Data Preprocessing and Transformation

Why ??

Garbage in -> Garbage out

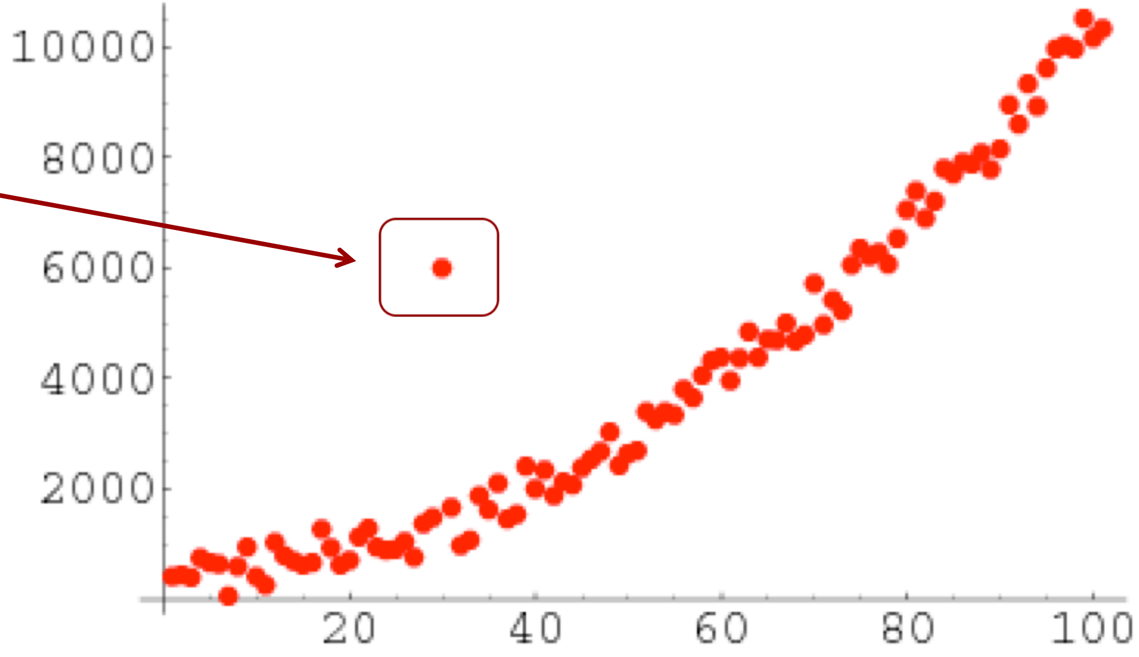
Data Preprocessing and Transformation

Pre-Processing	Cleaning
Selection	Sampling Feature subset selection
Transformation	Aggregation Dimensionality reduction Feature transformation Discretization and Binarization

Cleaning – removing data that we won't need, or will hinder the analysis

Numerical data

- Removing outliers



Numerical

Cleaning – removing data that we won’t need, or will hinder the analysis

If data with tags (e.g. <*title*>The Title”</*title*>):

- Removing Tags

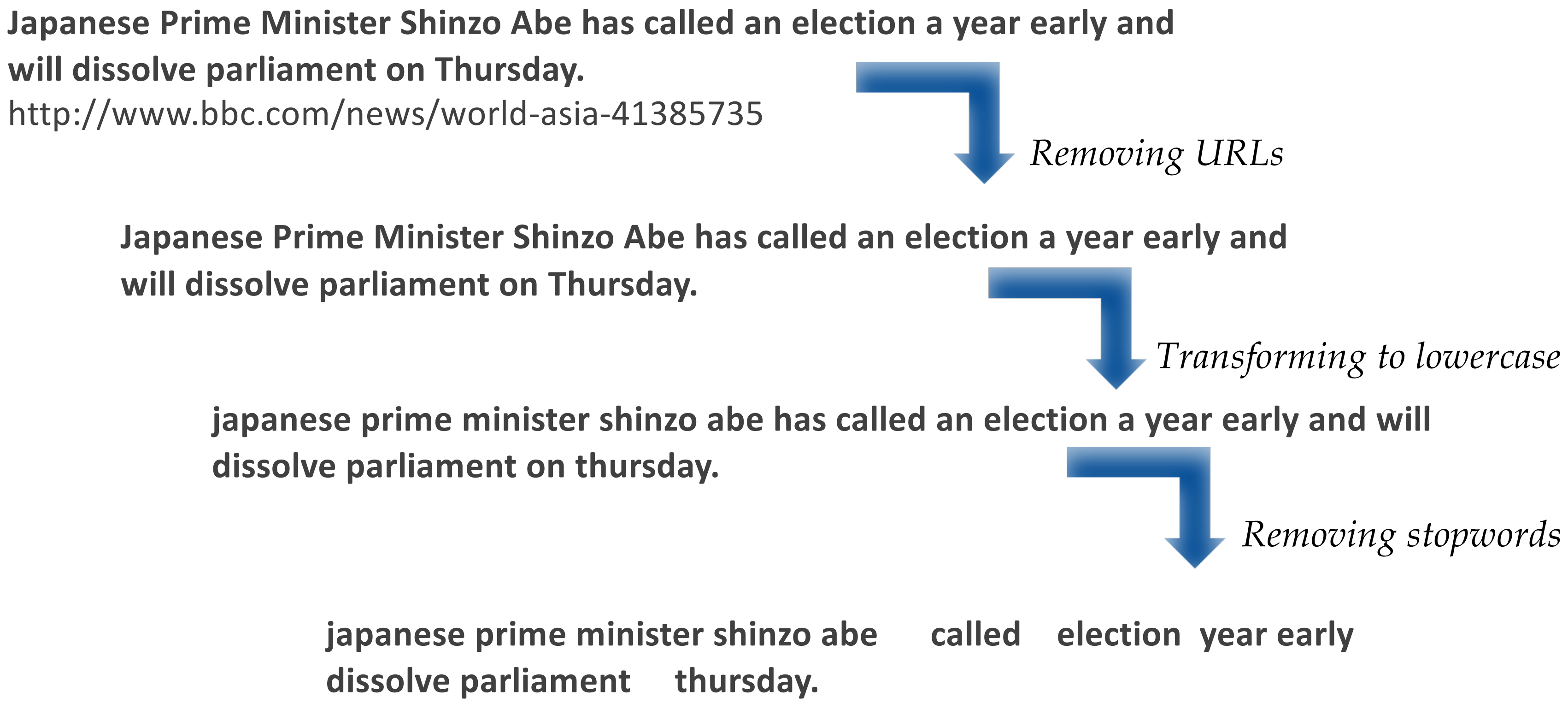
Common ways for any text - (based on the aim in mind) to preprocess:

