

# JM2050 – Natural Language Processing

Introduction to text analysis

September 12<sup>th</sup>, 2024

Prof.dr.ir. U. Kaymak

JM2050

Jheronimus  
Academy  
of Data Science

## Recap

- Challenges with text data
  - Ambiguity
  - Variation
  - World knowledge
  - Context
- Major types of learning
  - Supervised
  - Unsupervised
  - Reinforcement learning
- Performance measurement
  - Metrics derived from confusion matrix
- Balance between model fit and generalization

[2]

## Outline

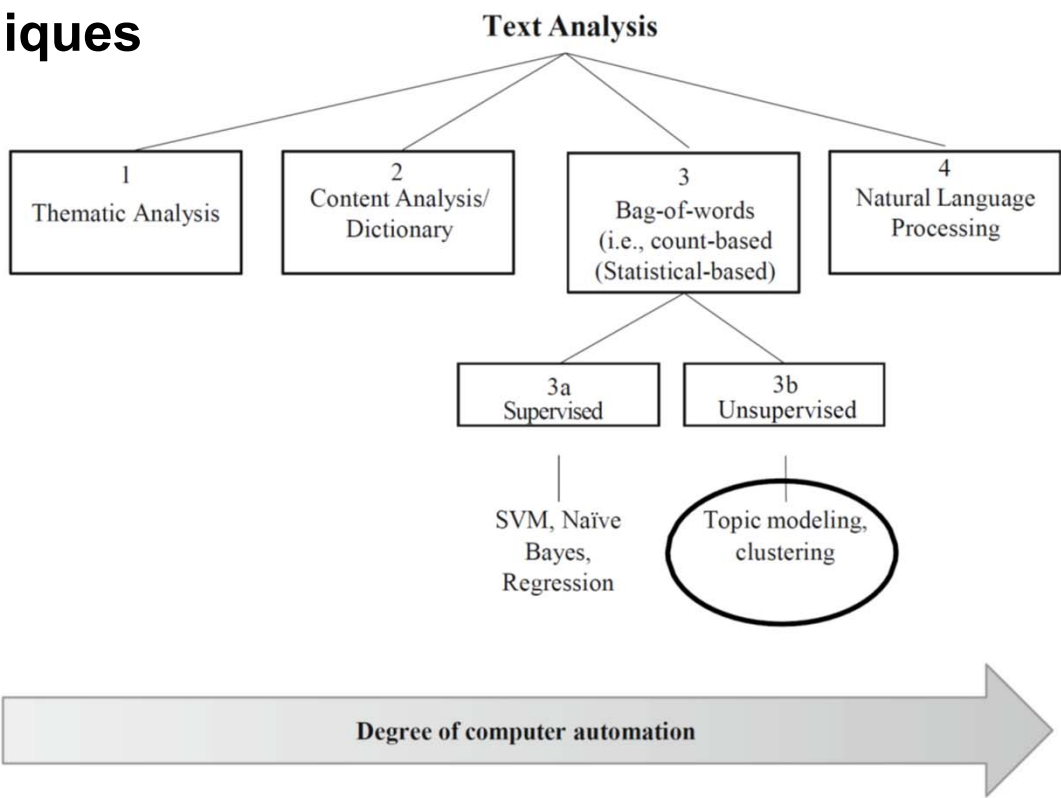
- Descriptive text analysis
- Pre-processing text
- Regular expressions

# Descriptive text analysis

Acknowledgement: slides adopted from K. Zervanou

JADS Jheronimus Academy of Data Science

# Text analysis techniques



(Banks et al., 2018)

[5]

Some common natural language processing tasks

<u>Text classification</u>	<u>Information retrieval</u>	<u>Information extraction</u>
spam filtering	<u>recommender systems</u>	Template-filling
<u>topic modeling</u>	<u>search engine</u>	<u>named entity recognition (NER)</u>
<u>sentiment analysis</u>	<u>question answering</u>	<u>relationship extraction</u>
	Summarization	<u>ontology extraction</u>

