MACQUARIE University

COMP3420/6420 AI for Text and Vision

Week 7 – Simple Text Processing

Dr Yuankai Qi

Outline

- Why do we learn text processing?
- Tasks in text processing
- Challenges in text processing
- Basics in text processing

Brief Introduction

- **Research interests**
  - Medical Image Processing
  - Large Language Model

Brief Introduction

MACQUARIE University

- **Research interests**
  - Robots

Brief Introduction

MACQUARIE University

- **How to contact me?**

  For private message:

  yuankai.qi@mq.edu.au

  For public questions: ilearn forum

  I will check regularly, and reply asap

Why do we learn text processing?

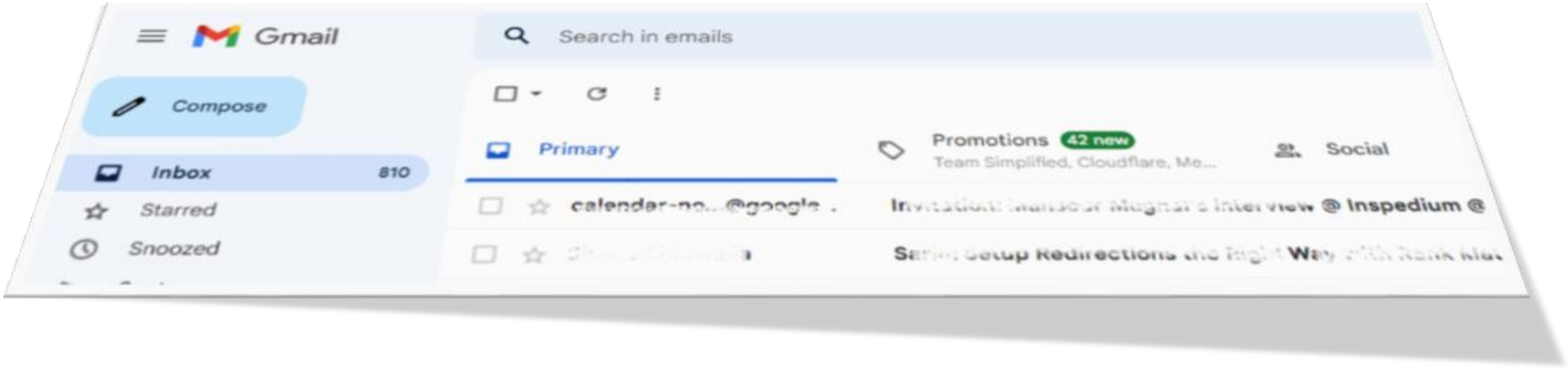
Applications

- **Sentiment analysis**

  - ❑ Customer feedback
  - ❑ Social Media Mon

Applications

- **Text Classification**
  - ❑ Spam detection (spam/focused/social/promotions/…)
  - ❑ Content categorization