- Lowercase
- Remove punctuations (.,!?-+…)
- Remove stop words (*and, but, not, no*, etc..)
- Lemmatize (*convert plurals to singulars*)
- Stemming

Specific for the project

- Removing URLs from text

Textual



Japanese Prime Minister Shinzo Abe has called an election a year early and will dissolve parliament on Thursday.
http://www.bbc.com/news/world-asia-41385735

Removing URLs

Japanese Prime Minister Shinzo Abe has called an election a year early and will dissolve parliament on Thursday.

Transforming to lowercase

japanese prime minister shinzo abe has called an election a year early and will dissolve parliament on thursday.

Removing stopwords

japanese prime minister shinzo abe   called   election year early dissolve parliament   thursday.

Textual

Data Selection – selecting only objects (units of analysis) or features (variables) that will be passed to the analysis

Usually done due to efficiency reasons…

saving time, computational resources

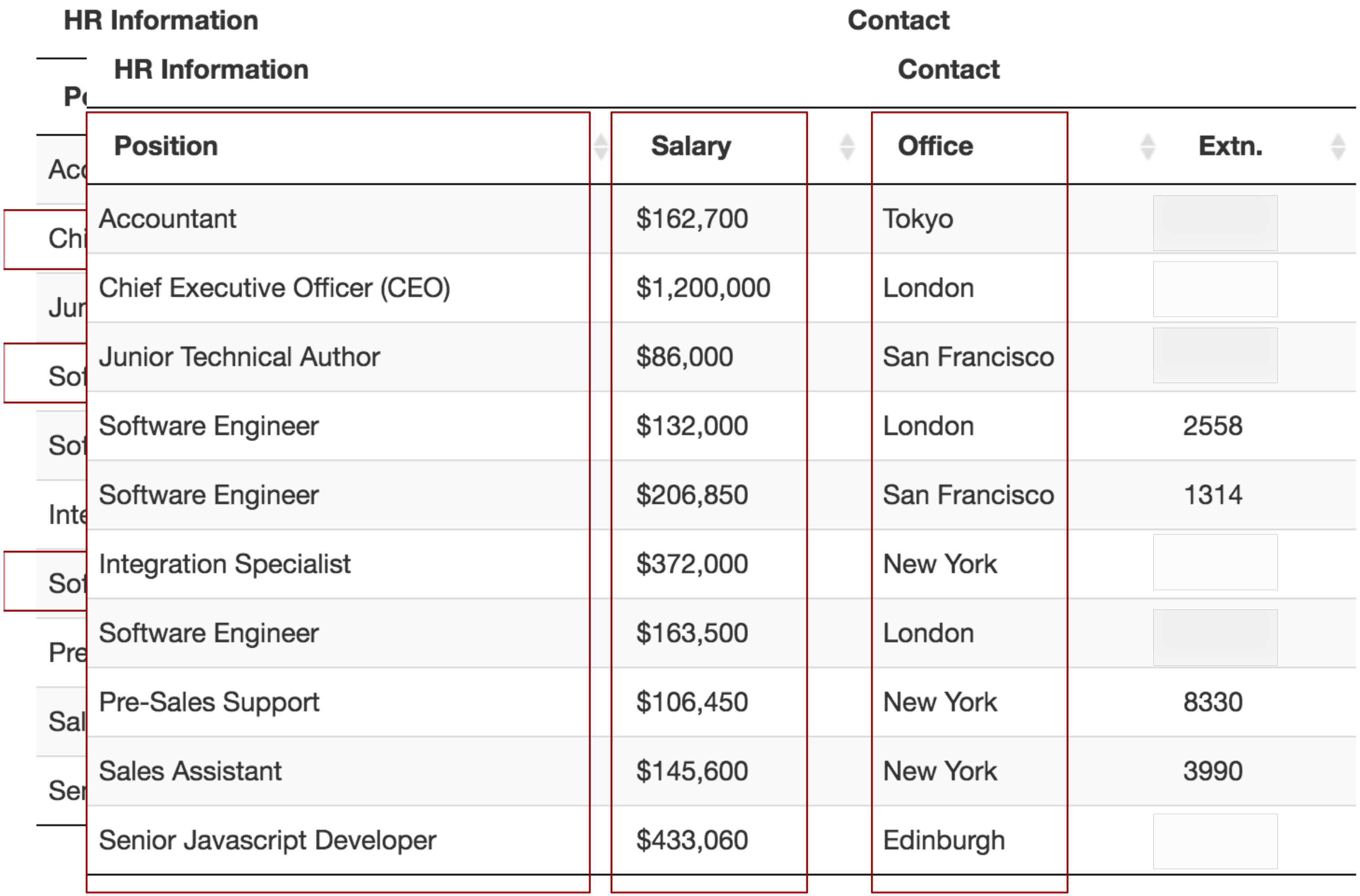
…also due to missing values

Sampling – selection of objects:

- Simple random sampling
    - with replacement
    - without replacement
- Stratified Sampling

Feature subset selection:

- Occurs naturally with algorithm
    - *E.g. Classification trees*
- Filter approach
    - *Researcher selects features based on his experience in the field*
- Using algorithm for feature selection only



HR Information Contact

HR Information		Contact	
Position	Salary	Office	Extn.
Accountant	$162,700	Tokyo	
Chief Executive Officer (CEO)	$1,200,000	London	
Junior Technical Author	$86,000	San Francisco	
Software Engineer	$132,000	London	2558
Software Engineer	$206,850	San Francisco	1314
Integration Specialist	$372,000	New York	
Software Engineer	$163,500	London	
Pre-Sales Support	$106,450	New York	8330
Sales Assistant	$145,600	New York	3990
Senior Javascript Developer	$433,060	Edinburgh	

Numerical

Data Transformation – feature or object creation

Aggregation – combining objects (units of analysis), or features (variables)

Dimensionality reduction– merging many features (variables) into few.

Principal component analysis - for 2d visualization or computation efficiency

Numerical

Feature transformation – shaping values of particular feature usually by some mathematical formula, can be text transformation as well, for a better precision, or emphasizing/denying differences in values

x^2, log(x), standardization, stemming, sensitive data to non

Power

HRS worked per week	(HRS worked per week)^2
40.000	1600
40.000	1600
58.000	3364
40.000	1600
40.000	1600
24.000	576
16.000	256
40.000	1600
15.000	225
50.000	2500
40.000	1600
45.000	2025
60.000	3600
60.000	3600
9.000	81