6

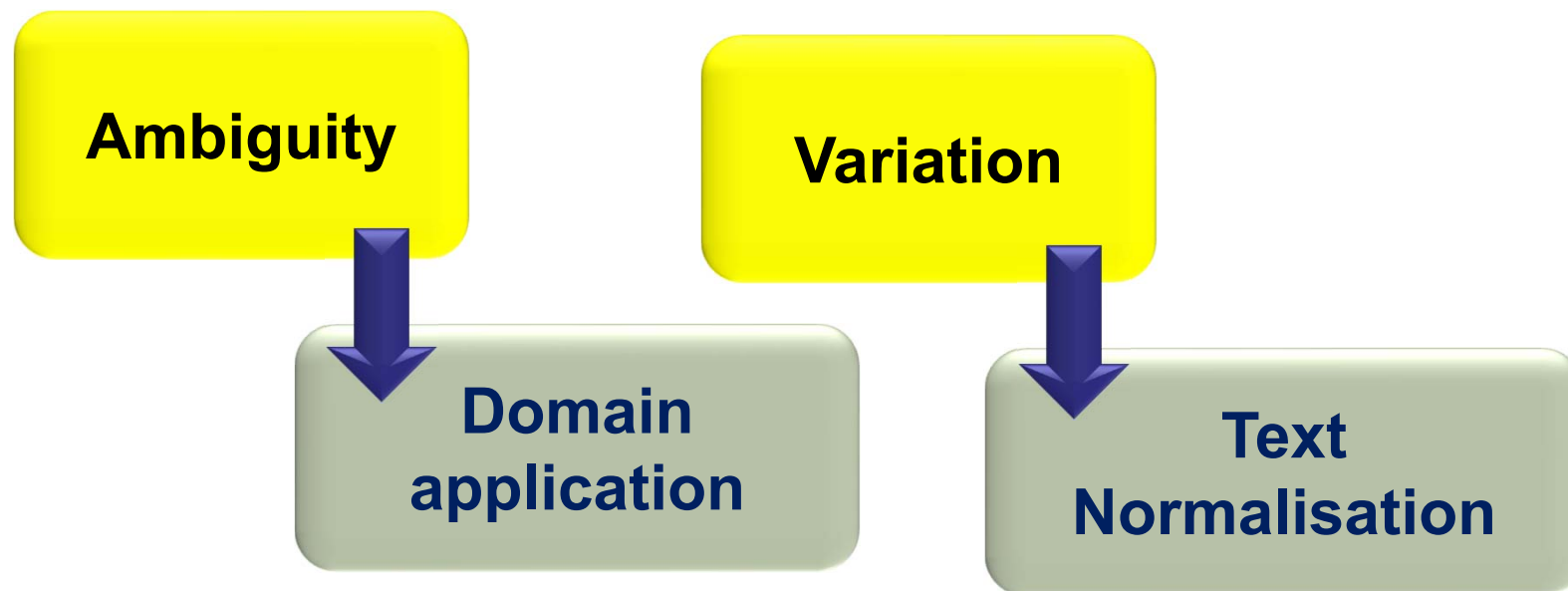
## Basic terminology

- Text – a series of symbols/characters
- Token – a sequence of symbols (characters) that form a useful semantic unit for processing
- Document – a collection of tokens
- Corpus – a collection documents



[7]

## Reduce ambiguity and variation: constraints & pre-processing

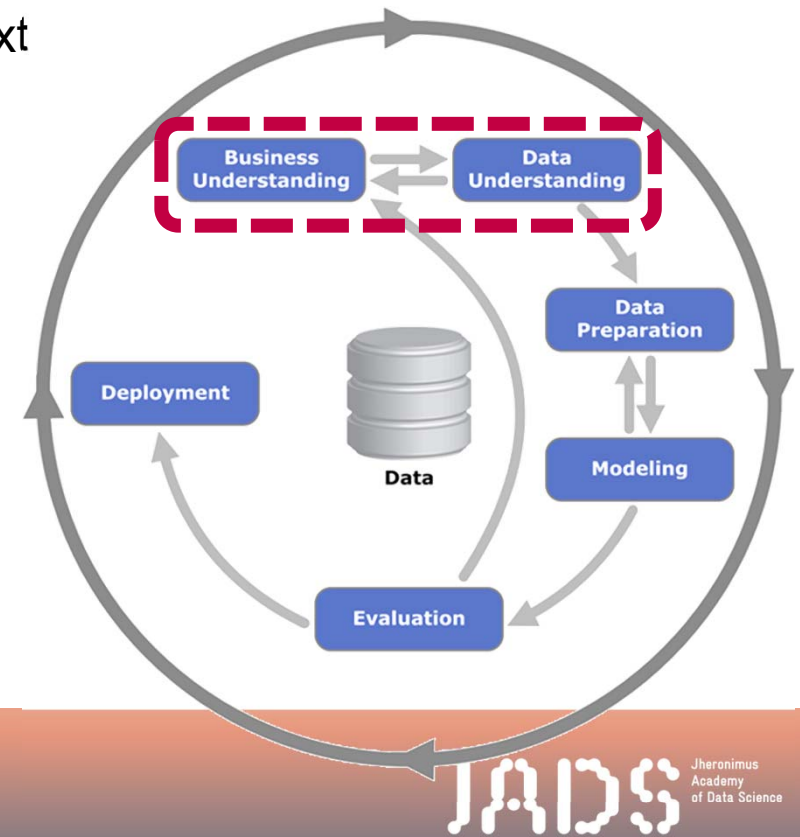




## Domain: *Text type or communication context*

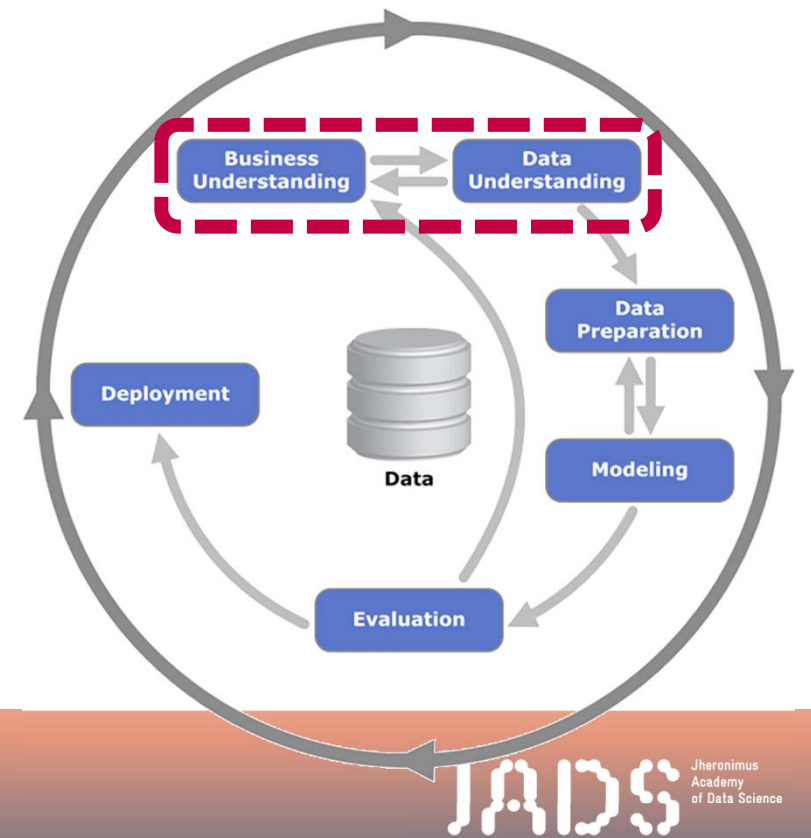
- extra-linguistic/pragmatic document context

