MANCHESTER
1824
The University of Manchester

COMP34711
Week 2

Tokenisation

Goran Nenadic

Tokenisation

- **Goal**: break input into basic units = **tokens**
- What can be a token?
  - Words? Numbers? Punctuation? Emoji?

- Easy approximation is *token = "whitespace-delimited sequence"*?
- How about breaking on punctuation signs?

200g/7oz parsnips, chopped

This is language- and possibly domain-dependent!

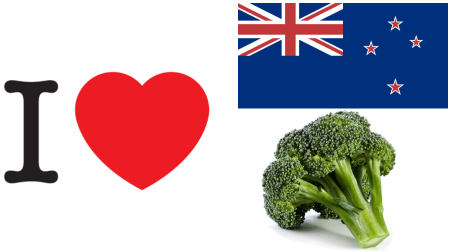
Tokenisation

- Not all languages use spaces between tokens

我喜欢新西兰花。

我 喜欢 新西兰 花。

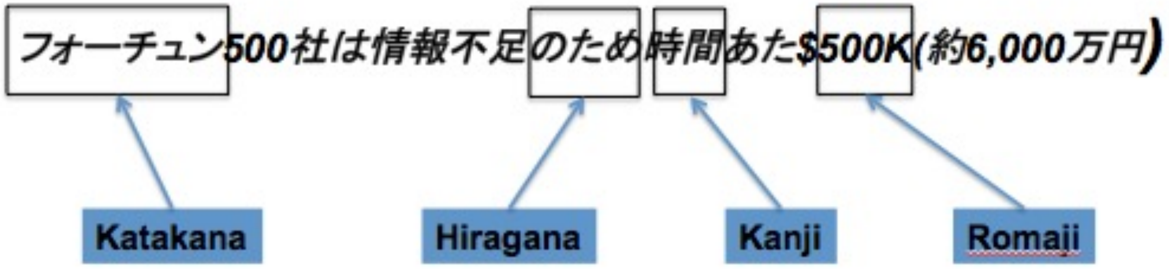
我 喜欢 新 西兰花。



Unique tokenisation not guaranteed?

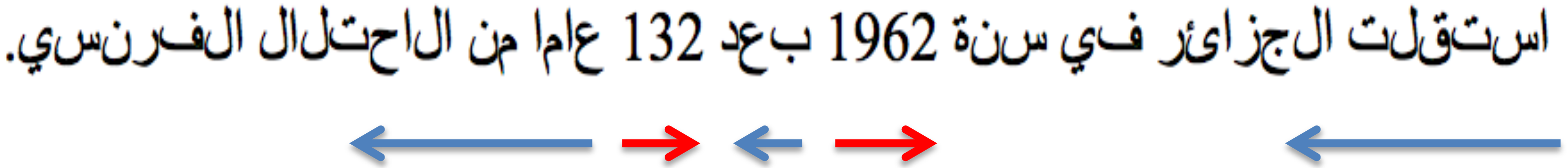
Tokenisation

- Japanese uses several alphabets…



フォーチュン500社は情報不足のため時間あた$500K(約6,000万円)

- Arabic, Hebrew: generally written right to left, but not always
  - Words separated, but complex ligatures used within words



استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

  - Right to left, but numbers written left to right

IIR book

Tokenisation – example 1

- San Francisco → one token or two?
- in spate of → one, two or three tokens?

Not all white spaces are the same!

- Hewlett-Packard → Hewlett + Packard ?
- Manchester-based → Manchester + based ?

Not all dashes are the same!

Tokenisation – example 2

Humans Process Dog and Human Facial Affect in Similar Ways

The crate was 106 Lx71 Wx79 H cm in size and appropriate for all the dogs used in this study (i.e., they could stand and move around in it comfortably)... Addi-tionally, women responded faster to positive than to negative words following positive primes (F(1,30) = 9.3, p<.01) and they responded faster to negative than to positive words following negative primes (F(1,30) = 13.3, p<.001).