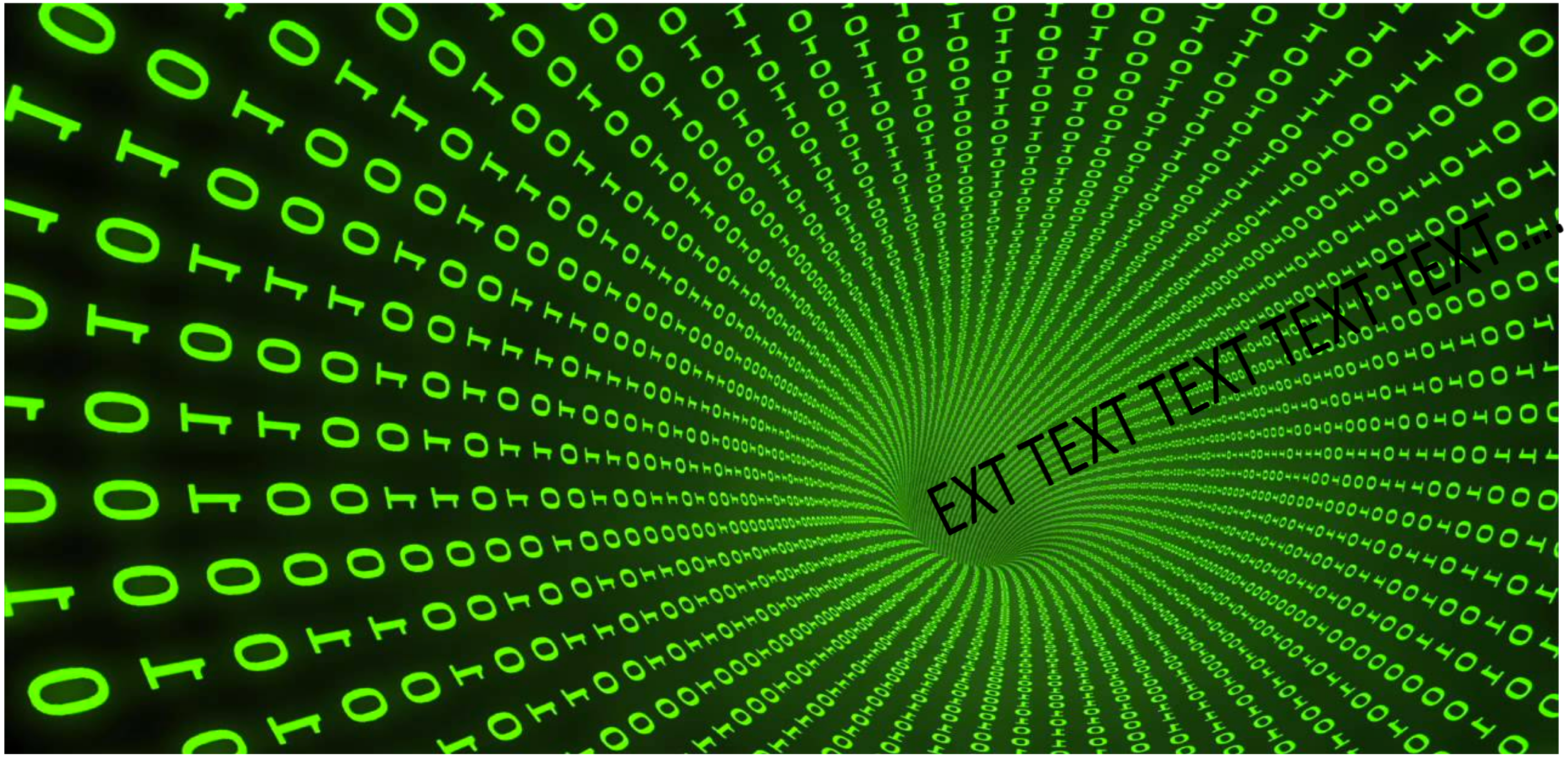
Square root

Salary a month $	sqrt(Salary a month $)
0	0
0	0
0	0
0	0
4064	64
1055	32
0	0
0	0
0	0
0	0
0	0
7688	88
15024	123
15024	123
0	0

Numerical

How we can apply linear regression on Text

Transforming text to number



Examples for text transformation

- Calculating amount of characters that a piece of text has
  - a dog – 5chars, lesson – 6chars, going for a walk. – 17chars
- Sentiment analysis – transforming text to positive or negative number based on the mood or appearance of positive/negative words
  - good very good sushi, but bad ramen - +1, the sailor looked through his depressed eyes… -1
- Calculating amount of appearing particular words
  - and – 5 times, aye – 10 times, 1978 – 3 times
- Tracking appearance of words together
  - Bag of words
- Calculating similarity between texts
  - "Fax" and "tax" has 1 on Leveinshtein distance

Discretization and binarization – transforming one type of data to different type of data. Continuous numbers to integers, text to numbers.

Dictionary

Word	ID
japanese	1
prime	2
minister	3
shinzo	4
abe	5
called	6
election	7
year	8
……	

Document1

Minister of japan shinzo abe. shinzo has called an election a year early and shinzo will dissolve parliament on thursday.

Transforming

Document1 -[(1,0)(2,0)(3,1)(4,3)….]

Textual

Text Summarization techniques

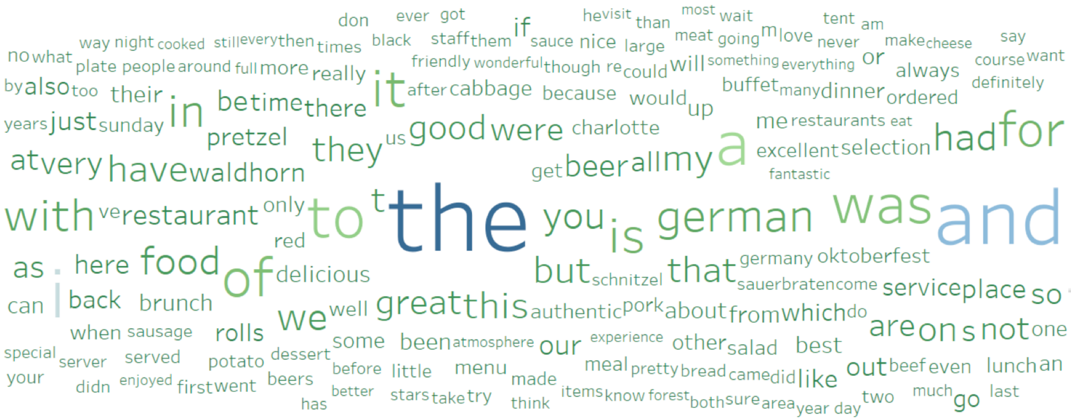
Why???

