Tools for quantitative text analysis

DARIAH-DE/CLARIAH-DE Workshop
Digital tools and methods for historical research

Mainz, 25.09.219
Demival Vasques Filho

DARIAH-DE
Digitale Forschungsinfrastruktur für die
Geistes- und Kulturwissenschaften

IEG
Leibniz-Institut für
Europäische Geschichte

Agenda

- Visualising textual data with Voyant Tools
  - Getting started
  - Exploring the tools
  - Preparing the corpus

- DigiVoy: analysing DARIAH TextGrid data with Voyant tools
  - Importing documents

- Brief introduction to DKProWrapper
  - More advanced text analysis

Goals

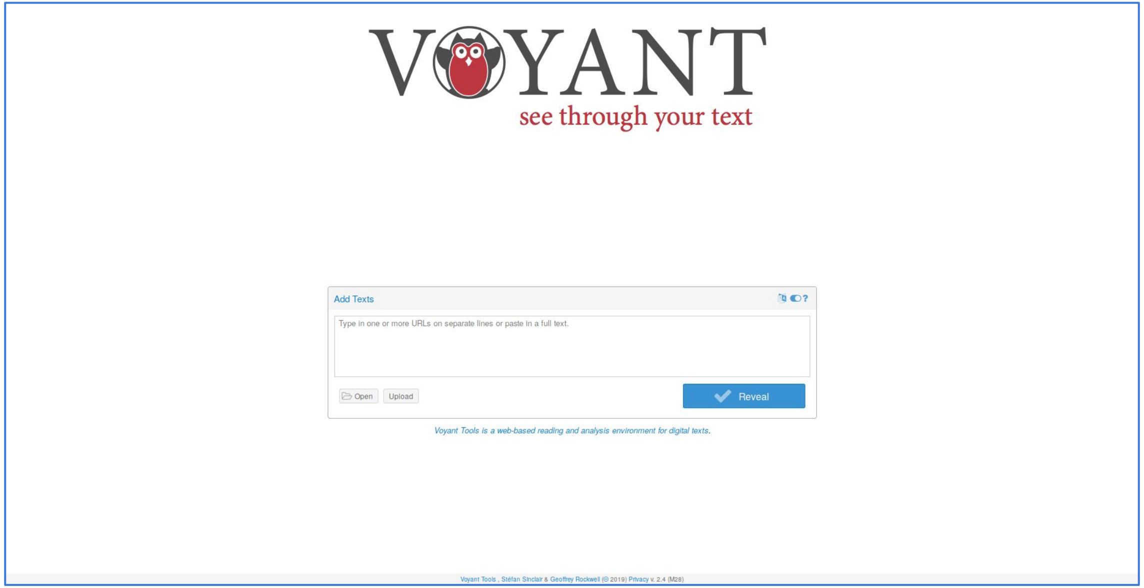
1. To explore the main tools of the Voyant suite. The number of functionalities is vast, and can be overwhelming.

2. To show how simple using the Voyant (text analysis) tools is. Hopefully, at the end of the workshop, you will feel confident and motivated enough to perform text analysis more frequently.

3. To clarify when it is time to look for more advanced tools.

Getting started...

Voyant tools is a web-based application: http://voyant-tools.org/



Type in one or more URLs on separate lines or paste in a full text.

Voyant Tools is a web-based reading and analysis environment for digital texts.

Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2019) Privacy v. 2.4 (M28)

Getting started… Sources input

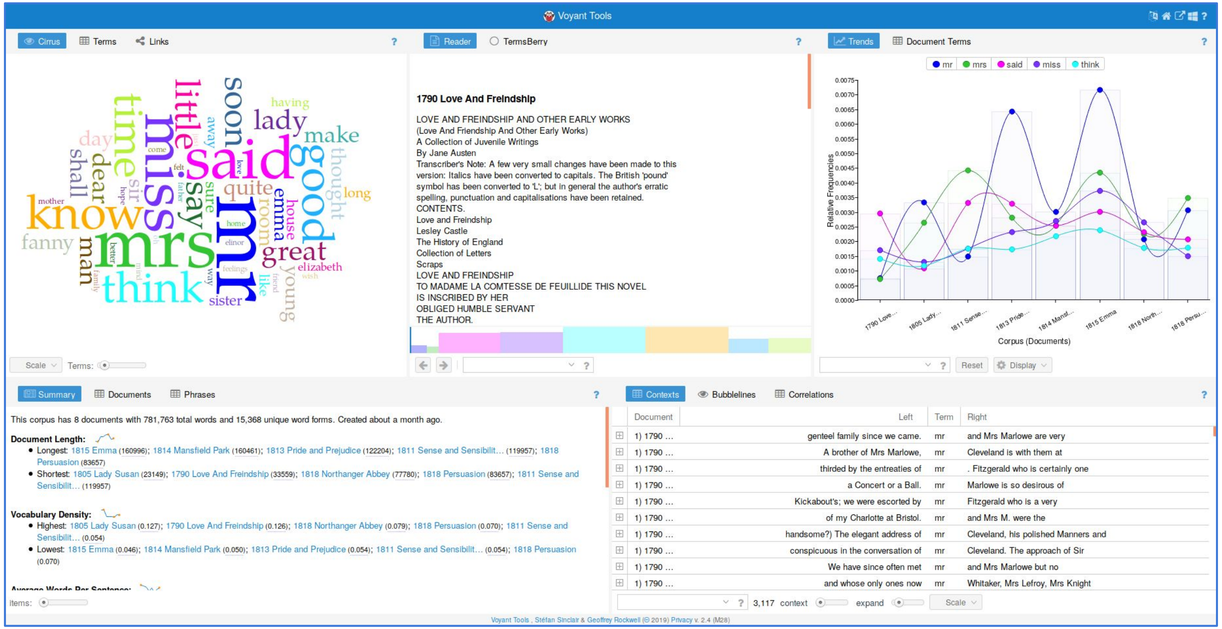
- Several options for inputting document sources. You can:
  - paste a large body of text in the initial box,
  - paste one (or several) URL (website address),
  - upload a single file (plain text, HTML, XML, MS Word, RTF, JSON and PDF),
  - upload multiple files (individually or zip archive),
  - open an available corpus (just Shakespeare’s plays and Austen’s novels).

PS: We can add, remove, or reorder documents after a corpus has been created (Document tool)

Getting started… Voyant skin

- The default Voyant skin has five tools: Cirrus, Reader, Trends, Summary, and Context.
- Tools can interact with one another (choosing a word from one tool, does so for another tool).
- With the upper right icon in each tool, we can:
  - export the information on the panel (icon with an arrow),
  - change tools in each panel (“window” icon),
  - define options for the current tool (“on-off” icon),
  - read what the tool is about (question mark icon).

Getting started… Default Voyant skin



1790 Love And Freindship

LOVE AND FREINDSHIP AND OTHER EARLY WORKS
(Love And Friendship And Other Early Works)
A Collection of Juvenile Writings
By Jane Austen
Transcriber's Note: A few very small changes have been made to this version: Italics have been converted to capitals. The British 'pound' symbol has been converted to 'L'; but in general the author's erratic spelling, punctuation and capitalisations have been retained.
CONTENTS.
Love and Freindship
Lesley Castle
The History of England
Collection of Letters
Scraps
LOVE AND FREINDSHIP
TO MADAME LA COMTESSE DE FEUILLIDE THIS NOVEL
IS INSCRIBED BY HER
OBLIGED HUMBLE SERVANT
THE AUTHOR.

This corpus has 8 documents with 781,763 total words and 15,368 unique word forms. Created about a month ago.

Document Length:

- Longest: 1815 Emma (160996); 1814 Mansfield Park (160461); 1813 Pride and Prejudice (122204); 1811 Sense and Sensibilit… (119957); 1818 Persuasion (83657)
- Shortest: 1805 Lady Susan (23149); 1790 Love And Freindship (33559); 1818 Northanger Abbey (77780); 1818 Persuasion (83657); 1811 Sense and Sensibilit… (119957)

Vocabulary Density:

- Highest: 1805 Lady Susan (0.127); 1790 Love And Freindship (0.126); 1818 Northanger Abbey (0.079); 1818 Persuasion (0.070); 1811 Sense and Sensibilit… (0.054)
- Lowest: 1815 Emma (0.046); 1814 Mansfield Park (0.050); 1813 Pride and Prejudice (0.054); 1811 Sense and Sensibilit… (0.054); 1818 Persuasion (0.070)

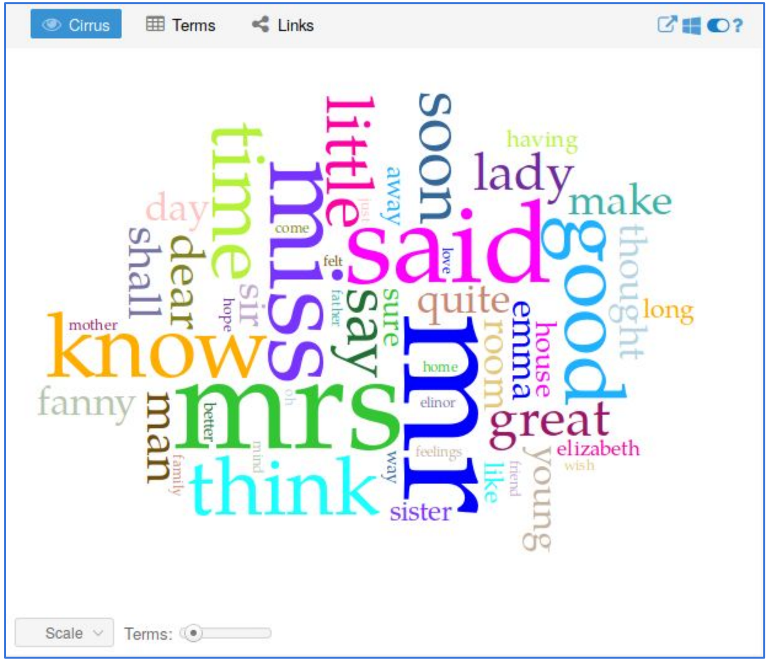
Average Words Per Sentence:

Document	Left	Term	Right
1) 1790 …	genteel family since we came.	mr	and Mrs Marlowe are very
1) 1790 …	A brother of Mrs Marlowe,	mr	Cleveland is with them at
1) 1790 …	thirded by the entreaties of	mr	. Fitzgerald who is certainly one
1) 1790 …	a Concert or a Ball.	mr	Marlowe is so desirous of
1) 1790 …	Kickabout's; we were escorted by	mr	Fitzgerald who is a very
1) 1790 …	of my Charlotte at Bristol.	mr	and Mrs M. were the
1) 1790 …	handsome?) The elegant address of	mr	Cleveland, his polished Manners and
1) 1790 …	conspicuous in the conversation of	mr	Cleveland. The approach of Sir
1) 1790 …	We have since often met	mr	and Mrs Marlowe but no
1) 1790 …	and whose only ones now	mr	Whitaker, Mrs Lefroy, Mrs Knight

3,117

Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2019) Privacy v. 2.4 (M28)

Exploring the tools… Cirrus



Cirrus:
- word cloud that visualizes the top frequency words of a corpus or document,
- most frequent are centrally positioned (but it will fill in empty spaces with small words),
- rovering over a word will show its frequency,
- clicking in a word will interact with other tools.

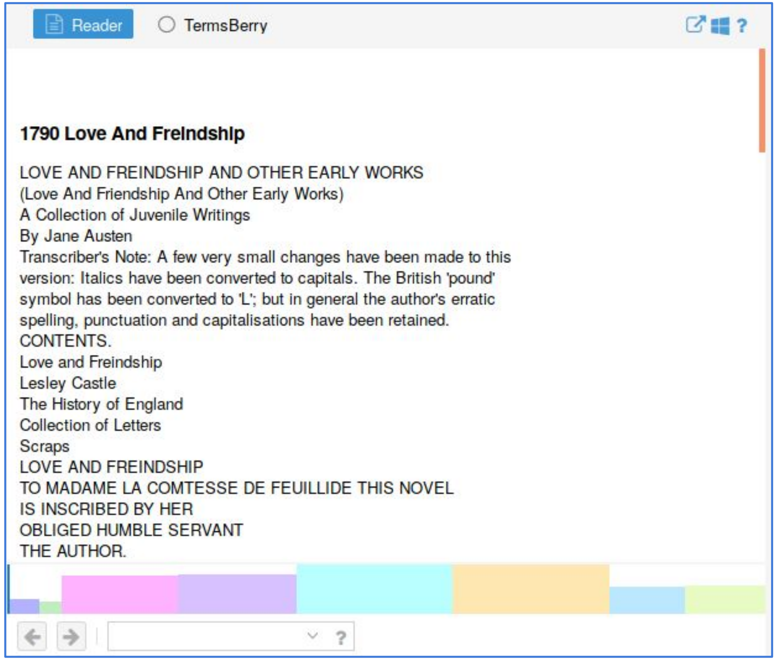
Options:

- stopwords,
- white list (set of allowed words – only these will appear),
- max terms,
- font family,
- palette.

Buttons:

- scale (corpus or specific document),
- terms (number of words appearing).

Exploring the tools… Reader



1790 Love And Freindship

LOVE AND FREINDSHIP AND OTHER EARLY WORKS
(Love And Friendship And Other Early Works)
A Collection of Juvenile Writings
By Jane Austen
Transcriber's Note: A few very small changes have been made to this version: Italics have been converted to capitals. The British 'pound' symbol has been converted to 'L'; but in general the author's erratic spelling, punctuation and capitalisations have been retained.
CONTENTS.
Love and Freindship
Lesley Castle
The History of England
Collection of Letters
Scraps
LOVE AND FREINDSHIP
TO MADAME LA COMTESSE DE FEUILLIDE THIS NOVEL
IS INSCRIBED BY HER
OBLIGED HUMBLE SERVANT
THE AUTHOR.

Reader:
- provides a way of reading documents in the corpus,
- shows a overview of the length of the documents in the corpus.

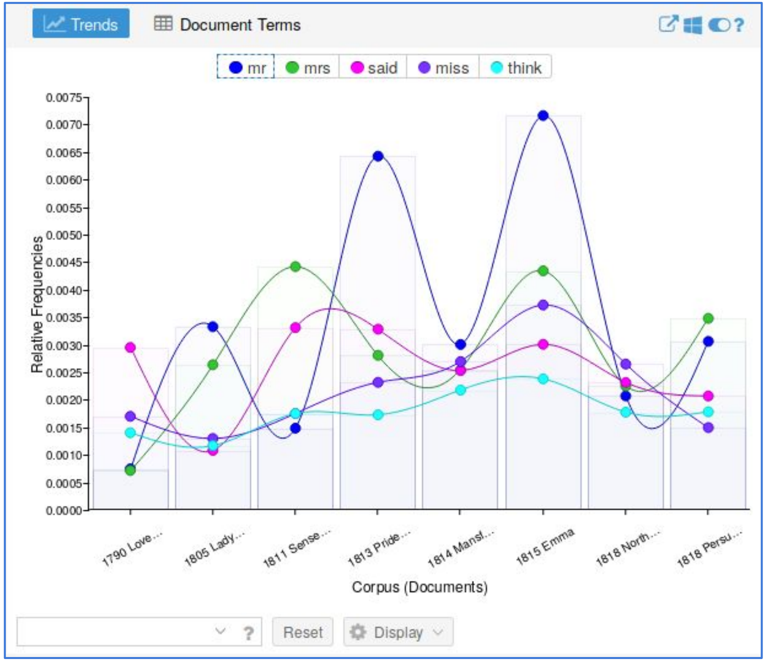
Options:

- no options.

Buttons:

- previous/next page (document),
- search box.

Exploring the tools… Trends



Trends:

- shows the distribution of words' occurrence across a corpus or document (document is split in segments).
- legend displays which words are in the graph,
- clicking in a word in the legend will show/hide that word in the graph,
- rovering over a point in the graph will call a box with information about that point,
- double-clicking on a dot will open two options:
    1) Terms: show the selected term for all documents.
    2) Documents: show all terms (in the legend) for the selected document.

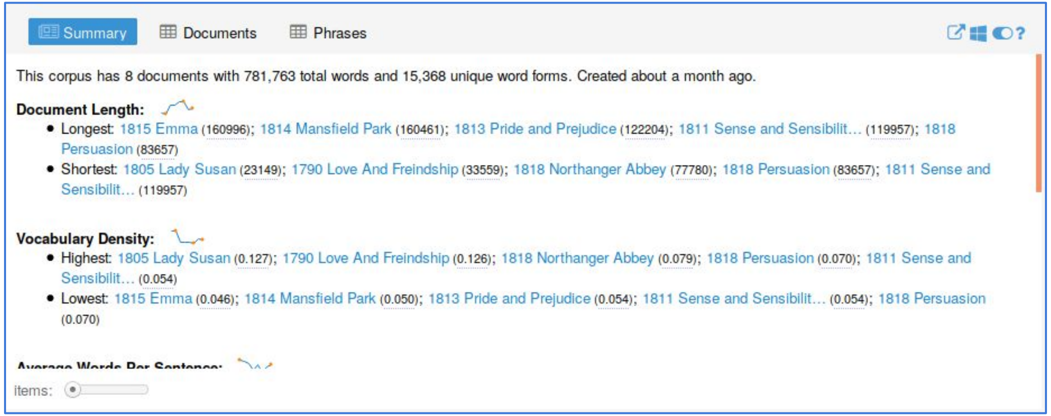
Options:

- stopwords,
- segments (number of segments into which number of documents are divided),
- frequencies (relative or absolute),
- palette.

Buttons:

- search box (can add more words to the graph too),
- reset (to initial configuration with five most frequent words).

Exploring the tools… Summary



This corpus has 8 documents with 781,763 total words and 15,368 unique word forms. Created about a month ago.

Document Length:

- Longest: 1815 Emma (160996); 1814 Mansfield Park (160461); 1813 Pride and Prejudice (122204); 1811 Sense and Sensibilit… (119957); 1818 Persuasion (83657)
- Shortest: 1805 Lady Susan (23149); 1790 Love And Freindship (33559); 1818 Northanger Abbey (77780); 1818 Persuasion (83657); 1811 Sense and Sensibilit… (119957)

Vocabulary Density:

- Highest: 1805 Lady Susan (0.127); 1790 Love And Freindship (0.126); 1818 Northanger Abbey (0.079); 1818 Persuasion (0.070); 1811 Sense and Sensibilit… (0.054)
- Lowest: 1815 Emma (0.046); 1814 Mansfield Park (0.050); 1813 Pride and Prejudice (0.054); 1811 Sense and Sensibilit… (0.054); 1818 Persuasion (0.070)

Average Words Per Sentence:

items:

Summary (some interesting statistics!):

- overview of the corpus with: number of words, number of unique words, longest and shortest documents, highest and lowest vocabulary density, average number of words per sentence, most frequent words, notable peaks in frequency, and distinctive words.