❖ *letters,*  
❖ *tweets,*  
❖ *chat,*  
❖ *reports,*  
❖ *news stories,*  
❖ *scientific articles,...*



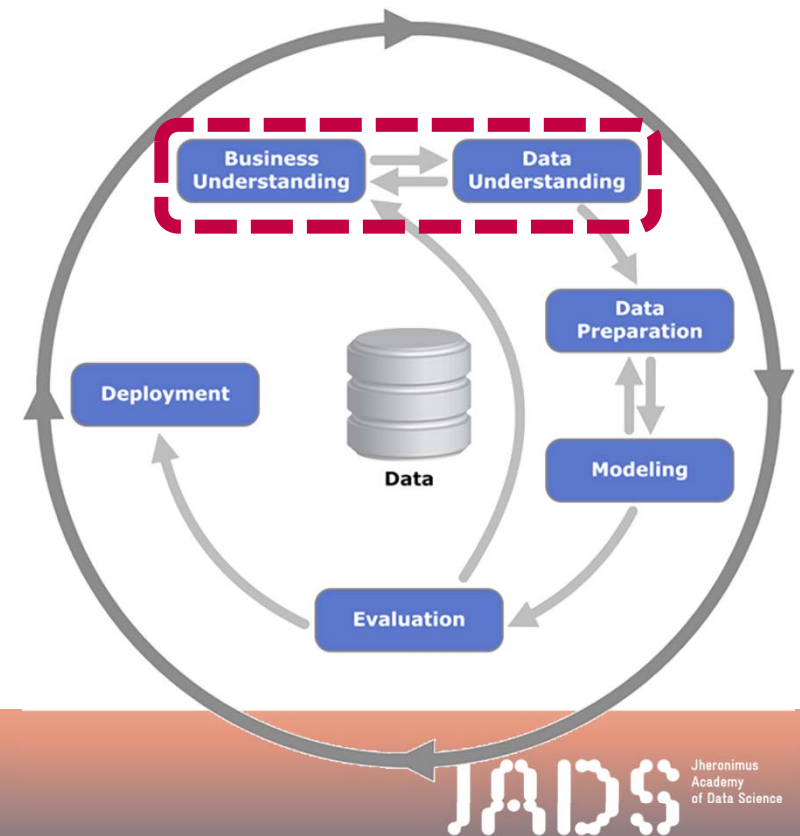
## Domain: *application domain*

- application domain: area of application
- ❖ topics & content
- ❖ vocabulary use: terminology, jargon, general
- ❖ writing style: formal, informal, 🥰 ...
- ❖ language(s)



## Domain: *Corpus characteristics*

- **Corpus**: Document collection
- ❖ **text format**: annotations? Text, XML, HTML, ...?
- ❖ **text encoding**: ASCII? UTF-8?
- ❖ **text unit(s)** of interest: documents? paragraphs? sentences? phrases?
- ❖ **text units length**



## Corpus characteristics: *Know your text data*

What You See Is **NOT** What You Get!



## Corpus characteristics: *Know your text data*

*Documents optimized  
for visual  
presentation or  
machine readability...*

## What is your text document format?

- XML, HTML, text, pdf, MS word document,...

## What is your information structure?

Plain text, XML,  
tabular data,  
combined/distributed  
information

## What is your text encoding?

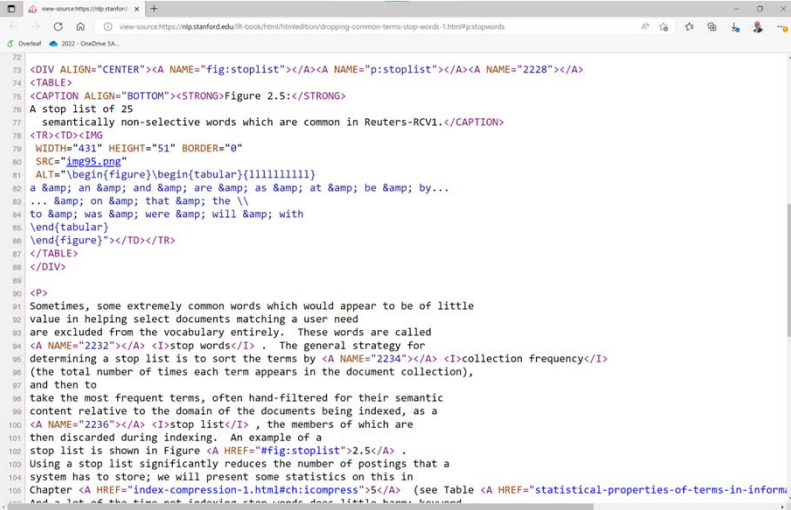
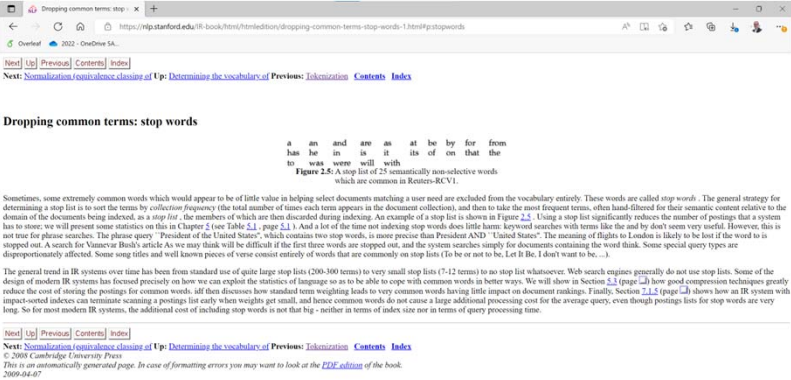
UTF-8, UTF-16,  
ISO 8859, ...