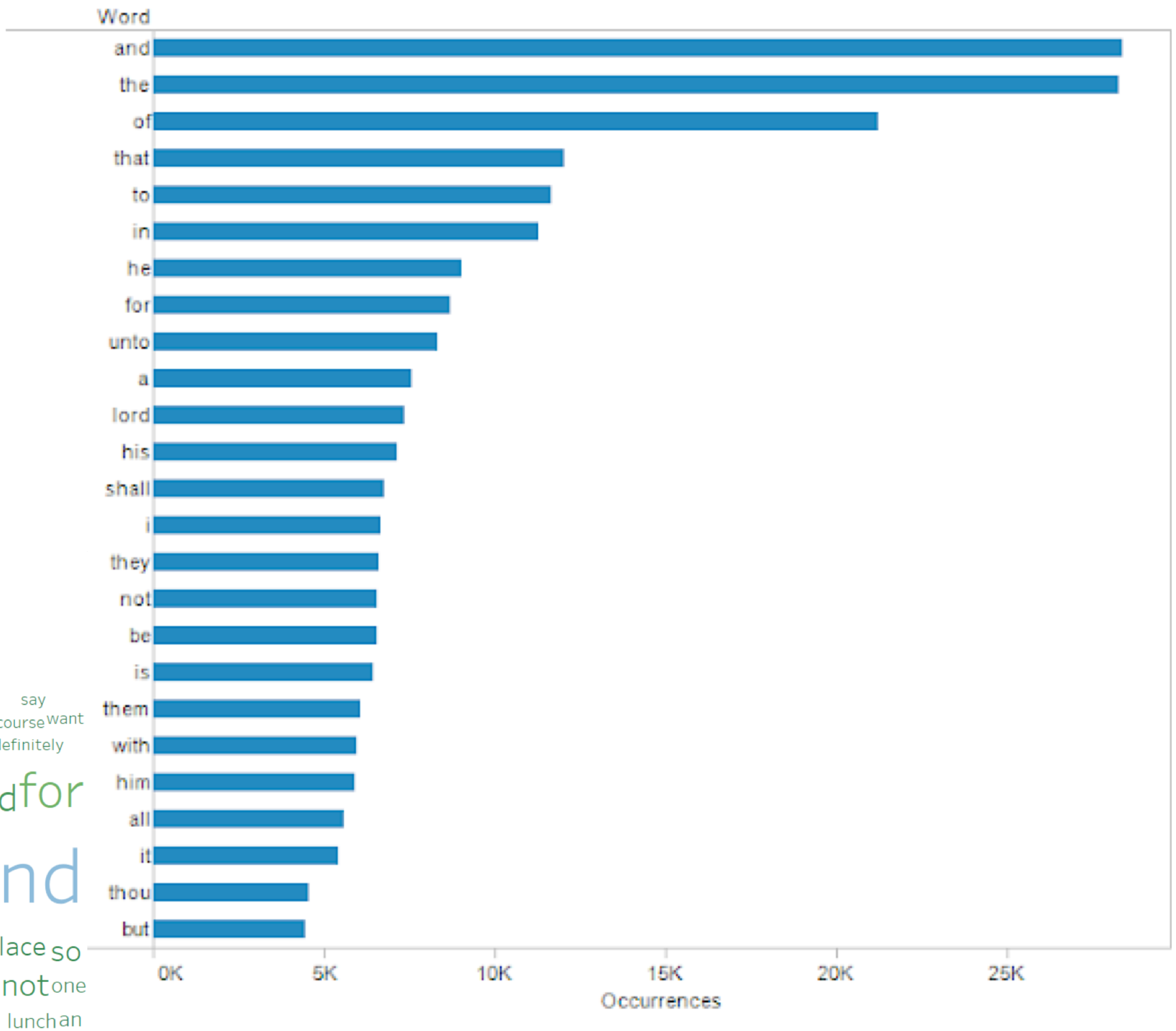
Understanding a piece of text without reading it

Word cloud