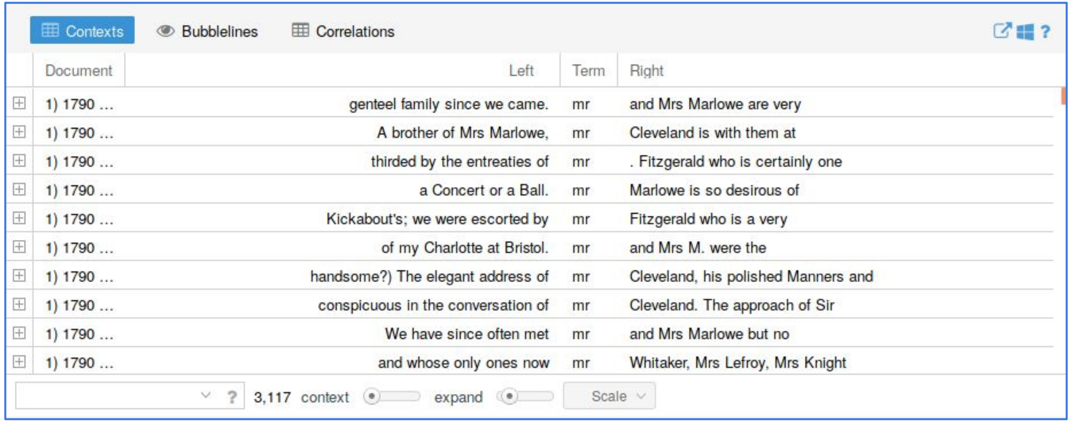
Options:

- stopwords.

Buttons:

- items (number of longest/shortest documents, highest/lowest vocabulary density, and so on).

Exploring the tools… Contexts



	Document	Left	Term	Right
⊞	1) 1790 …	genteel family since we came.	mr	and Mrs Marlowe are very
⊞	1) 1790 …	A brother of Mrs Marlowe,	mr	Cleveland is with them at
⊞	1) 1790 …	thirded by the entreaties of	mr	. Fitzgerald who is certainly one
⊞	1) 1790 …	a Concert or a Ball.	mr	Marlowe is so desirous of
⊞	1) 1790 …	Kickabout's; we were escorted by	mr	Fitzgerald who is a very
⊞	1) 1790 …	of my Charlotte at Bristol.	mr	and Mrs M. were the
⊞	1) 1790 …	handsome?) The elegant address of	mr	Cleveland, his polished Manners and
⊞	1) 1790 …	conspicuous in the conversation of	mr	Cleveland. The approach of Sir
⊞	1) 1790 …	We have since often met	mr	and Mrs Marlowe but no
⊞	1) 1790 …	and whose only ones now	mr	Whitaker, Mrs Lefroy, Mrs Knight

Contexts:

- shows each occurrence of a keyword with a bit of surrounding text (the context).

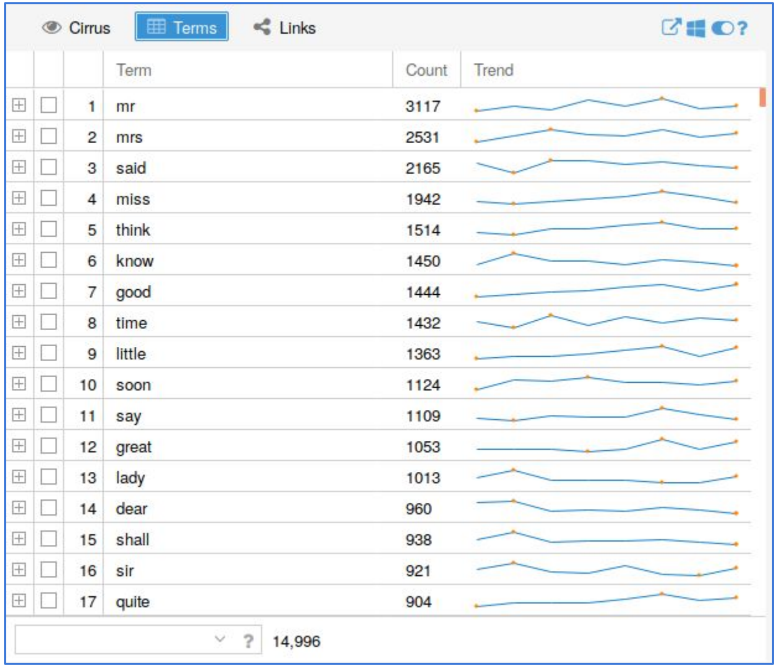
Options:

- no options.

Buttons:

- search box,
- context (how many words before and after the highlighted term),
- expand (to expand each row to that number of words).

Exploring the tools… Terms



			Term	Count	Trend
⊞	□	1	mr	3117	
⊞	□	2	mrs	2531	
⊞	□	3	said	2165	
⊞	□	4	miss	1942	
⊞	□	5	think	1514	
⊞	□	6	know	1450	
⊞	□	7	good	1444	
⊞	□	8	time	1432	
⊞	□	9	little	1363	
⊞	□	10	soon	1124	
⊞	□	11	say	1109	
⊞	□	12	great	1053	
⊞	□	13	lady	1013	
⊞	□	14	dear	960	
⊞	□	15	shall	938	
⊞	□	16	sir	921	
⊞	□	17	quite	904	

14,996

Terms:

- frequency of words (terms) in the entire corpus.

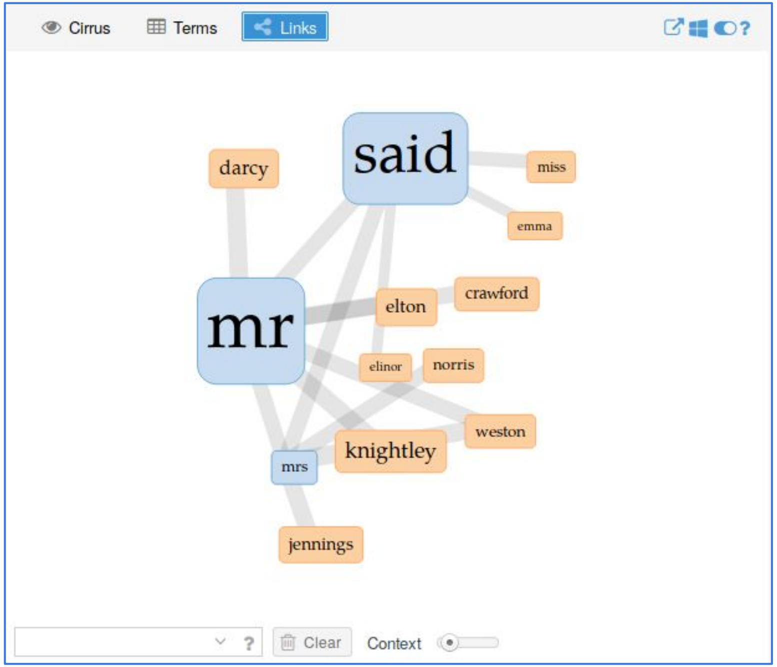
Options:

- stopwords,
- comparative (compare the relative frequency with another corpus).

Buttons:

- search box.

Exploring the tools… Links



Links:
- network of words that co-occurs in close proximity,
- color code: blue words are the main terms and orange words are the collocates.

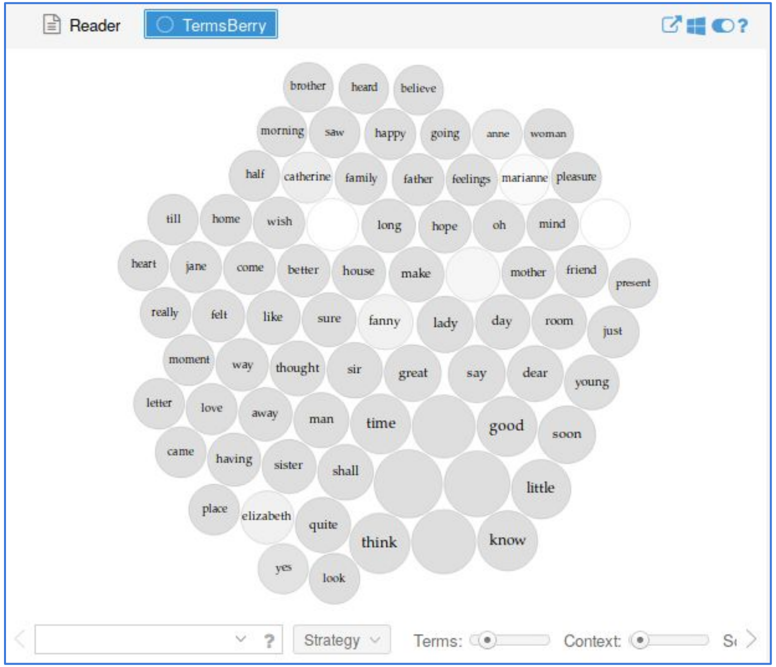
Options:

- stopwords,
- categories (it is possible to assign categories to words).

Buttons:

- search box,
- clear (to start from scratch),
- context (number of surrounding words).

Exploring the tools… TermsBerry



TermsBerry:

- high frequency terms and their collocates (words that occur in proximity),
- works like Cirrus plus the collocates of highly frequent terms,
- the darker the bubble, the larger the number of documents in which the term appears,
- when hovering over a term, the term becomes green and its collocates become pink. The darker the shade the more frequently the collocates appear in the context.