Applications

MACQUARIE University

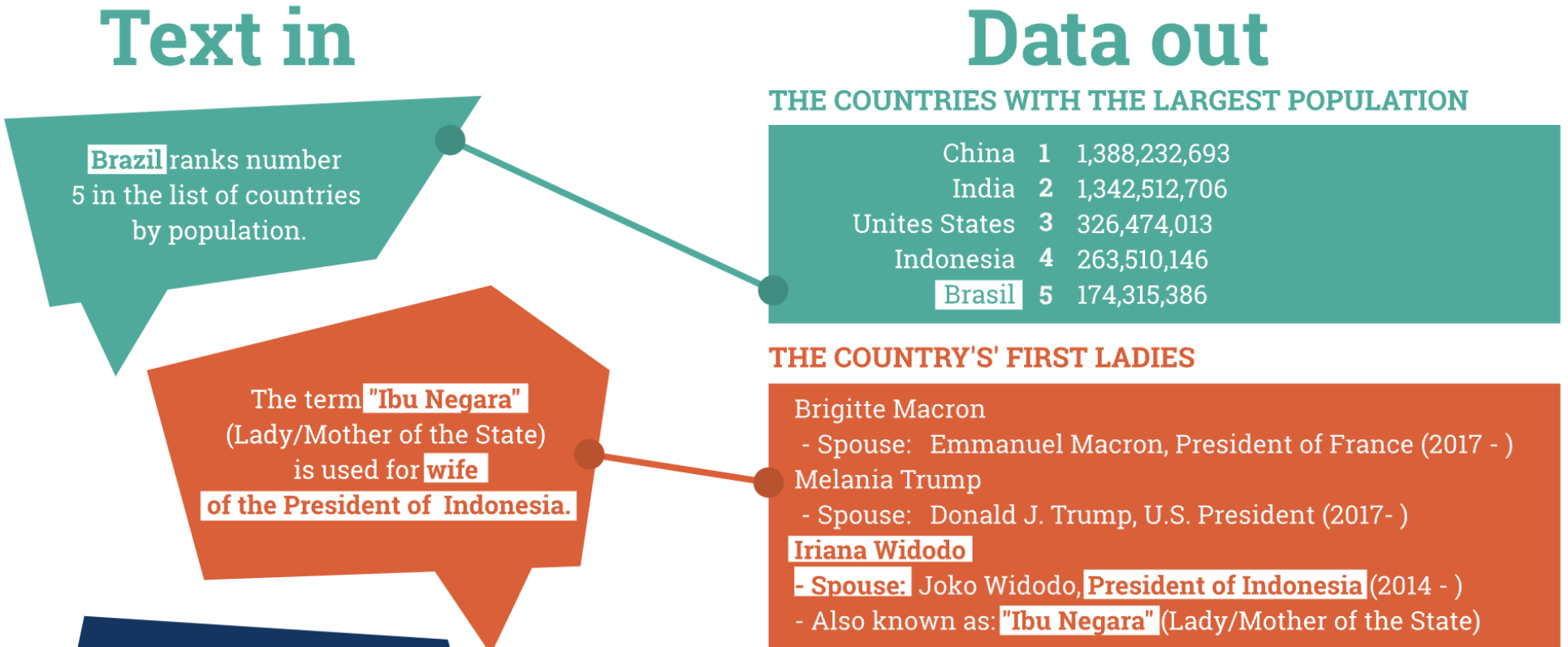
- **Named Entity Recognition (NER)**
  - ❑ People
  - ❑ Locations
  - ❑ Organizations
  - ❑ Dates
  - ❑ Phone number
  - ❑ Bank account

Elon Musk PERSON apparently wasn't aware that his company SpaceX had a Facebook ORG page. The SpaceX and Tesla PRODUCT CEO has responded to a comment on Twitter GPE calling for him to take down the SpaceX, Tesla and Elon Musk ORG official pages in support of the #deletefacebook movement by first ORDINAL acknowledging he didn't know one existed, and then following up with promises that he would indeed take them down.

He's done just that, as the SpaceX NORP Facebook page is now gone, after having been live earlier today DATE (as you can see from the screenshot included taken at around 12:10 PM ET) TIME .

