47	
6	☐	SOVIET	702		86598	0.81	5.27	
7	☐	PROPAGANDA	685		30353	2.26	6.75	
8	☐	INTERNATIONAL	552		224102	0.25	3.56	
9	☐	RUSSIA	546		105374	0.52	4.63	
10	☐	COMMUNIST	478		23725	2.01	6.59	
11	☐	RUSSIAN	417		55264	0.75	5.17	
12	☐	BLOC	416		7406	5.62	8.07	
13	☐	PARTIES	415		238199	0.17	3.06	
14	☐	SPREAD	325		49652	0.65	4.97	
15	☐	AGGRESSION	317		18447	1.72	6.36	
16	☐	INFLUENCE	298		132912	0.22	3.42	
17	☐	FASCIST	265		4612	5.75	8.10	
18	☐	SOCIALIST	263		44898	0.59	4.80	
19	☐	COMMUNISM	253		8618	2.94	7.13	
20	☐	THREAT	252		63602	0.40	4.24	
21	☐	EASTERN	250		45746	0.55	4.70	
22	☐	SOCIALISM	249		15755	1.58	6.24	
23	☐	MOVEMENT	243		98345	0.25	3.56	
24	☐	LEADERS	240		65545	0.37	4.13	
25	☐	Ré	230		12304	1.87	6.48	
26	☐	COMMUNISTS	217		8042	2.70	7.01	
27	☐	INFILTRATION	209		1382	15.12	9.50	
28	☐	SOCIALISTS	197		8860	2.22	6.73	
29	☐	ACTIVITIES	194		105498	0.18	3.13	
30	☐	FASCISM	192		2260	8.50	8.66	

Discussion

- Given the methods we've so discussed so far:
  - How might each of these methods be useful for you?
    - Word frequency
    - Concordances
    - Collocation

Move over to UCloud now!

https://cloud.sdu.dk

References

- **Church, K.W. & Hanks, P. (1990).** 'Word association norms, mutual information, and lexicography', *Computational Linguistics,* 16(1), 22-29.

- **Cohen, M. (2009).** 'Narratology in the Archive of Literature', *Representations*, 108(1), 51-75.

- **Evert, S. (2010).** *Collocations.de – Computational Approaches to Collocation*. http://collocation.de [Accessed February 2022]

- **Evert, S. & Krenn, B. (2001).** 'Methods for the qualitative evaluation of lexical association measures', *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, 188-195.

  - **(2003).** 'Computational approaches to collocations', Introductory course at the *European Summer School on Logic, Language, and Information (ESSLLI 2003)*, Vienna.

- **Firth, J. R. (1957).** 'A synopsis of linguistic theory 1930-55', *Studies in Linguistic Analysis* (special volume of the Philological Society), 1-32. [Reprinted in: Palmer, F. R. (ed.) (1968). *Selected Papers of J. R. Firth 1952-59*, pages 168-205. Longmans, London.]

- * **Hunston, S. (2002).** *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press