## What You See Is **NOT** What You Get!



[13]

# Example: processing HTML text

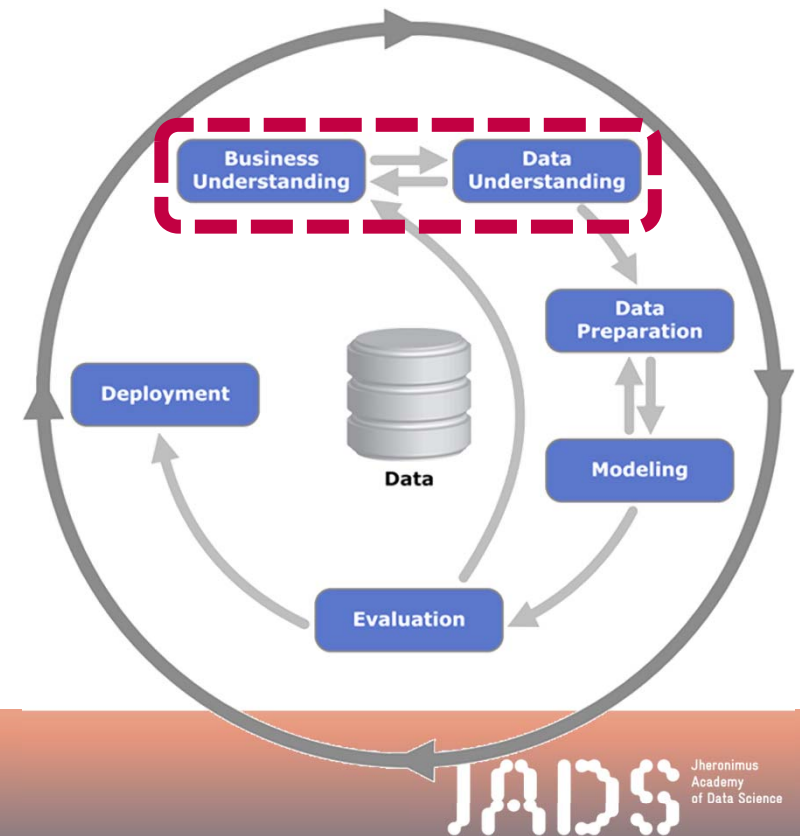


[14]

## Domain: *Corpus characteristics*

**Corpus**: Document collection

- ❖ **vocabulary** richness/variation  
(i.e. unique vs. total “words” number)
- ❖ document **structure**,  
e.g. CS articles, wikipedia, etc.
- ❖ corpus **homogeneity**,  
e.g. wikipedia, news



## Corpus data understanding: *Descriptive statistics*

- How many documents?
- How many “words”?
- Which “words” occur very frequently?
- How much lexical variation do your texts have?
  - type / token ratio: unique words vs. total words
- Average sentence length?
- Average document length?



# Corpus data understanding

- Descriptive statistics are not enough
- Explore: read some documents yourself, **look for patterns**

## Domain considerations

- **Data size**, *big? small?*
- **Private and sensitive data**,  
*e.g. military, police, healthcare, banking*
- **Ethical issues**  
*e.g. fake news promotion, user exploitation, surveillance*

- **data storage**
- **processing memory limitations**
- **type of processing**
- **available tools & resources**
- **ethical & legal constraints**



[18]

# Pre-processing text



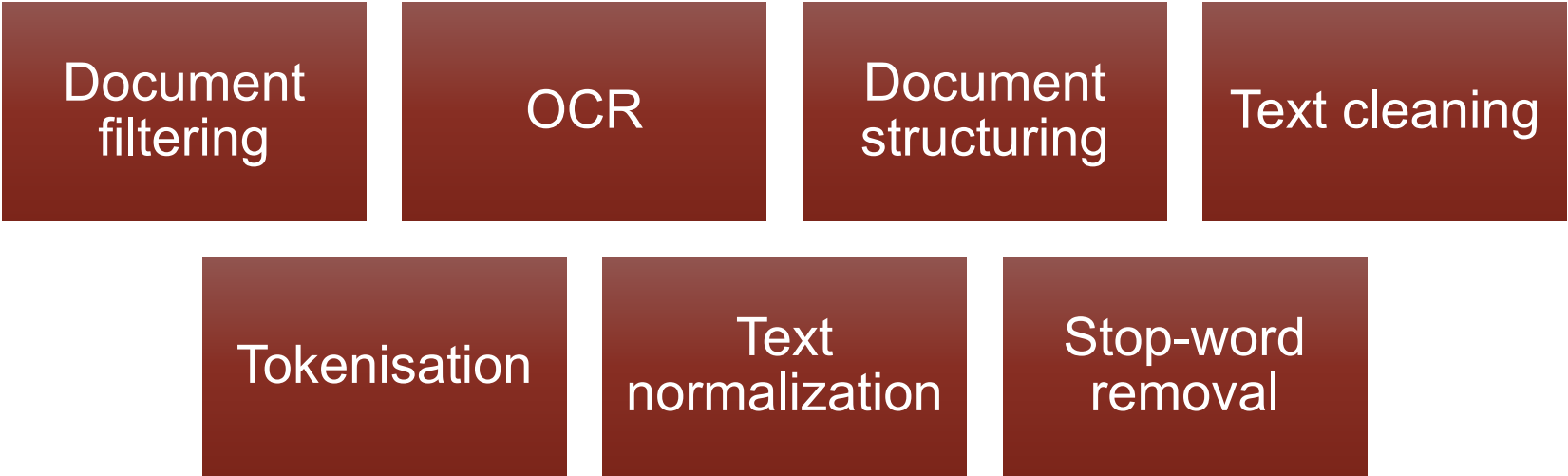
Acknowledgement: slides adopted from K. Zervanou