Applications

- **Information Extraction**
  - ❑ Events
  - ❑ Relationships



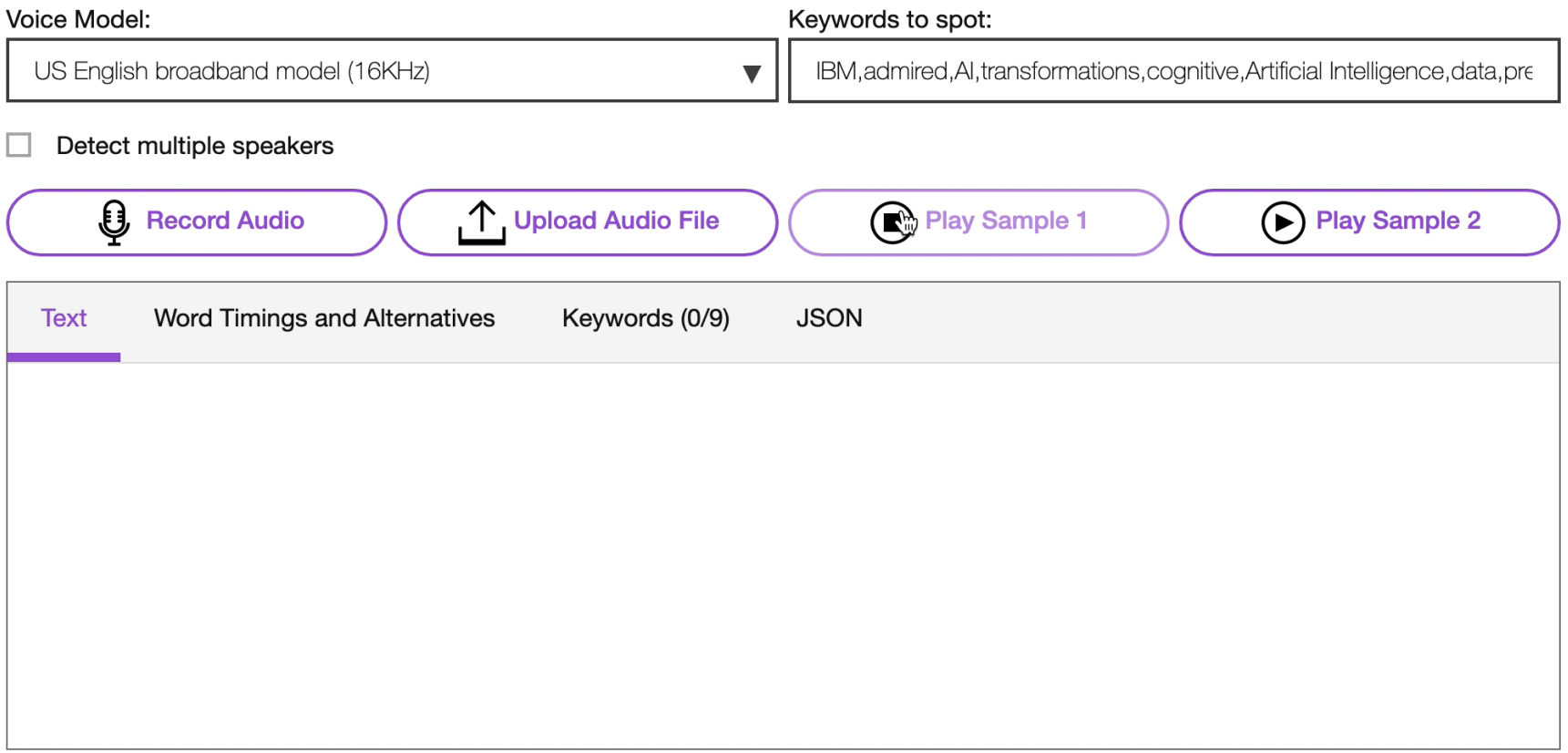
Applications

MACQUARIE University

- **Speech Recognition** audio-> text

Challenges:
- ❑ Reduced form
- ❑ Connected speech
- ❑ Accent

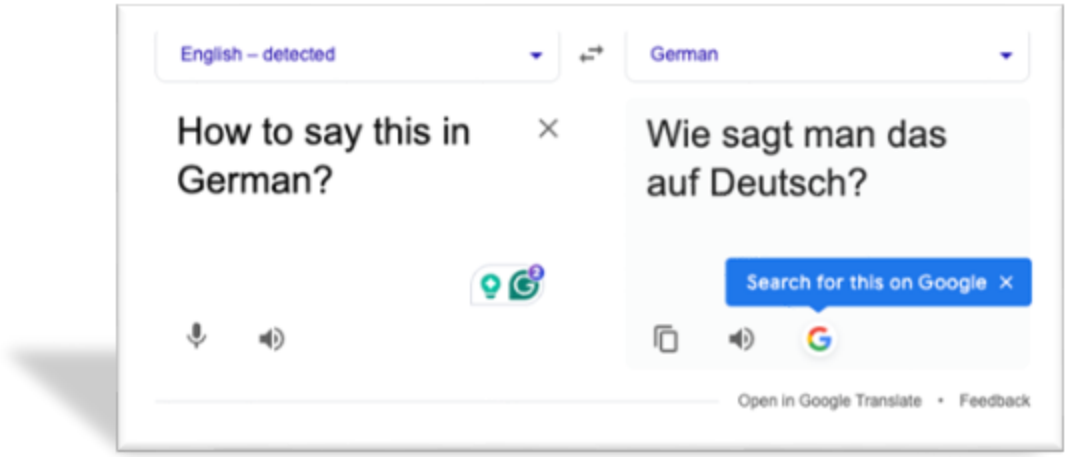
How to overcome
- ❑ Grammar
- ❑ Syntax



Applications

- **Document Summarization**

- **Machine Translation**



Applications

- **Chat Robot and Virtual Assistant**
  - ChatGPT, Claude
  - Alexa, Google Assistant, Siri

Tasks in Text/NLP

MACQUARIE University

Phonetics and phonology	The study of language sounds
Ecology	The study of language conventions for punctuation, text mark-up and encoding
Morphology	The study of meaningful components of words
Syntax	The study of structural relationships among words
Lexical semantics	The study of word meaning
Compositional semantics	The study of the meaning of sentences
Pragmatics	The study of the use of language to accomplish goals
Discourse conventions	The study of conventions of dialogue