the most frequent text

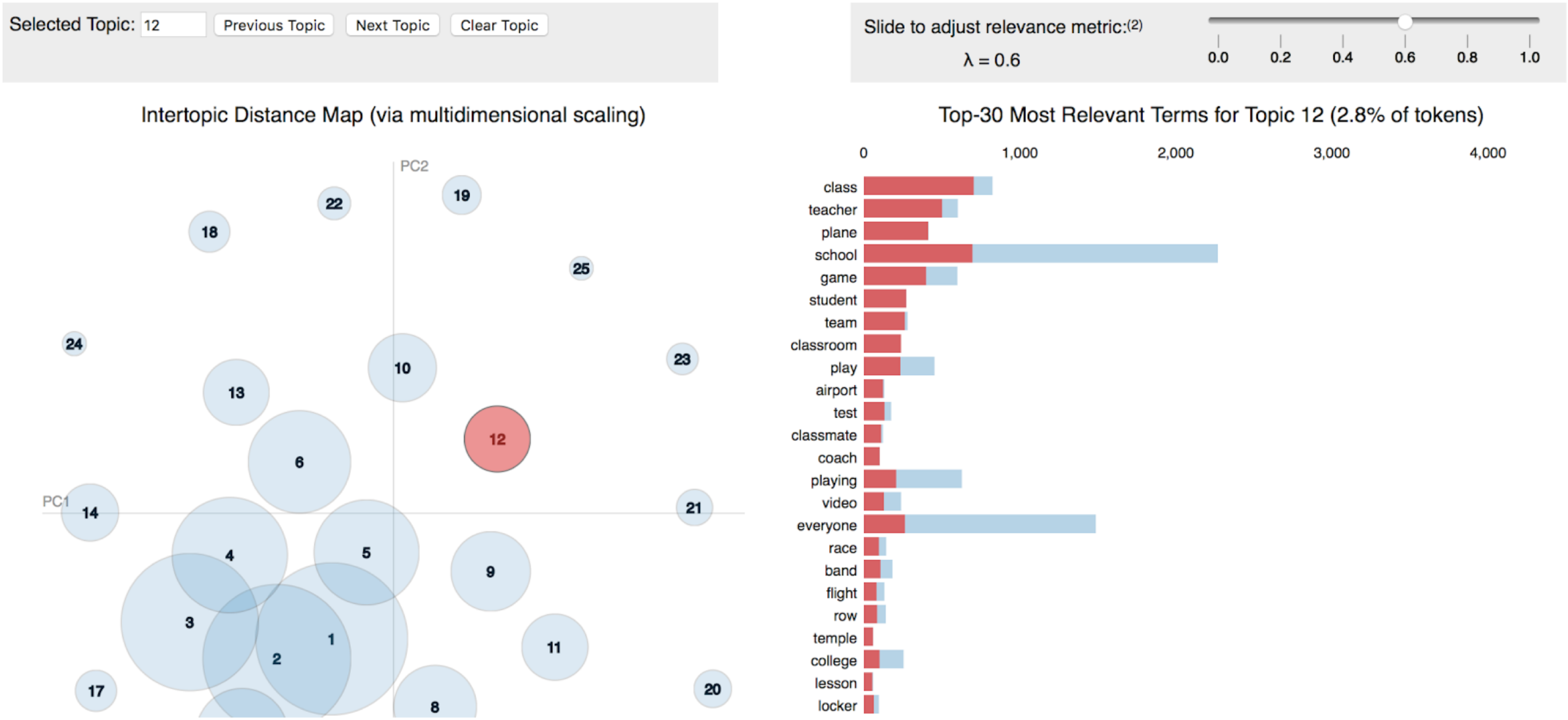
(single word or a phrase)