Jheronimus  
Academy  
of Data Science

[19]

# Pre-processing



[20]

## Document filtering

- Select relevant documents, *e.g.*
  - retrieve tweets about coronavirus using [#coronavirus](#) tag
  - retrieve wikipedia pages about TV series

## Optical Character Recognition (OCR)



CONVERT SCANNED  
TEXT IMAGES INTO TEXT



MAY INTRODUCE A LOT  
OF ERRORS

[22]

# OCR

- Mixing headlines with plain text
- Advertisements
- Image captions

Subscribe and get your weekday paper for 50% less

Page 24

THE TIMES

Wednesday February 17 2021 | thetimes.co.uk | No. 73297

£2.20 (€1.10) to subscribers

Save on the first year

£2.20 (€1.10) to subscribers

Save on the first year

Emergence in style

Essential fashion for facing the outside world

The 10 best interviews in TV history

Andrew Billen

Mass testing blitz as PM plans easing of lockdown

Millions to receive kits amid national campaign

Henry Zellman  
Chief Political Correspondent  
Chris Linstead  
Deputy Political Editor

Prime Minister Boris Johnson is drawing up plans for a testing blitz of millions of people in England as the Covid-19 lockdown is eased.

NIHS Test and Trace is preparing for a nationwide 'surge' of testing which will see more than 400,000 rapid lateral flow kits sent to homes and workplaces every day. The Times has learned. The aim is to put the country to 'a new normal'.

A campaign previously called 'Are you ready? Get testing. Go!' will begin before schools return, to encourage people to have a test even where they do not show symptoms. Ministers hope that it will be led by household names.

to other developments.

• The number of new coronavirus cases reported yesterday was 10,625, with the seven-day total down 27% per cent on the previous week. There were 799 deaths, with the seven-day total down by 25% per cent.

• The number of additional vaccine doses administered in England on Monday was 23,144. Of these 99 per cent were first doses. The total number of first doses reached 1,500,000.

• An extra 1.7 million people will be advised to shield, including 800,000 who will go to the front of the queue for a job.

• Teachers and other key workers are unlikely to be given priority for a vaccination, with the next stage of the process to be determined by age.

The prime minister will give details of the testing programme on Monday as he publishes his road map for easing the lockdown. The assumption of a surge in the use of lateral flow kits, which deliver results in 15 minutes, but are less sensitive than PCR laboratory tests, the minister has previously expressed extreme caution about lateral flow being conducted by untrained people and it is unclear whether the government has formally sought its advice.

Officials at Test and Trace are working on the assumption that schools will return on March 6, despite uncertainty, and further education will open in late April, and that hospitals, leisure and sports will follow in early May. These dates are thought to be fluid and continue to depend on the vaccine programme and infection numbers.

All schoolchildren will be offered a test a week once they go back, as will their parents and support bubbles. A similar policy is likely to apply to teachers and their close contacts.

Combined with testing for people having to leave home for work during the current lockdown, about 10 per cent of England's population will be eligible for repeat rapid tests next month, according to documents seen by The Times.

A further surge is being planned for of their shared love of horse racing and

Princess Laila bint Mohammed al Maktoum claims in a secretly recorded video to have been threatened with being shot

I'm a hostage, says Dubai ruler's daughter

David Brown, Catherine Philip

The daughter of the billionaire ruler of Dubai has accused him of keeping her hostage after she was abducted from a yacht during an escape attempt.

In a series of secretly recorded videos, Princess Laila bint Mohammed al Maktoum claims to have been threatened with being shot unless she co-operates with official statements issued by her father.

Sheikh Mohammed bin Rashid al Maktoum is the ruler of Dubai because of their shared love of horse racing and

is an influential ally of Britain in the Middle East. A US court ruled last year that he abducted Laila and her sister Princess Shamsa. The Times reported after the ruling that the Queen was to distance herself from the deaths and avoid being photographed beside him.

Laila's videos, which are being handed to the United Nations, were recorded in the bathroom of a villa close to a beach in the tourist destination. They were taken on a phone she was given about a year after her father offered to release Dubai in 2015.

This villa has been converted into a

just, she says in one recording. 'All the windows are barred shut. There's five police men outside and two police women inside. I can't even go out to get fresh air. So basically I'm a hostage.'

Lawyers for the princess said last night that they were preparing a motion to force the British government to impose sanctions against the sheikh.