Tasks in Text/NLP

MACQUARIE University

- Phonetics, Phonology
  - Pronunciation Modeling

SOUNDS	Th	i	a	si	e	n

Tasks in Text/NLP

MACQUARIE University

- Word
  - Language Modeling
  - Tokenization
  - Spelling correction

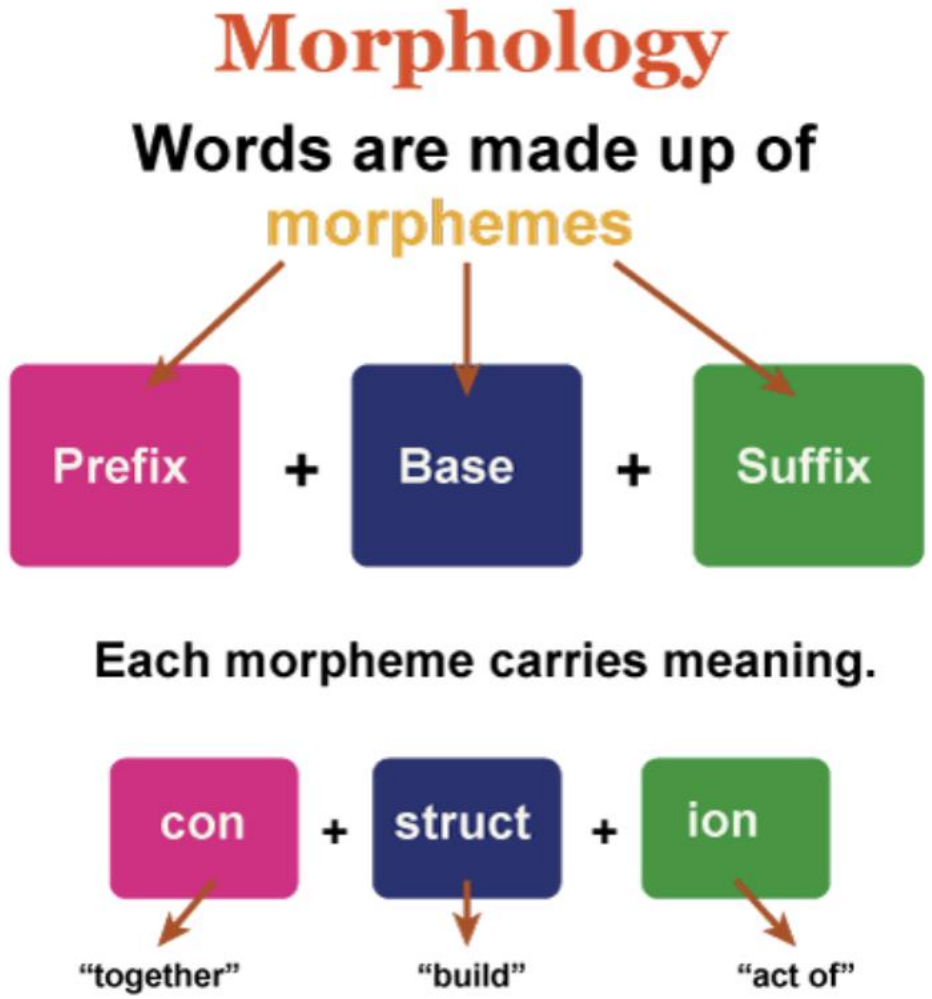
Tasks in Text/NLP

- Morphology
  - Morphology analysis
  - Tokenization
  - Lemmatization

He runs far away

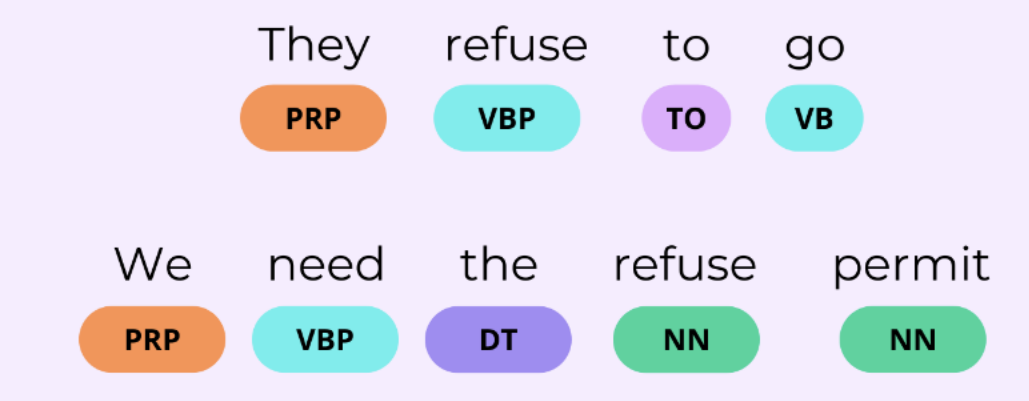
Run

3sg



Tasks in Text/NLP

- Part of Speech

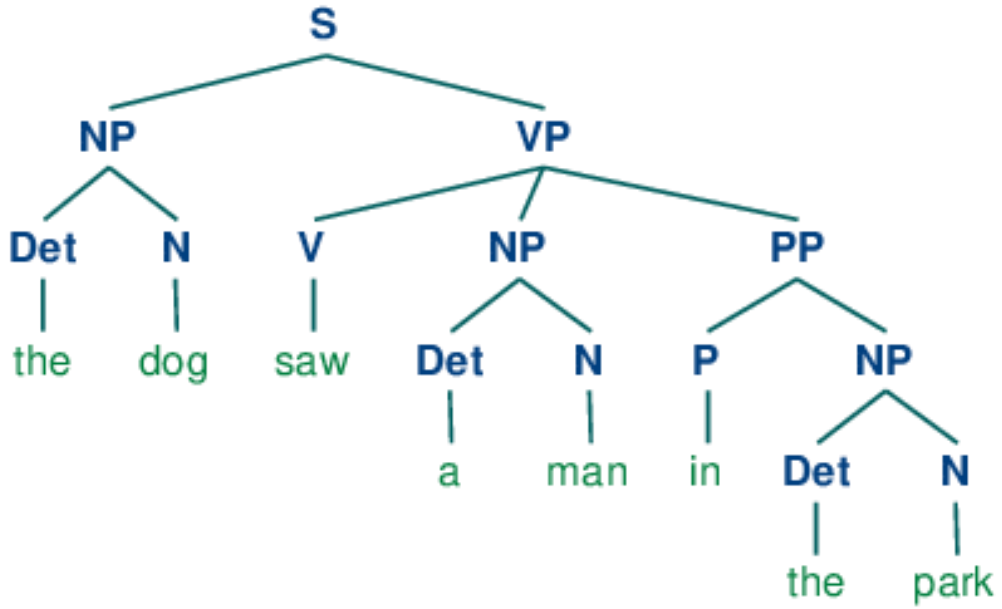


Tag	Description	Example	Tag	Description	Example
PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
RB	adverb	*quickly*	WP$	wh-possess.	*whose*
RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
RBS	superlatv. adverb	*fastest*	$	dollar sign	*$*
RP	particle	*up, off*	#	pound sign	*#*
SYM	symbol	*+,%, &*	“	left quote	‘ or “
TO	“to”	*to*	”	right quote	’ or ”
UH	interjection	*ah, oops*	(	left paren	[, (, {, <
VB	verb base form	*eat*	)	right paren	],), }, >
VBD	verb past tense	*ate*	,	comma	,
VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