Topic Modeling (associations between text)

discovering the abstract “topics” that occur in a collection of documents.

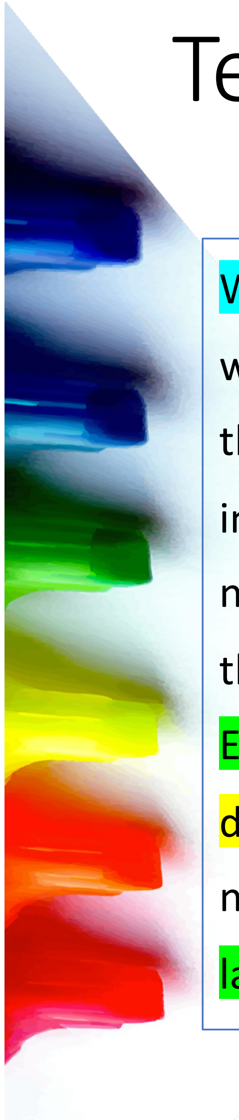


Selected Topic: 12

Slide to adjust relevance metric:(2)

λ = 0.6

Texts are collections of different topics



WoS is also dominated by publications and authors writing in English. In the data set that we analyzed, more than 73% of the publications are written in English, 12% in French, and the other 15% in 24 other languages. BJA, naturally, is all in English. However, it is not reasonable to think that aesthetic issues would only be addressed in English, especially because many of them are highly dependent on culture and language. In the future, we need digital databases that better cover several languages.