Roderic Owen, QC, who presented Laila's case to the UN, urged ministers to end Britain's 'close relationship' with the United Arab Emirates to end the princess's immediate release. He added:

Continued on page 7, col 6

IN THE NEWS

China 'critical threat'

Britain increasingly sees China as a critical threat and opposes deepening economic ties with Beijing, a senior aide has said. Two little support challenging its human rights record. Page 5

Stark losses in court

Kew Star, the photographer who was once Prince Andrew's girlfriend, faces bankruptcy after being in court to win a £50,000-a-year settlement from her disbarred father. Page 5

Free-speech payouts

Universities have ended control of guest talks to promote the education secretary and as he set out plans to strengthen people who had no intention to speak with them. Page 5

Iran test for Biden

Joe Biden's administration faces an early challenge from Iran after a rocket attack on a base in Iraq. Iran has said it will continue working for the US-led coalition. Page 9

Bitcoin hits \$50,000

Bitcoin, the world's largest cryptocurrency, passed the \$50,000 mark for the first time after it was strengthened again by signs of its shift to the financial mainstream. Page 25

Root apologises

The England cricket captain Joe Root is understood to have apologised to the all-rounder Mervyn Dymally after suggesting he had 'cheated' to win the first Test in India. Page 41

[23]

www.jads.nl

JADS

Jheronimus Academy of Data Science

Uzay Kaymak

23

## Document structuring

- Identify & select document sections,  
*e.g. Abstract, Title, Conclusion*



[www.jads.nl](http://www.jads.nl)

**JADS** Jheronimus  
Academy  
of Data Science



## Text cleaning

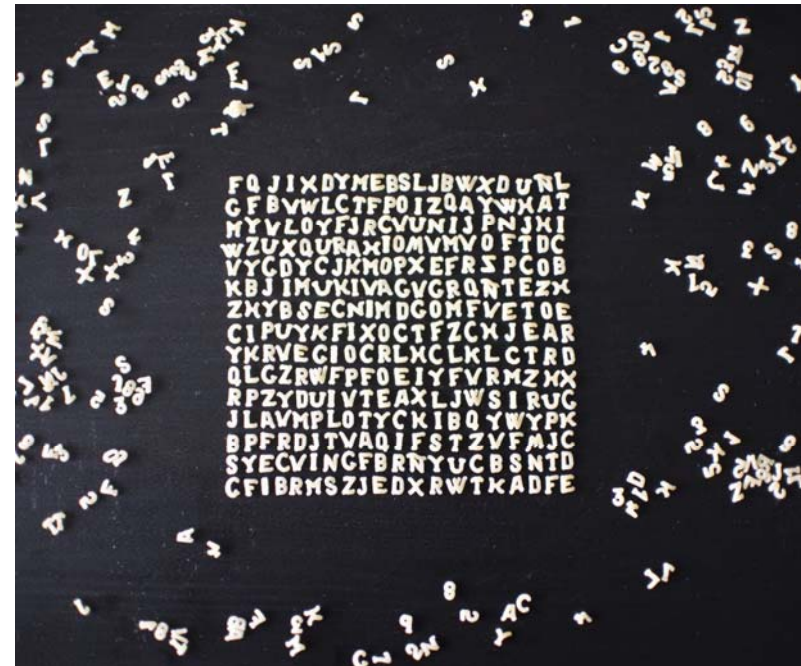
### Remove non-relevant:

- ✓ metadata (e.g. author, edit history, etc.)
- ✓ mark-up (e.g. HTML, javascript code etc.)
- ✓ headers, redundant spaces
- ✓ intervening page numbers, footnotes
- ✓ tables
- ✓ duplicate documents
- ✓ noise / non-word characters



## Text cleaning: “corrections”

- hyphenated words
- spelling correction
- OCR error correction
- convert irregular language  
e.g. abbreviations
- character encoding
- anonymise



[26]

# Tokenisation

- split text into tokens (“words”) – based on spaces & punctuation

From 1997 to 2011, the number of adults aged 18 years or older with diagnosed diabetes who reported taking diabetes medication increased for those taking either insulin, pills, or both.

Diabetes Patients Medication Status  
Source: Centers for Disease Control and Prevention (CDC)  
<https://www.cdc.gov/>

Token
From
1997
to
2011
,
the
number

## Tokenisation: *issues*

- **multi-word tokens?**
  - New York, stock exchange
- **what about punctuation?**
  - E.U., EU
  - COVID-19, Murphy's law
  - \$4.4 billion, 18.5°C, 31/03/2020...
- **Assumption: words separated by non-letters**
- Not always true **but** practical



[28]

## Text normalisation

**If needed** reduce vocabulary variation by

- ✓ removing numbers

*(but what if you need to find dates & amounts?)*

- ✓ removing punctuation & special characters (e.g. @#, -, \*, ...)

*(but what if you need to identify sentiment?)*

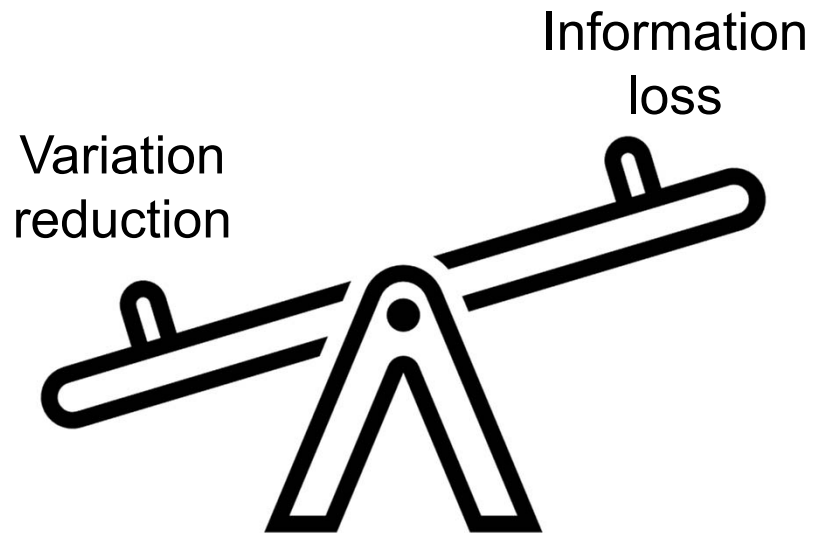
- ✓ convert into lower case

*(but what if you need to find names of people & products?)*

- ✓ lemmatization or stemming

*(but what if you unnecessarily increase ambiguity?)*

## Text normalisation



### Some points to consider

- Does my corpus have a lot of variation?
  - What is the ratio of unique tokens vs. my total token number?
- Is it likely that I lose information that I need?
- How is modelling affected by the tokens I remove or normalize?
- Do I remove important text context?