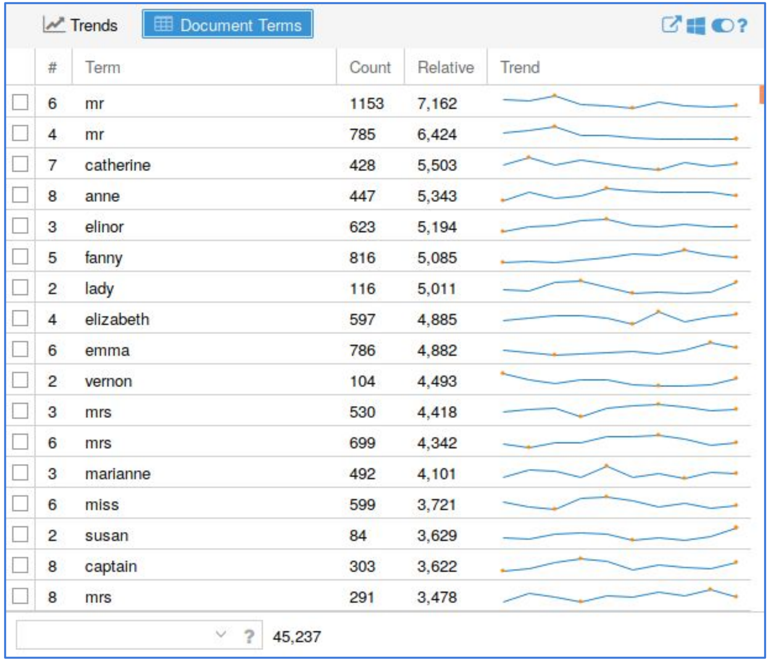
Options:

- stopwords,
- categories.

Buttons:

- search box,
- strategy. With two options:
  1) Top terms: highest frequencies in the corpus, or
  2) Distinct terms: distinct terms for each document.
- terms (number of “bubbles”),
- context,
- scaling (for adjusting the size of he bubbles according to the frequency of the terms).

Exploring the tools… Document terms



	#	Term	Count	Relative	Trend
□	6	mr	1153	7,162	
□	4	mr	785	6,424	
□	7	catherine	428	5,503	
□	8	anne	447	5,343	
□	3	elinor	623	5,194	
□	5	fanny	816	5,085	
□	2	lady	116	5,011	
□	4	elizabeth	597	4,885	
□	6	emma	786	4,882	
□	2	vernon	104	4,493	
□	3	mrs	530	4,418	
□	6	mrs	699	4,342	
□	3	marianne	492	4,101	
□	6	miss	599	3,721	
□	2	susan	84	3,629	
□	8	captain	303	3,622	
□	8	mrs	291	3,478	

45,237

Document terms:
- frequency of word for each document.

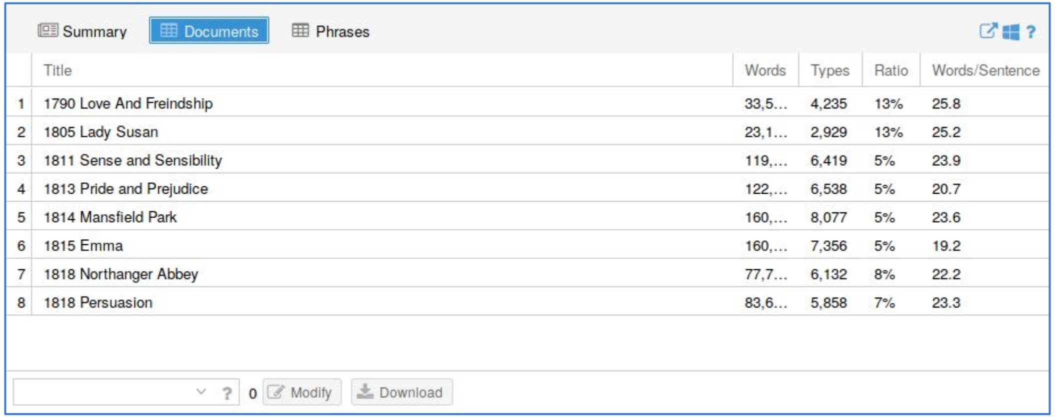
Options:

- stopwords,

Buttons:

- search box.

Exploring the tools… Documents



	Title	Words	Types	Ratio	Words/Sentence
1	1790 Love And Freindship	33,5…	4,235	13%	25.8
2	1805 Lady Susan	23,1…	2,929	13%	25.2
3	1811 Sense and Sensibility	119,…	6,419	5%	23.9
4	1813 Pride and Prejudice	122,…	6,538	5%	20.7
5	1814 Mansfield Park	160,…	8,077	5%	23.6
6	1815 Emma	160,…	7,356	5%	19.2
7	1818 Northanger Abbey	77,7…	6,132	8%	22.2
8	1818 Persuasion	83,6…	5,858	7%	23.3

Documents:

- show some information and basic statistics for every document of the corpus.

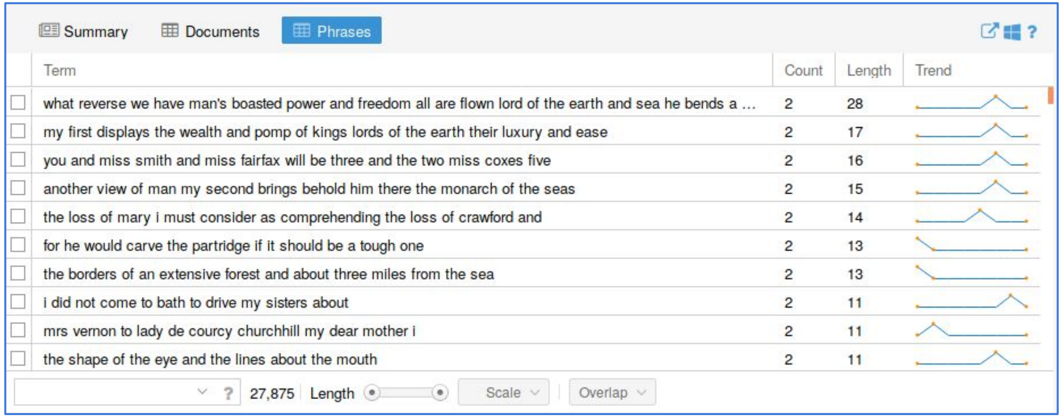
Options:

- no options.

Buttons:

- search box,
- modify (to add new documents to the corpus),
- download (the corpus).

Exploring the tools… Phrases



	Term	Count	Length	Trend
□	what reverse we have man's boasted power and freedom all are flown lord of the earth and sea he bends a …	2	28	
□	my first displays the wealth and pomp of kings lords of the earth their luxury and ease	2	17	
□	you and miss smith and miss fairfax will be three and the two miss coxes five	2	16	
□	another view of man my second brings behold him there the monarch of the seas	2	15	
□	the loss of mary i must consider as comprehending the loss of crawford and	2	14	
□	for he would carve the partridge if it should be a tough one	2	13	
□	the borders of an extensive forest and about three miles from the sea	2	13	
□	i did not come to bath to drive my sisters about	2	11	
□	mrs vernon to lady de courcy churchhill my dear mother i	2	11	
□	the shape of the eye and the lines about the mouth	2	11	

Phrases:

- shows repeating sequences of words organized by frequency of repetition or number of words in each repeated phrase.

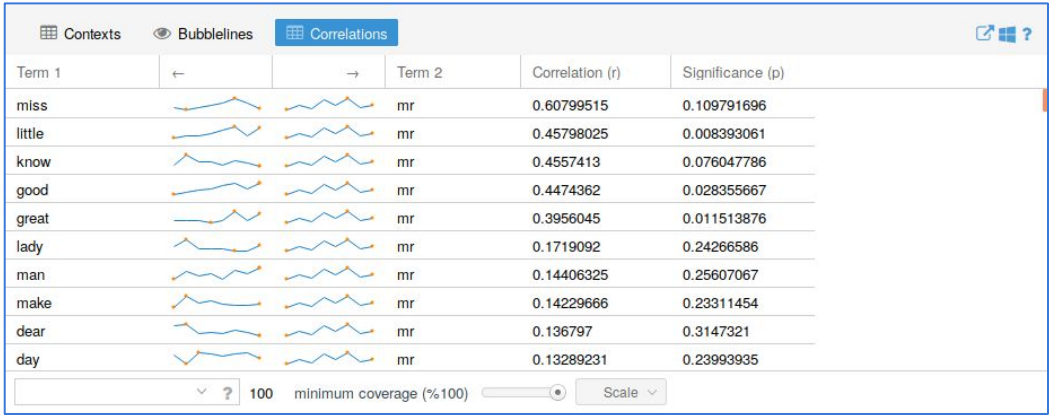
Options:

- no options,

Buttons:

- length (of the phrase),
- scale (corpus or documents),
- overlap (show all, prioritize by length, prioritize by frequency).

Exploring the tools… Correlation



Term 1	←	→	Term 2	Correlation (r)	Significance (p)
miss			mr	0.60799515	0.109791696
little			mr	0.45798025	0.008393061
know			mr	0.4557413	0.076047786
good			mr	0.4474362	0.028355667
great			mr	0.3956045	0.011513876
lady			mr	0.1719092	0.24266586
man			mr	0.14406325	0.25607067
make			mr	0.14229666	0.23311454
dear			mr	0.136797	0.3147321
day			mr	0.13289231	0.23993935

Correlation:

- terms whose frequencies rise and fall together or inversely.