Tasks in Text/NLP

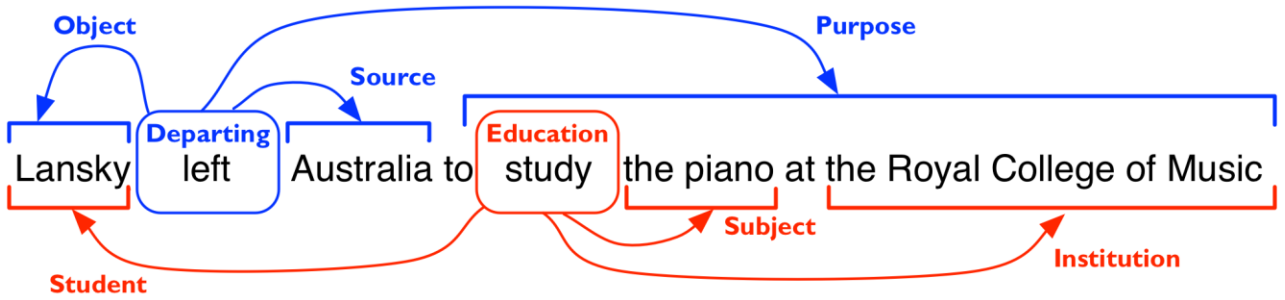
MACQUARIE University

- Syntax Parsing



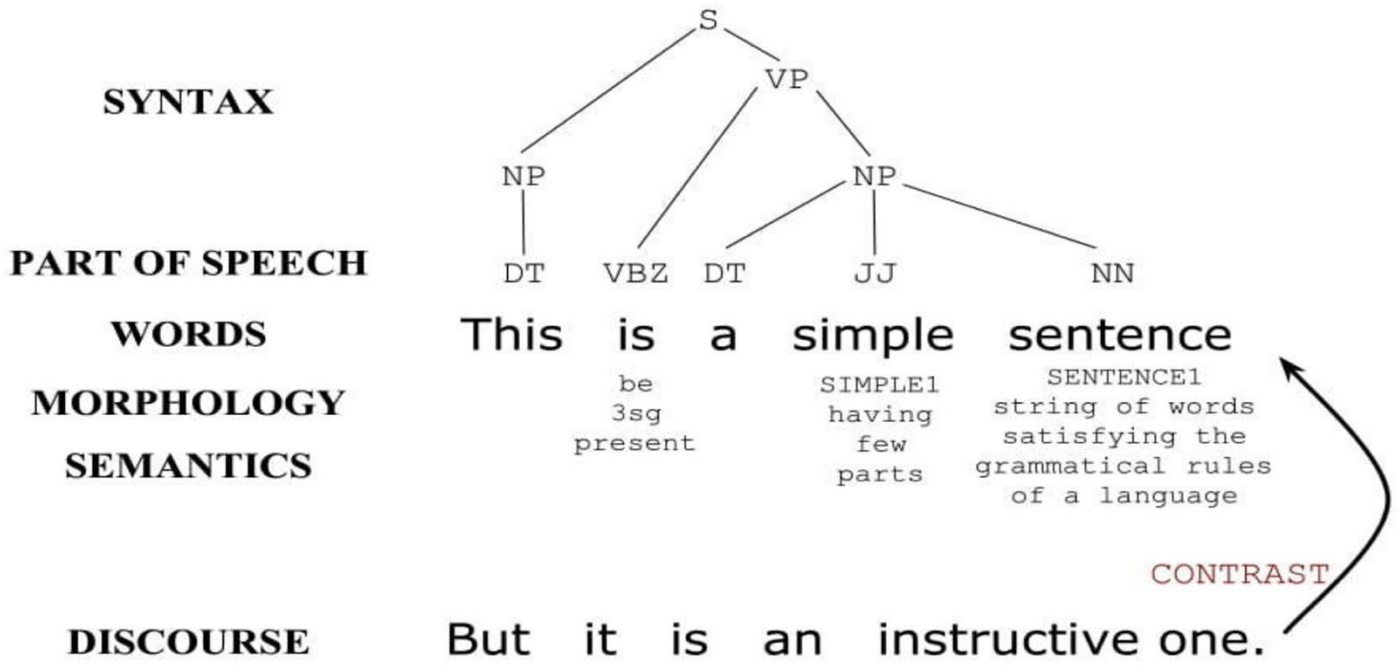
Tasks in Text/NLP

- Semantic
  - Named entity recognition
  - Word sense disambiguation
  - Semantic role labeling



Tasks in Text/NLP

- Discourse



MACQUARIE University

Why is understanding text challenging?

Why understanding text is challenging?

MACQUARIE University

- Ambiguity
- Sparsity
- Variation
- Expressivity
- Unknow representations

...

Why understanding text is challenging?

MACQUARIE University

- **Ambiguity at multiple levels**

Word senses: bank (finance or river?)
Part of speech: chair (noun or verb?)
Syntactic structure: I can see a man with a telescope

Why understanding text is challenging?

MACQUARIE University

“We saw the woman with the telescope wrapped in paper”

- ❑ Who has the telescope?
- ❑ Who or what is wrapped in paper?
- ❑ An event of perception or an assault?

Why understanding text is challenging?