30

# Stemming

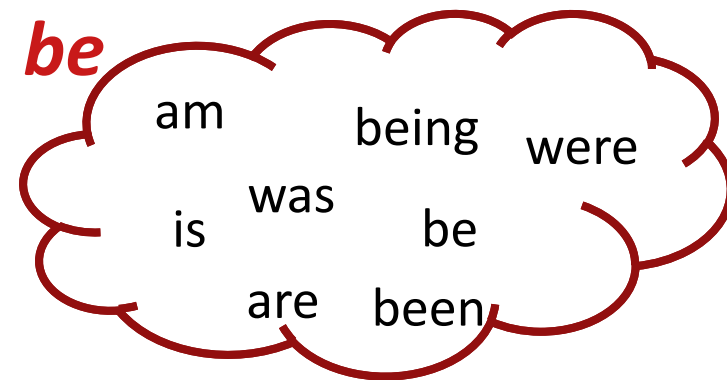
**Assumption:** a fixed number of characters ending a token are suffixes

Token	Stem
worker	work
working	work
worked	work

Token	Stem
are	ar
requirement	requir
aged	ag
afterwards	afterward

## Lemmatisation

- convert word to dictionary lemmas
- requires dictionary & part-of-speech
- result linguistically correct





## Removal of (stop) words

- Where is the information required?
- **If needed**, filter out “non-informational” text
  - function words (e.g. could, will, be, and, both, in... )?
  - Stop words? (all very common words in general language)
  - all verbs?
  - all words except nouns & adjectives?

## List of common stop words in English

a an and are as at be by for from  
has he in is it its of on that the  
to was were will with

**Figure 2.5:** A stop list of 25 semantically non-selective words which are common in Reuters-RCV1.

<https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html#p:stopwords>

## NLTK stop words

your, yours, yourself, yourselves, he, him, his, himself, she, she's, her, hers, herself, it, it's, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, that'll, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at

[34]

## Recap

### Reduce ambiguity and variation

- Ambiguity → domain application
- Variation → text normalization

### Text normalization

- removing numbers
- removing punctuation & special characters
- convert into lower case
- lemmatization or stemming

[35]

## Recap

### Pre-processing text

- Document filtering
- OCR
- Document structuring
- Text cleaning
- Tokenization
- Text normalization
- Stop word removal

[36]

# Rule-based pre-processing of text: regular expressions

Regular expression material based on slides from Jurafsky et al. (2020)

JADS Jheronimus Academy of Data Science [37]

## Regular expressions



- A formal (regular) language for specifying text strings
- How can we search for any of these?
  - woodchuck
  - woodchucks
  - Woodchuck
  - Woodchucks

[38]

# Regular Expressions: Disjunctions

- Letters inside square brackets []

Pattern	Matches
[wW]oodchuck	Woodchuck, woodchuck
[1234567890]	Any digit

- Ranges [A-Z]

Pattern	Matches	
[A-Z]	An upper case letter	<u>D</u> renched Blossoms
[a-z]	A lower case letter	<u>m</u> y beans were impatient
[0-9]	A single digit	Chapter <u>1</u> : Down the Rabbit Hole

# Regular Expressions: Negation in Disjunction

- Negations `[^Ss]`
  - Carat means negation only when first in []

Pattern	Matches	
<code>[^A-Z]</code>	Not an upper case letter	O <u>y</u> fn pripetchik
<code>[^Ss]</code>	Neither 'S' nor 's'	<u>I</u> have no exquisite reason"
<code>[^e^]</code>	Neither e nor ^	<u>L</u> ook here
<code>a\^b</code>	The pattern a carat b	Look up <u>a^b</u> now



## Regular Expressions: More Disjunction

- Woodchuck is another name for groundhog!
- The pipe | for disjunction

Pattern	Matches
<code>groundhog   woodchuck</code>	groundhog
<code>groundhog   Woodchuck</code>	Woodchuck
<code>a   b   c</code>	= <code>[abc]</code>
<code>[gG] roundhog   [Ww] oodchuck</code>	Woodchuck



# Regular Expressions: ? \*+.

Pattern	Matches	
colou?r	Optional previous char	<u>color</u> <u>colour</u>
oo*h!	0 or more of previous char	<u>oh!</u> <u>ooh!</u> <u>oooh!</u> <u>ooooh!</u>
o+h!	1 or more of previous char	<u>oh!</u> <u>ooh!</u> <u>oooh!</u> <u>ooooh!</u>
o{2}h!	Precisely 2 times previous char	<u>ooh!</u> <u>oooh!</u> <u>ooooh!</u>
baa+		<u>baa</u> <u>baaa</u> <u>baaaa</u> <u>baaaaa</u>
beg.n		<u>begin</u> <u>begun</u> <u>begun</u> <u>beg3n</u>