Tokenisation – example 2

Humans Process Dog and Human Facial Affect in Similar Ways

The crate was 106 Lx71 Wx79 H cm in size and appropriate for all the dogs used in this study (i.e., they could stand and move around in it comfortably)… Addi-tionally, women responded faster to positive than to negative words following positive primes (F(1,30) = 9.3, p<.01) and they responded faster to negative than to positive words following negative primes (F(1,30) = 13.3, p<.001).

Tokenisation – example 3

- Packed with goodness, tasty vegetables and layered slices of potato, this hotpot is easy on the wallet too. This is designed to be a low cost recipe for all in M/cr.
- Ingredients
  - 1 tbsp olive oil
  - 4 carrots, chopped
  - 1/2 swede, chopped
  - 200g/7oz parsnips, chopped
  - 50g/1¾ oz plain flour
  - 900g/2lb potatoes, sliced into 2mm slices

Activity

Task 1: How many tokens are there in 900g/2lb

1, 2, 3, 4, 5? More?

Task 2: try Example 3 (previous slide) with tokenisers available at:

http://text-processing.com/demo/tokenize/

Tokenisation – example 4

Abbreviated forms

- what’re, I’m → what + are, I + am
- King’s coming → King + is + coming

Possessives

- King’s speech → King + ‘s + speech ?
- Finland’s capital → Finland + ’s + capital ?

Typical tokenisation steps

1. Initial segmentation
2. Handling abbreviations and apostrophes
3. Handling hyphenation
4. Dealing with (other) special expressions

Step 1: Initial segmentation

- In languages such as English, word tokens are mainly delimited by **white spaces & punctuation**

- Simple token boundary identification

  1. replace white spaces with word boundaries

  2. cut off leading and trailing quotation marks, parentheses & punctuation

- Produces reasonable performance, but there are issues with

  - **over-segmentation** (e.g. *ad hoc*)
  - **under-segmentation** (e.g. *London-based*)

Step 2: Handling abbreviations

- When a period follows an abbreviation it is an integral part of this abbreviation and should be tokenised together with it

  U.S.A. → U.S.A. (1 token)

  e.g. → e.g. (1 token)

- Previously unseen or unusual abbreviations may be problematic
  - M/cr or M’cr

Step 2: Handling abbreviations

- When a **period** (**.**) is directly attached to the previous word, it is usually a separate token which signals the end of the sentence
  - All punctuation (comma, semi-colon, quotation marks, etc.) should be a separate token

- Problem when a full stop is at end of sentence (e.g. He brought bread, milk, etc.)
  - – Is it part of the token? Should there be two full stops?

Step 2: Handling apostrophes

- Apostrophes are ambiguous as they can be used as:
  - quotative markers, e.g. *'All Quiet on the Western Front’*
  - genitive markers, e.g. *Remarque's book*
  - enclitics
    - e.g. *she's* → *she has* or *she is*

Step 2: Handling apostrophes

- **Enclitics** – abbreviated forms typically of auxiliary verbs (*be, will*) that is pronounced with so little emphasis that it is shortened and forms part of the preceding word

- ***to be***
  - *'m* in *I'm*
  - *'re* in *you're*
  - *'s* in *she's*

- auxiliary verbs
  - *'ll* in *they'll*
  - *'ve* in *they've*
  - *'d* in *you'd*

- *n’t* **(not)** as in *can’t, won’t*

Step 3: Handling hyphenation

- Different types of hyphens
    - end-of-line hyphens (?)
    - true hyphens
        - lexical hyphens
- **Lexical hyphen** is a true hyphen used in compound words which have made their way into standard vocabulary (and should be kept)

Step 3: Handling hyphenation

- **Lexical hyphens:** certain prefixes (e.g. *co–, pre–, meta–, multi–*, etc.) are often written hyphenated but they make a single token with the word after it
  - e.g. *meta-analysis* (1 token)
  - e.g. *multi-disciplinary* (1 token)
  - e.g. *self–assessment* (1 token)
- Some hyphens are not lexical

  e.g. *UK–based* (3 tokens)

Step 4: Other special expressions

e-mail addresses	xx@yahoo.co.uk
URLs	http://www.manchester.ac.uk/
telephone numbers	+44(0)161 306 0000
date & time	07/03/2013
measures	16 km/h
vehicle licence numbers	BD5I SMR
citations	(Author et al., 2012)
hashtags, emojis	#datasaveslives
. . .	