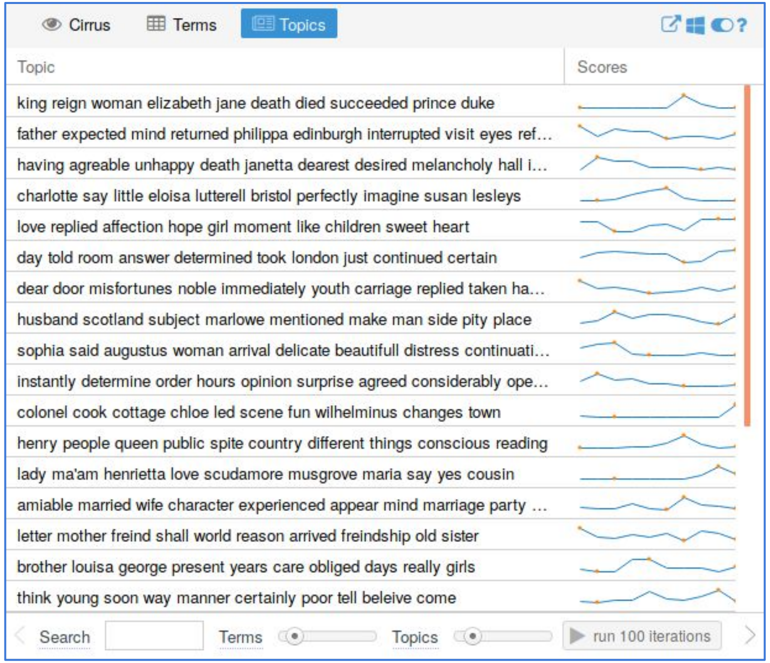
Options:

- no options.

Buttons:

- search box,
- minimum coverage (minimum relative frequency of the term in the documents. For instance, if a corpus has 10 documents and the minimum coverage is 20%, at least two of the documents must contain the term or it will be ignored),
- scale (corpus or documents).

Exploring the tools… Topics



Topic	Scores
king reign woman elizabeth jane death died succeeded prince duke	
father expected mind returned philippa edinburgh interrupted visit eyes ref…	
having agreable unhappy death janetta dearest desired melancholy hall i…	
charlotte say little eloisa lutterell bristol perfectly imagine susan lesleys	
love replied affection hope girl moment like children sweet heart	
day told room answer determined took london just continued certain	
dear door misfortunes noble immediately youth carriage replied taken ha…	
husband scotland subject marlowe mentioned make man side pity place	
sophia said augustus woman arrival delicate beautifull distress continuati…	
instantly determine order hours opinion surprise agreed considerably ope…	
colonel cook cottage chloe led scene fun wilhelminus changes town	
henry people queen public spite country different things conscious reading	
lady ma'am henrietta love scudamore musgrove maria say yes cousin	
amiable married wife character experienced appear mind marriage party …	
letter mother freind shall world reason arrived freindship old sister	
brother louisa george present years care obliged days really girls	
think young soon way manner certainly poor tell beleive come	

Topics:

- LDA topic modelling. Words form clusters according to their frequency and how they appear together in documents. Each row of the table is one of these clusters, that represent one specific topic.

Options:

- stopwords,
- terms per document (number of first words that are being used, e.g first 1000 words),
- iterations (number of times the algorithm reads the corpus to identify topics).

Buttons:

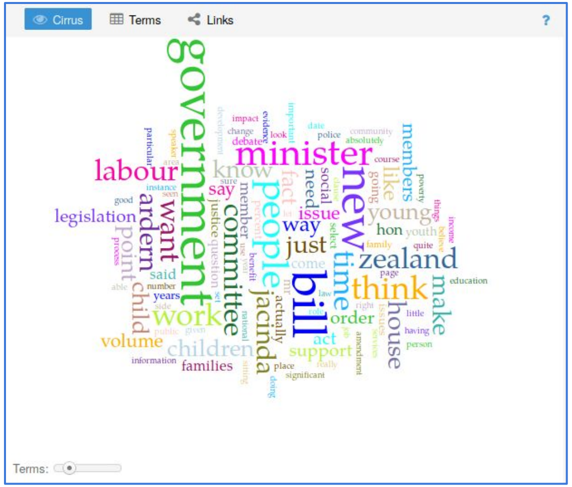
- search box,
- terms (number of keywords),
- topics (number of topics),
- run 100 iterations (to refine already defined topics).

Preparing the corpus…

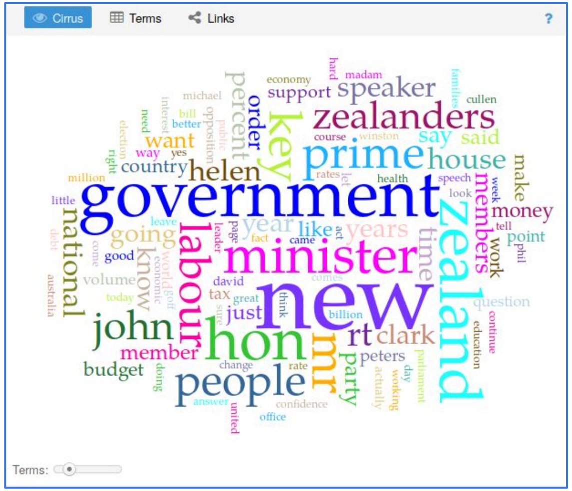
- With a good knowledge of what the tools can do, we can plan ahead how we expect to use our corpus, with questions like:
  - Am I only interested on how terms (and keywords) appear in the set of documents?
  - Do I want to track the differences the documents present over time?
  - Do I want to find different writing / authoring styles?
  - Perhaps, compare styles across different geographical locations?
  - Or to find pattern of distinct cultures?
  - And so on...

Preparing the corpus…

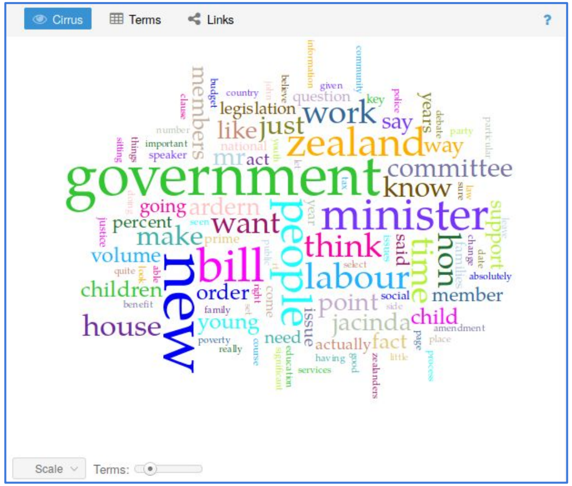
Speeches from 2008 to 2016 given by former NZ PM John Key and current PM Jacinda Ardern



Jacinda Ardern



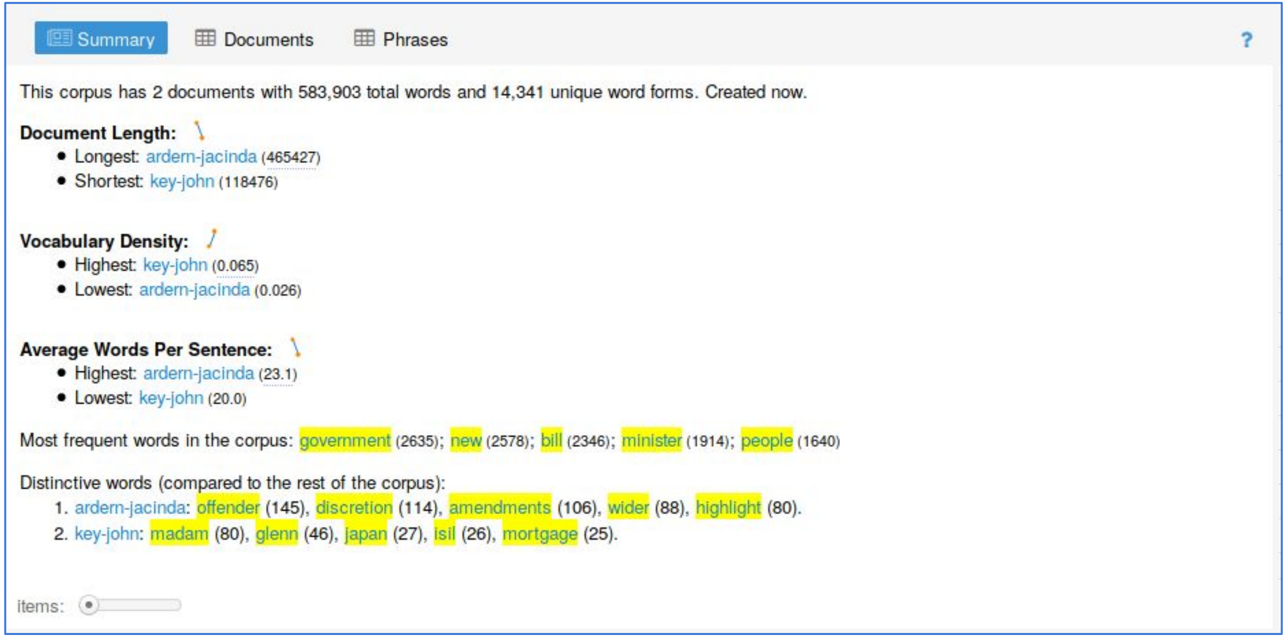
John Key



Both

Preparing the corpus… a few examples

- Differences in style



This corpus has 2 documents with 583,903 total words and 14,341 unique word forms. Created now.