Stephen C. Kleene

# Regular Expressions: Anchors ^ \$

Pattern	Matches
<sup>^</sup> [A-Z]	<u>P</u> alo Alto
<sup>^</sup> [^A-Za-z]	<u>_</u> "Hello"
\. <sup>\$</sup>	The end <u>.</u>
. <sup>\$</sup>	The end <u>?</u> The end <u>!</u>

- <sup>^</sup> : starts with
- <sup>\$</sup> : ends with

[43]

## Example

- Find me all instances of the word “the” in a text.

the

Misses capitalized examples

[tT]he

Incorrectly returns other or theology

[^a-zA-Z][tT]he[^a-zA-Z]

[^a-zA-Z]?[tT]he[^a-zA-Z]

## Errors

- Note that we fixed **two kinds of errors**:
  1. Matching strings that we should not have matched  
(**there**, **then**, **other**)  
**False positives (Type I errors)**
  2. Not matching things that we should have matched (The)  
**False negatives (Type II errors)**

## Errors cont.

- In NLP we are always dealing with these kinds of errors.
- Reducing the error rate for an application often involves two antagonistic efforts:
  - Increasing accuracy or precision  
(minimizing false positives)
  - Increasing coverage or recall  
(minimizing false negatives)

## Power of regular expressions

- Regular expressions play a surprisingly large role
  - Sophisticated sequences of regular expressions are often the first model for any text processing task
- For hard tasks, we use machine learning classifiers
  - But regular expressions are still used for pre-processing, or as features in the classifiers
  - Can be very useful in capturing generalizations

## Splits and substitutions

- `txt = "The rain is another matter in the theology in Spain."`
- Split at white spaces: `re.split("\s",txt)`  
`['The', 'rain', 'is', 'another', 'matter', 'in', 'the', 'theology', 'in', 'Spain.']`
- Split when string does not contain word characters:  
`re.split("\W",txt)`  
`['The', 'rain', 'is', 'another', 'matter', 'in', 'the', 'theology', 'in', 'Spain', '']`
- Substitute a pattern with a string  
`x = re.sub("\sthe\s", " a ", t)`  
The rain is another matter in **a** theology in Spain.



## Capture Groups

- Say we want to put angles around all numbers:

*the 35 boxes* → *the <35> boxes*

- Use parens ( ) to "capture" a pattern into a numbered register (1, 2, 3...)
- Use \1 to refer to the contents of the register  
`re.sub("([0-9]+)", "<\\1>", txt)`

## Capture groups: multiple registers

the `(.*)`er they `(.*)`, the `\\1`er we `\\2`

Matches

*the faster they ran, the faster we ran*

*But not*

*the faster they ran, the faster we ate*

## But suppose we don't want to capture?

Parentheses have a double function: grouping terms, and capturing

Non-capturing groups: add a ?: after paren:

`(?:some|a few) (people|cats) like some \\1`

matches

`some cats like some cats`

but not

`some cats like some some`

## Lookahead assertions

- `(?= pattern)` is true if pattern matches, but is **zero-width; doesn't advance character pointer**
- `(?! pattern)` true if a pattern does not match
- How to match, at the beginning of a line, any single word that doesn't start with "Volcano":

```
^(?!Volcano) [A-Za-z] +
```

```
t = "Some patterns do not look like a volcano."  
print(re.search("^(?!Volcano)[A-Za-z]+", t))  
<re.Match object; span=(0, 4), match='Some'>
```

## Example with repeating numbers

- text = 'Some number 7785 and another number 34 with a digit 9.'
- pattern = '[0-9]{2}'  
print(re.findall(pattern, text))  
['77', '85', '34']  
Find two consecutive digits, continue after match
- pattern = '[0-9]{3}'  
print(re.findall(pattern, text))  
['778']  
Find three consecutive digits, continue after match
- pattern = '[0-9]{2,3}'  
print(re.findall(pattern, text))  
['778', '34']  
Find two to three consecutive digits, continue after match

**Greedy search: try to find the longest match**

[53]

## Example with repeating numbers and look ahead assertion

- text = 'Some number 7785 and another number 34 with a digit 9.'
- pattern = `'(?=([0-9]{2}))'`  
Find two consecutive digits, continue with next character  
`print(re.findall(pattern,text))`  
`['77', '78', '85', '34']`
- pattern = `'(?=([0-9]{3}))'`  
Find three consecutive digits, continue with next character  
`print(re.findall(pattern,text))`  
`['778', '785']`
- pattern = `'(?=([0-9]{2,3}))'`  
Find two to three consecutive digits, continue with next character  
`print(re.findall(pattern,text))`  
`['778', '785', '85', '34']`

**Greedy search: try to find the longest match**

[54]

## Exercise

- text = 'Some number 7785 and another number 34 with a digit 9.'
- pattern = "\\b[0-9]{2}\\b"
- Determine what the Python command  

```
print(re.findall(pattern, text))
```

  
will return
- What does \\b do in the pattern? (search on the Internet)

[55]

## Simple rule-based system: ELIZA

- Early NLP system that imitated a Rogerian psychotherapist
  - Joseph Weizenbaum, 1966.
- Uses pattern matching to match, e.g.,:
  - "I need X"

and translates them into, e.g.

- "What would it mean to you if you got X?"



## Simple Application: ELIZA

Men are all alike.

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

## How ELIZA works

- s/. \* I'M (depressed|sad) . \*/ I AM SORRY TO HEAR YOU ARE \1./
- s/. \* I AM (depressed|sad) . \*/ WHY DO YOU THINK YOU ARE \1?/
- s/. \* all . \*/ IN WHAT WAY?/
- s/. \* always . \*/ CAN YOU THINK OF A SPECIFIC EXAMPLE?/