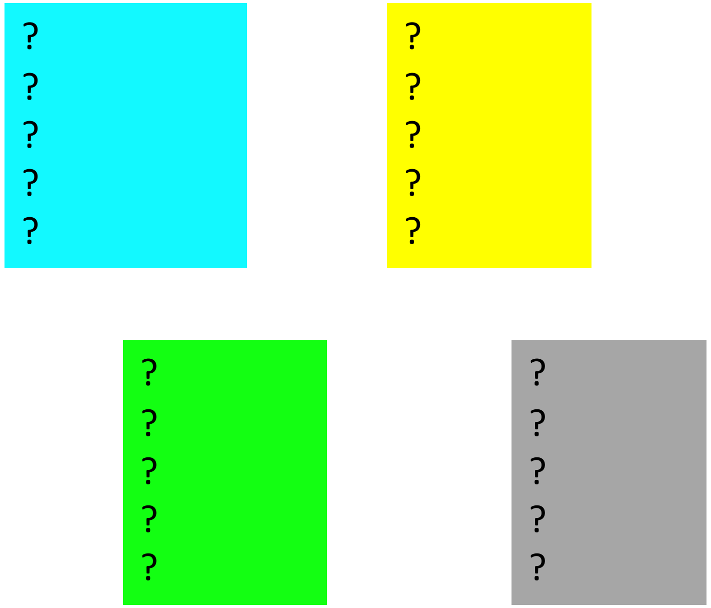
WoS
publications
data
BJA
databases

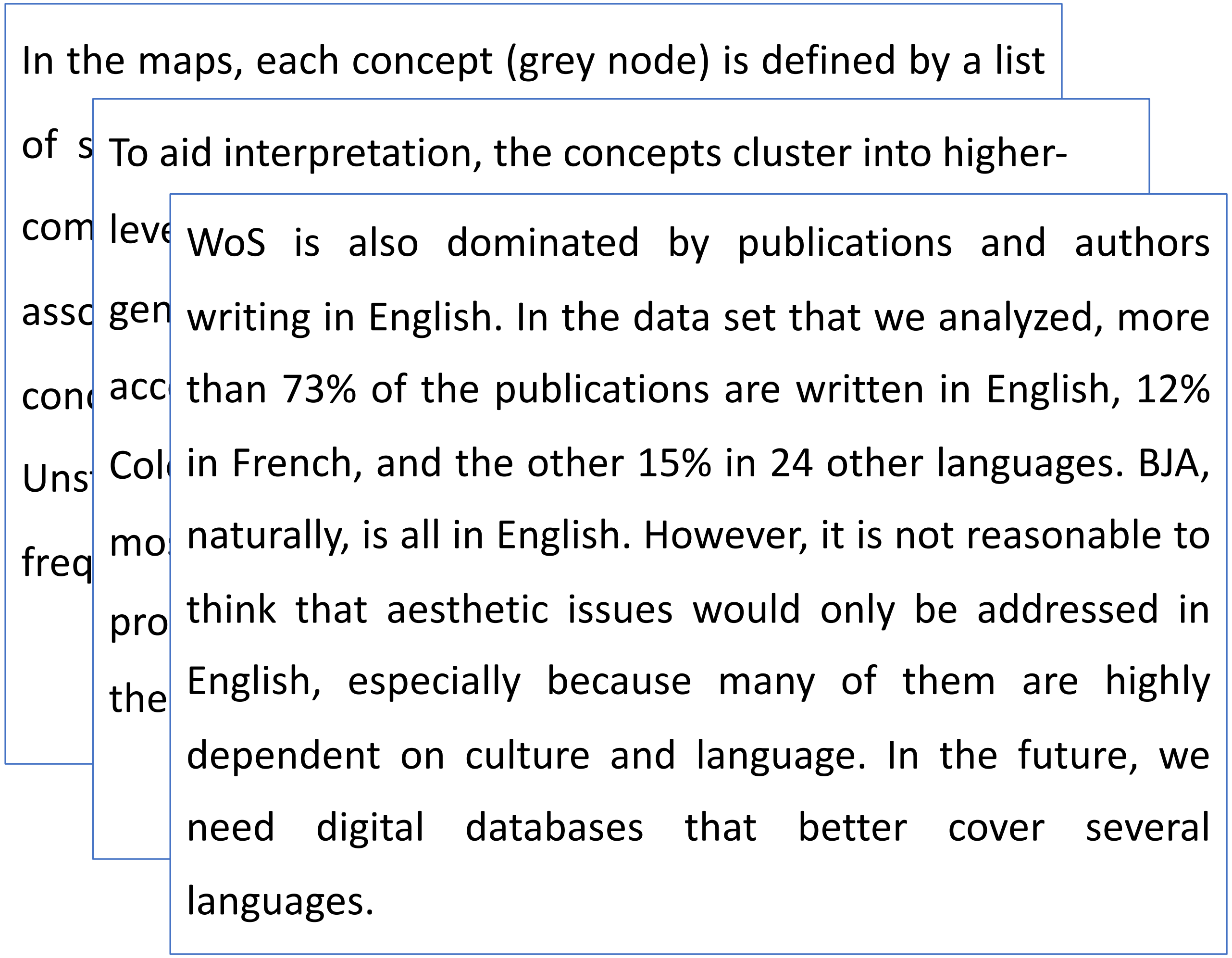
analysed
dependent
digital

English
French
languages
language

aesthetic
culture
future

Reversing markup

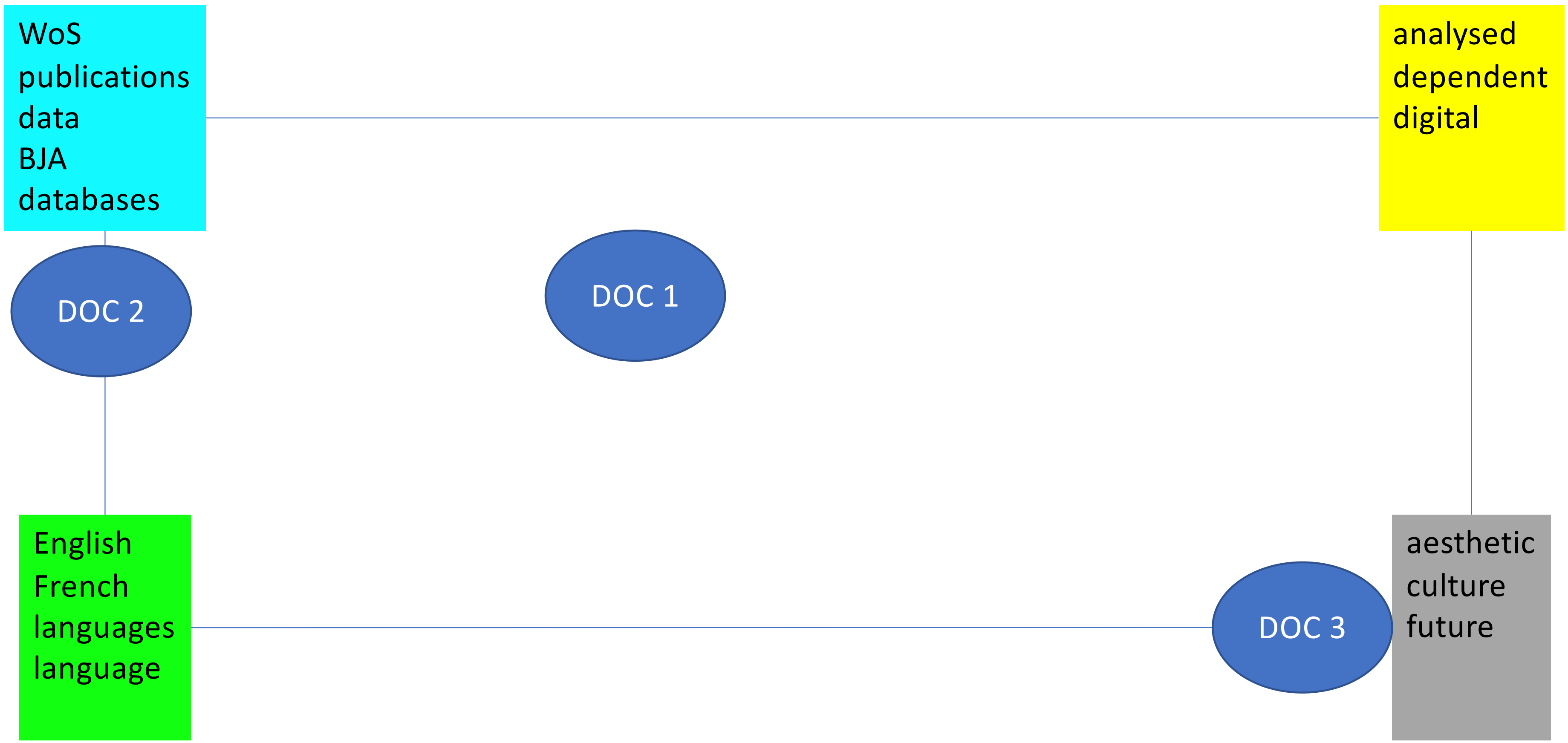




In the maps, each concept (grey node) is defined by a list
of s
com
assc
conc
Uns
freq

To aid interpretation, the concepts cluster into higher-
leve
gen
acc
Col
mo
pro
the

WoS is also dominated by publications and authors writing in English. In the data set that we analyzed, more than 73% of the publications are written in English, 12% in French, and the other 15% in 24 other languages. BJA, naturally, is all in English. However, it is not reasonable to think that aesthetic issues would only be addressed in English, especially because many of them are highly dependent on culture and language. In the future, we need digital databases that better cover several languages.


WoS
publications
data
BJA
databases
analysed
dependent
digital
DOC 2
DOC 1
English
French
languages
language
DOC 3
aesthetic
culture
future

How does it work?

WoS is also dominated by publications and authors writing in English. In the data set that we analyzed, more than 73% of the publications are written in English, 12% in French, and the other 15% in 24 other languages. BJA, naturally, is all in English. However, it is not reasonable to think that aesthetic issues would only be addressed in English, especially because many of them are highly dependent on culture and language. In the future, we need digital databases that better cover several languages.

dependent
language
BJA
databases

analysed