Document Length:

- Longest: ardern-jacinda (465427)
- Shortest: key-john (118476)

Vocabulary Density:

- Highest: key-john (0.065)
- Lowest: ardern-jacinda (0.026)

Average Words Per Sentence:

- Highest: ardern-jacinda (23.1)
- Lowest: key-john (20.0)

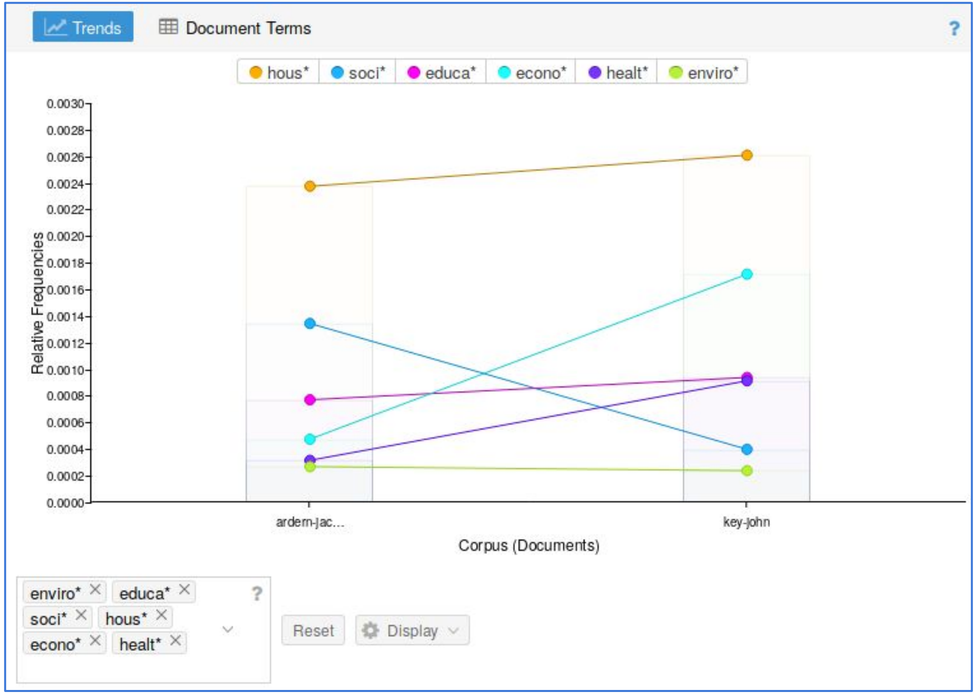
Most frequent words in the corpus: government (2635); new (2578); bill (2346); minister (1914); people (1640)

Distinctive words (compared to the rest of the corpus):

1. ardern-jacinda: offender (145), discretion (114), amendments (106), wider (88), highlight (80).
2. key-john: madam (80), glenn (46), japan (27), isil (26), mortgage (25).

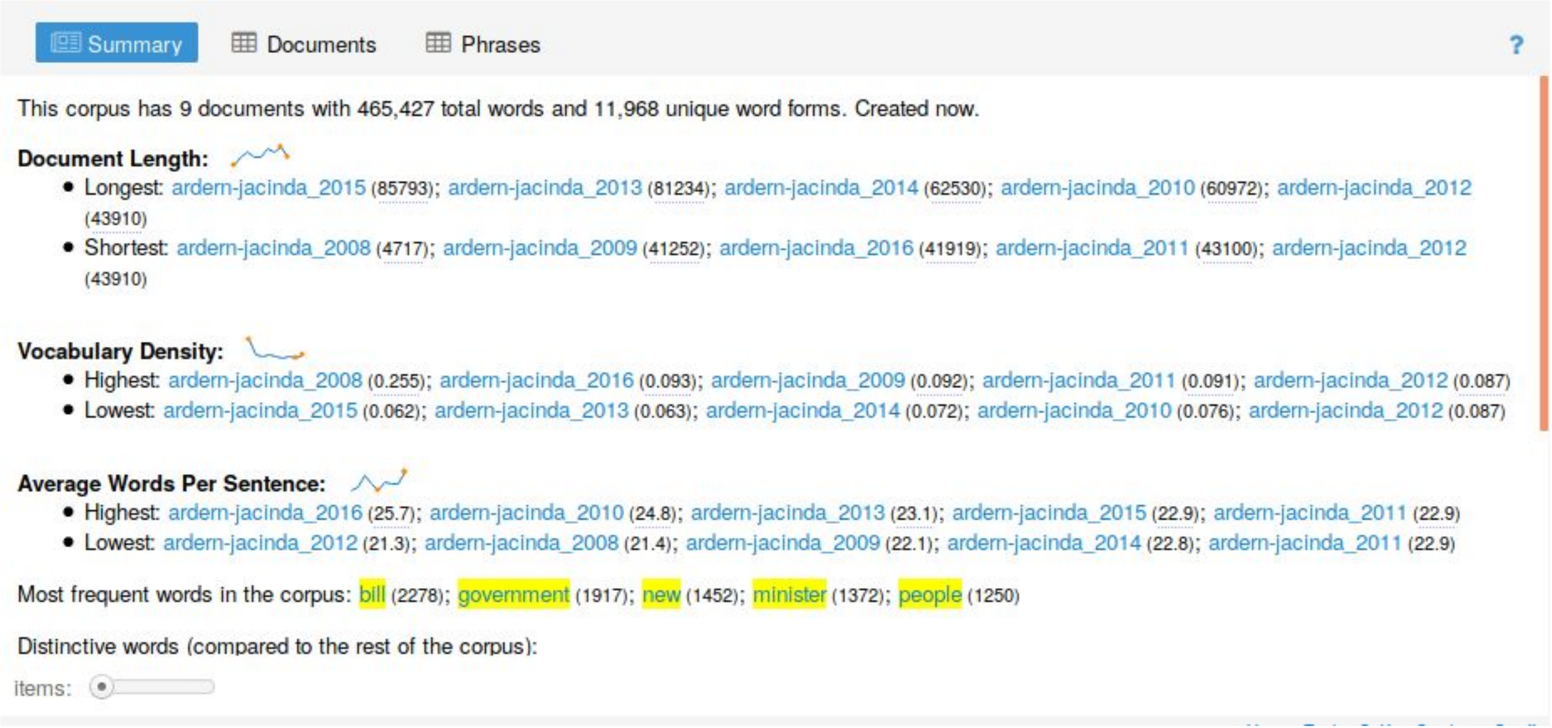
Preparing the corpus… a few examples

- Differences in style



Preparing the corpus… a few examples

- Evolution of style



This corpus has 9 documents with 465,427 total words and 11,968 unique word forms. Created now.

Document Length:

- Longest: ardern-jacinda_2015 (85793); ardern-jacinda_2013 (81234); ardern-jacinda_2014 (62530); ardern-jacinda_2010 (60972); ardern-jacinda_2012 (43910)
- Shortest: ardern-jacinda_2008 (4717); ardern-jacinda_2009 (41252); ardern-jacinda_2016 (41919); ardern-jacinda_2011 (43100); ardern-jacinda_2012 (43910)

Vocabulary Density:

- Highest: ardern-jacinda_2008 (0.255); ardern-jacinda_2016 (0.093); ardern-jacinda_2009 (0.092); ardern-jacinda_2011 (0.091); ardern-jacinda_2012 (0.087)
- Lowest: ardern-jacinda_2015 (0.062); ardern-jacinda_2013 (0.063); ardern-jacinda_2014 (0.072); ardern-jacinda_2010 (0.076); ardern-jacinda_2012 (0.087)

Average Words Per Sentence:

- Highest: ardern-jacinda_2016 (25.7); ardern-jacinda_2010 (24.8); ardern-jacinda_2013 (23.1); ardern-jacinda_2015 (22.9); ardern-jacinda_2011 (22.9)
- Lowest: ardern-jacinda_2012 (21.3); ardern-jacinda_2008 (21.4); ardern-jacinda_2009 (22.1); ardern-jacinda_2014 (22.8); ardern-jacinda_2011 (22.9)

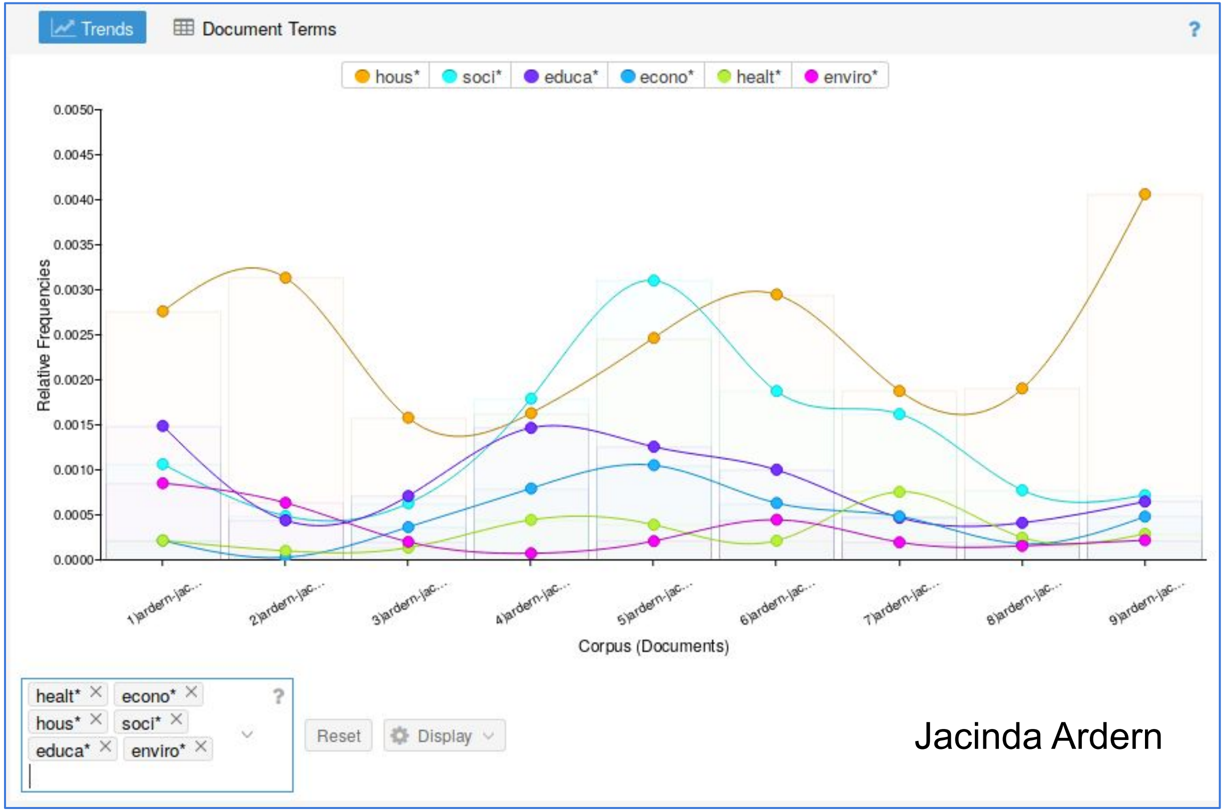
Most frequent words in the corpus: bill (2278); government (1917); new (1452); minister (1372); people (1250)