MACQUARIE University

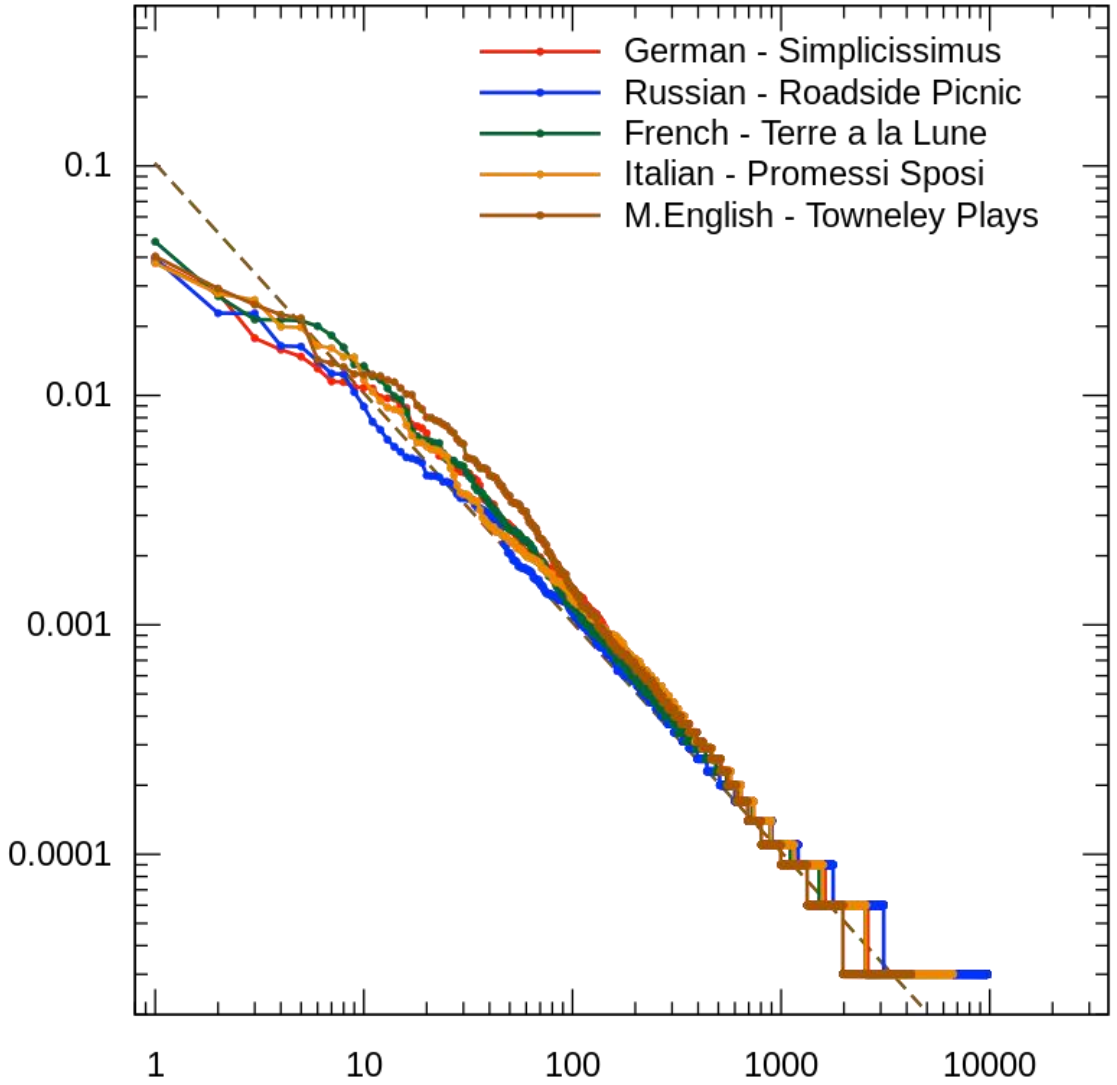
- **Sparsity**
  - **Zipf's Law**
  - **Corpus**

- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
- Examples
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of French/English sentences
  - Yelp reviews
  - The Web!

Why understanding text is challenging?

- **Zipf's Law**

The frequencies of certain words are inversely proportional to their ranks
--*George Kingsley Zipf*



Why understanding text is challenging?

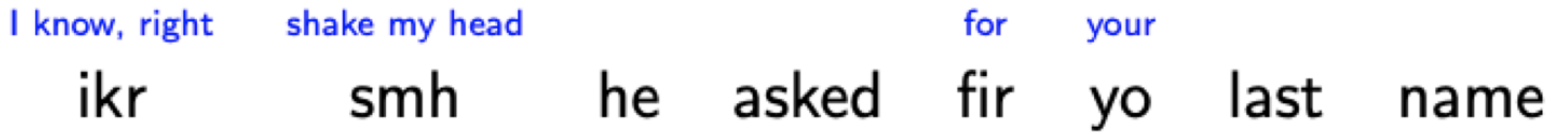
- **Morphological variation**

ikr smh he asked fir yo last name

so he can add u on fb lolololol

Why understanding text is challenging?