Step 4: Other special expressions

- **Numbers**
  - Decimals: 0.05, 3.4, .6
  - Language conventions for numbers:
    - English: 123,456.78
    - French: 123 456,78
    - German: 123.456,78
  - Telephone numbers: many different formats (language conventions)
- **Dates**
  - 12 Oct 2018: three tokens
  - 12/10/2018: one token? Or three? Or 5?

Tokenisation

- There are no firm rules for tokenisation – it has to be consistent with the rest of an NLP system
- Tokenisation: **knowing when to split, not when to combine**
  - Avoid over-segmentation
- Tokenisation is only the first step – should be simple, but will affect the rest of the steps.

Text pre-processing

Preparing text for further processing

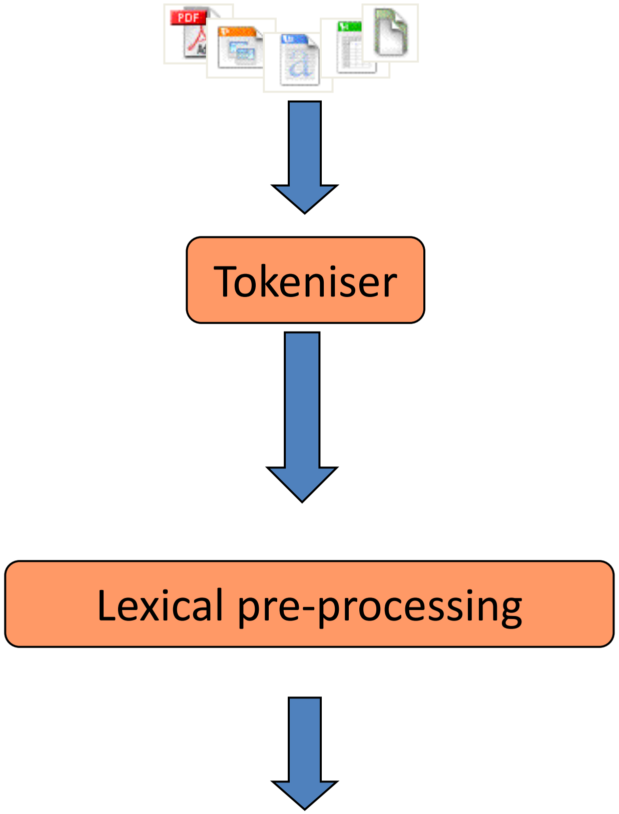
(week 2)

1) Document-level preparation
   - Document conversion
   - Language/domain identification

2) Tokenisation
   - Case folding

(week 4)

3) Basic lexical pre-processing
   - Lemmatization
   - Stemming
   - Spelling corrections



Case folding

- Convert everything to lowercase?
- What are we gaining and what are we loosing?
- Many proper nouns are derived from common nouns and so are distinguished only by case, including companies (General Motors, The Associated Press), government organizations (the Fed vs. fed) and person names (Bush, Black).
- Should we remove other orthographics (e.g. dots in acronyms):
  - C.A.T. -> CAT -> cat

Case folding

- Just lowercase some tokens?
  - Lowercase words at the beginning of a sentence/title
  - Leave mid-sentence capitalized words as capitalized?
  - This is known as **truecasing**
- We can learn (e.g. by a machine learning sequence model) when to case-fold
- May be useful for some applications, but not for all
  - e.g. *search engines*: users usually use lowercase regardless of the correct case of words. Thus, lowercasing everything seems a practical solution.
  - e.g. *identification of entity name (e.g. organization names, or people names):* preserving capitals would makes sense

Case folding

- How to deal with special characters
  - Ligatures?
  - Accents?
  - Emojis?

ꜵ

â

- What do your users type?
  - E.g. in a search engine? Even if normally use accents, they may not type them in query