Distinctive words (compared to the rest of the corpus):

items:

Preparing the corpus… a few examples

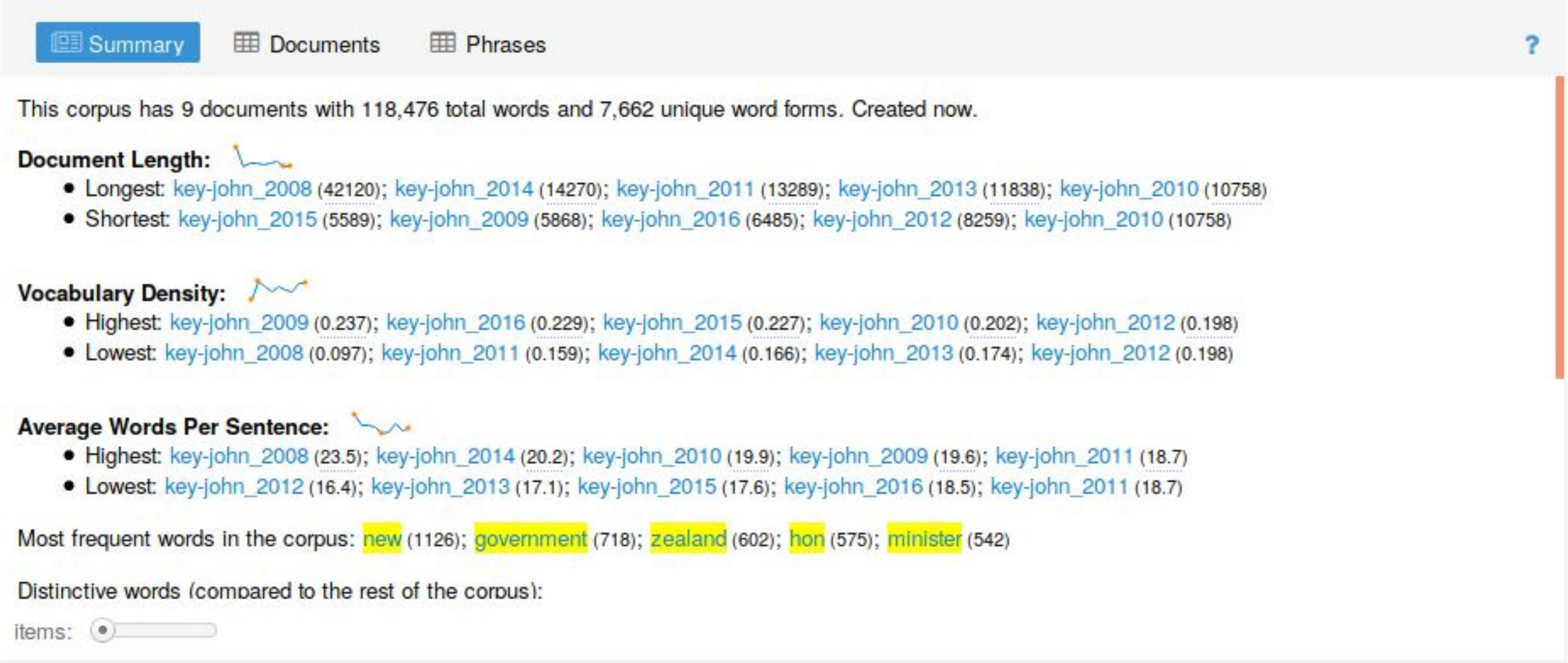
- Evolution of style



Jacinda Ardern

Preparing the corpus… a few examples

- Evolution of style



This corpus has 9 documents with 118,476 total words and 7,662 unique word forms. Created now.

Document Length:

- Longest: key-john_2008 (42120); key-john_2014 (14270); key-john_2011 (13289); key-john_2013 (11838); key-john_2010 (10758)
- Shortest: key-john_2015 (5589); key-john_2009 (5868); key-john_2016 (6485); key-john_2012 (8259); key-john_2010 (10758)

Vocabulary Density:

- Highest: key-john_2009 (0.237); key-john_2016 (0.229); key-john_2015 (0.227); key-john_2010 (0.202); key-john_2012 (0.198)
- Lowest: key-john_2008 (0.097); key-john_2011 (0.159); key-john_2014 (0.166); key-john_2013 (0.174); key-john_2012 (0.198)

Average Words Per Sentence:

- Highest: key-john_2008 (23.5); key-john_2014 (20.2); key-john_2010 (19.9); key-john_2009 (19.6); key-john_2011 (18.7)
- Lowest: key-john_2012 (16.4); key-john_2013 (17.1); key-john_2015 (17.6); key-john_2016 (18.5); key-john_2011 (18.7)

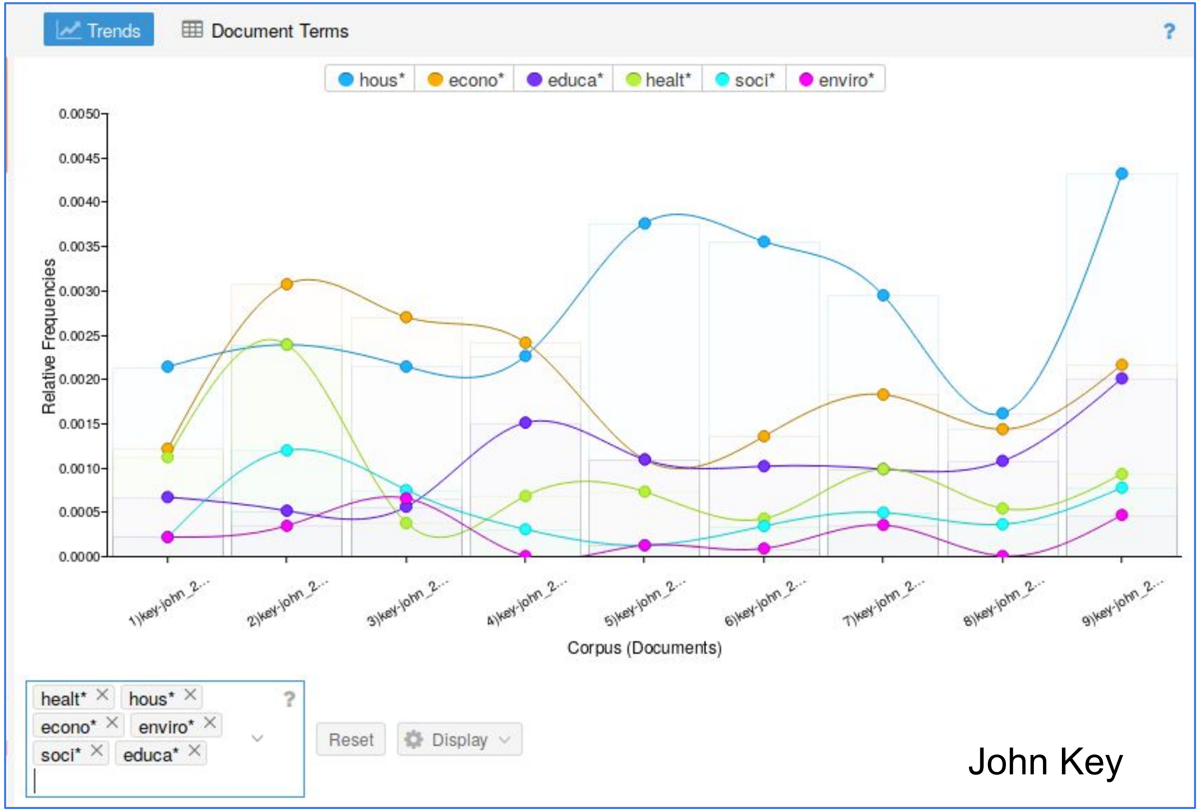
Most frequent words in the corpus: new (1126); government (718); zealand (602); hon (575); minister (542)

Distinctive words (compared to the rest of the corpus):

items:

Preparing the corpus… a few examples

- Evolution of style



John Key

DigiVoy - integration tool

- DigiVoy is a toll developed to facilitate text analysis by integrating Voyant Tools with:

https://textgridrep.org/

TextGrid Repository:

- provides an extensive, searchable and reusable collection of XML/TEI-coded texts, images and databases,
- includes works by around 600 authors of German-language fiction (prose, poetry, drama) and non-fiction.