culture
WoS
data

English
BJA
Culture
languages

aesthetic
publications
future
French

In the data set that we analyzed, more than 73% of the publications are written in English, 12% in French, and the other 15% in 24 other languages.

English | analysed | data | French | languages

Topic 1:

Topic 2:

Topic 3:

Topic 4:

dependent
language
BJA
databases

analysed
culture
WoS
data

English
BJA
Culture
languages

aesthetic
publications
future
French

In the data set that we analyzed, more than 73% of the publications are written in English, 12% in French, and the other 15% in 24 other languages.

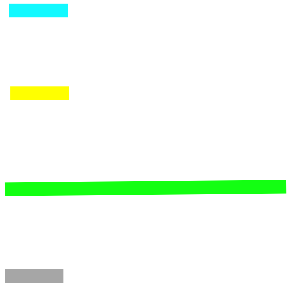
??English?? | analysed | data | French | languages

Topic 1: **0**

Topic 2: **0**

Topic 3: 1

Topic 4: **0**



dependent
language
BJA
databases

analysed
culture
WoS
data

English
BJA
Culture
language2

aesthetic
publications
future
French

The word occurrence in different topics

In the data set that we analyzed, more than 73% of the publications are written in English, 12% in French, and the other 15% in 24 other languages.

??English?? | **analysed** | **data** | **French** | **languages**

All words from the document occurrence in different topics

dependent
language
BJA
databases

analysed
culture
WoS
data

English
BJA
Culture
languages

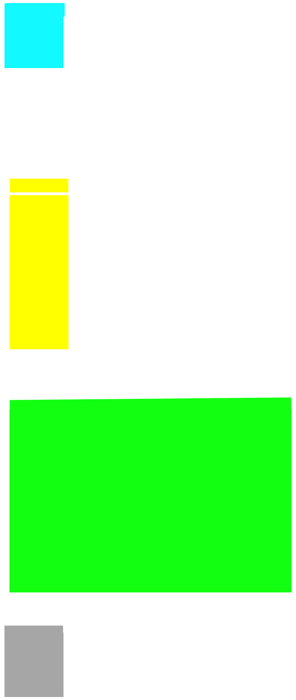
aesthetic
publications
future
French

Topic 1: **1**

Topic 2: **2**

Topic 3: 2

Topic 4: **1**



Analysing each word in each document and repeating multiple times the process

Final result:

WoS
publications
data
BJA
databases

analysed
dependent
digital

English
French
languages
language

aesthetic
culture
future