DigiVoy - integration tool

Three simple steps:

1) Search text on TextGrid Repository.
2) Add text to the shelf.
3) Export to Voyant Tools.

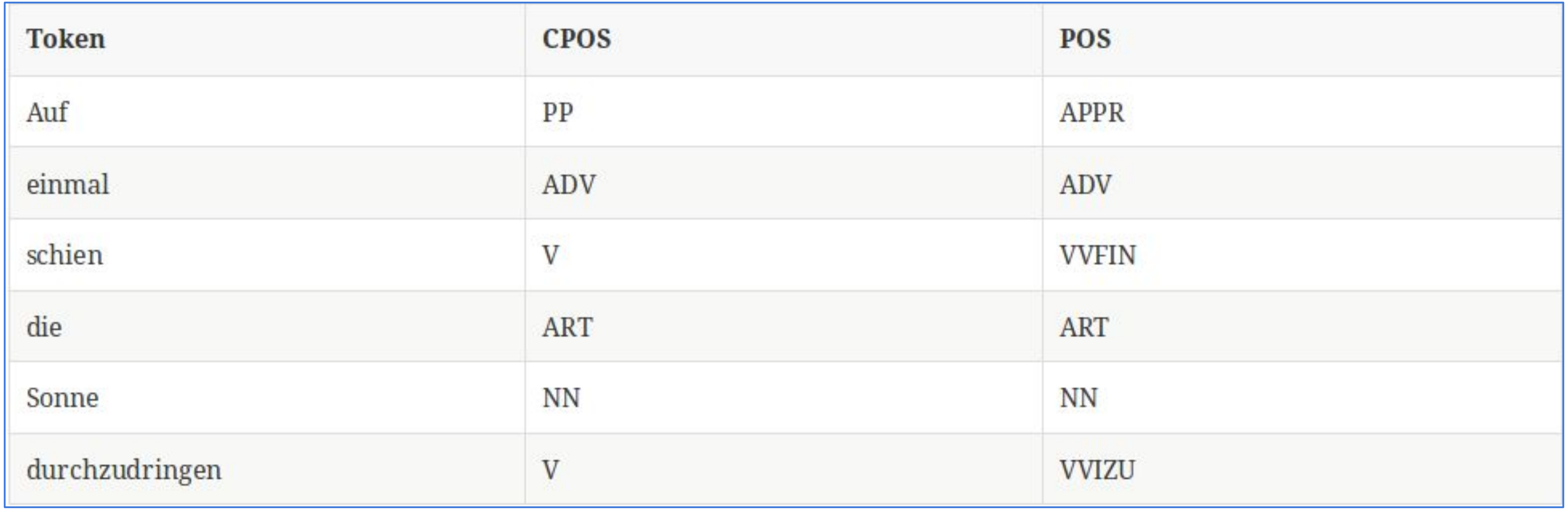
DKPro-Wrapper

Programming framework - a bundle of tools for complex methods of text analysis (NLP - Natural Language Processing), e.g:

- Segmentation
- Part-of-Speech Tagging
- Named Entity Recognition
- Topic modeling

DKPro-Wrapper

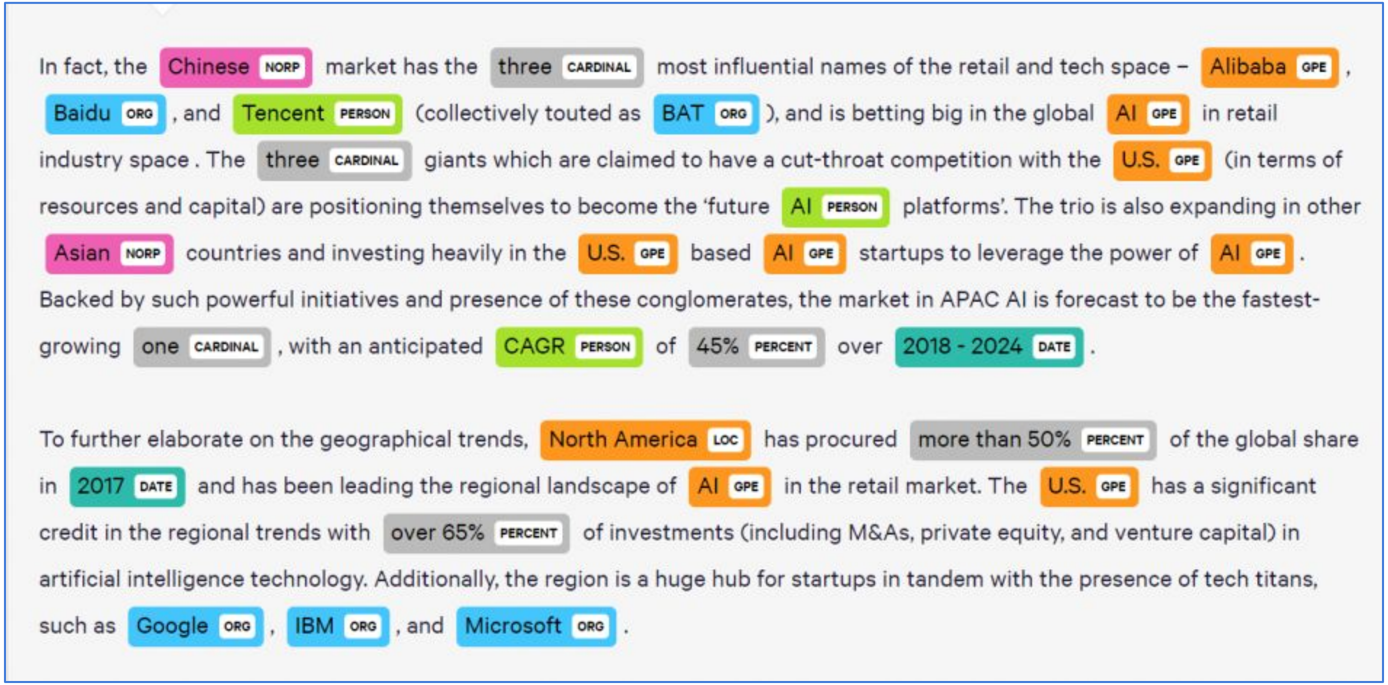
Part-of-Speech Tagging



Token	CPOS	POS
Auf	PP	APPR
einmal	ADV	ADV
schien	V	VVFIN
die	ART	ART
Sonne	NN	NN
durchzudringen	V	VVIZU

DKPro-Wrapper

Named Entity Recognition

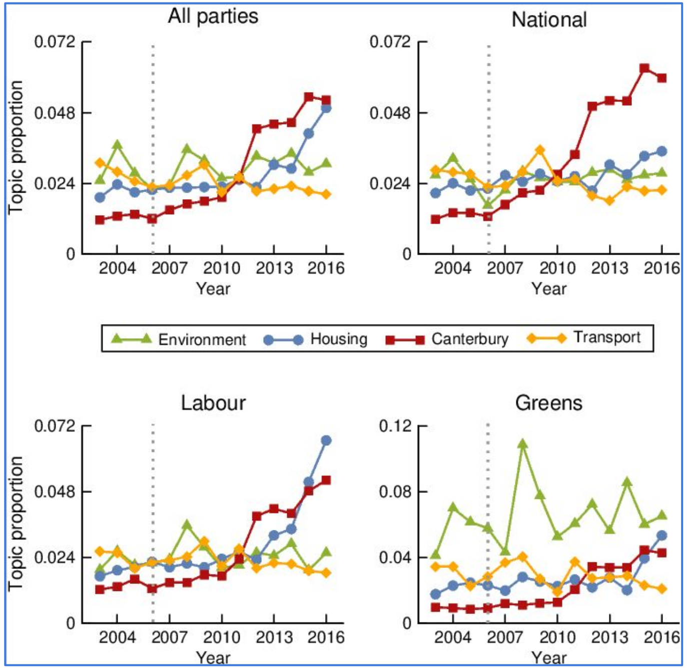


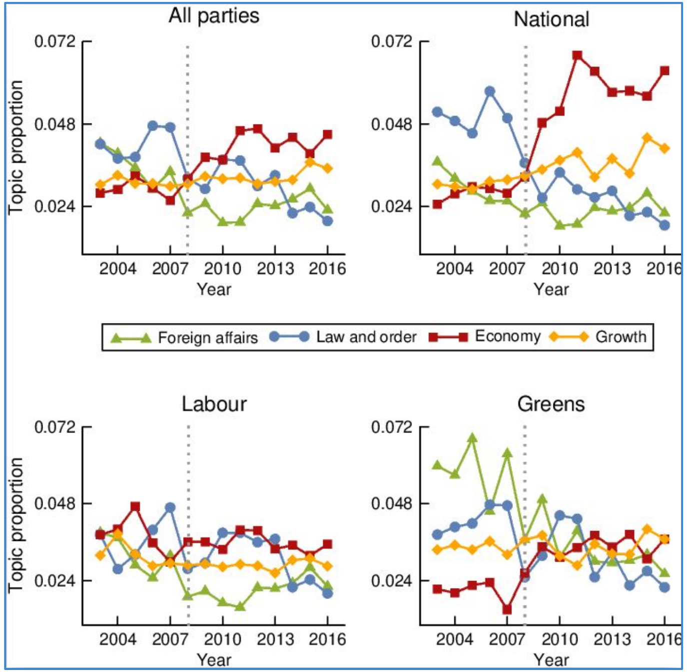
In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

DKPro-Wrapper

Topic modeling





Danke schön!

Fragen?

vasquesfilho@ieg-mainz.de

@demivasques