- **Morphological variation**



Why understanding text is challenging?

- **Morphological variation**

你好

Hello

Arabic

Why understanding text is challenging?

MACQUARIE University

- **Expressivity**

  - Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:
    - *She gave the book to Tom* vs. *She gave Tom the book*
    - *Some kids popped by* vs. *A few children visited*
    - *Is that window still open?* vs. *Please close the window*

Why understanding text is challenging?

MACQUARIE University

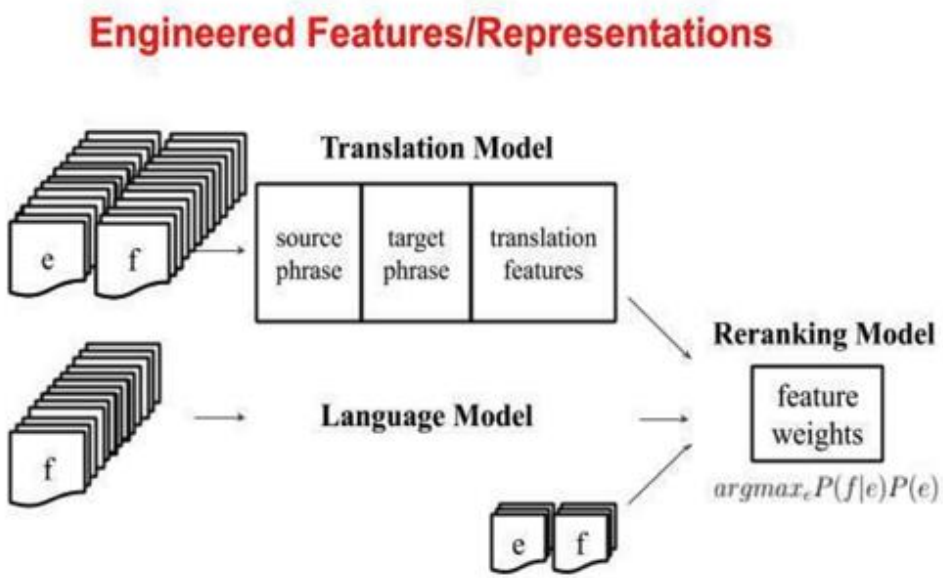
- **Unknown representations**

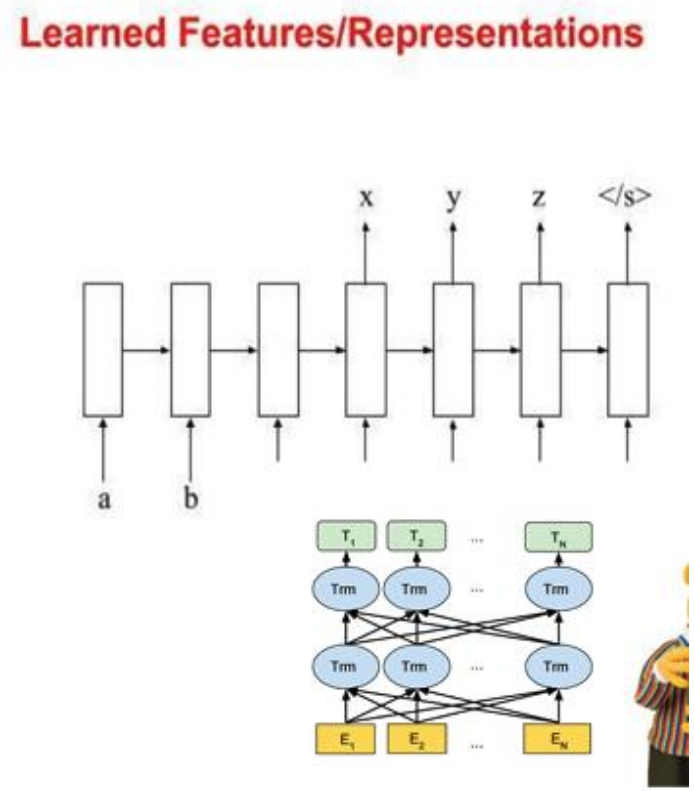
C++
Python
Java
R

V?

NLP Methods Overview

- **Probabilistic and Connectionist Methods**





Factors Changing NLP Landscape

MACQUARIE University

- Increases in computing capability
- The rise of the web and social web
- Advances in machine learning / artificial intelligence

MACQUARIE University

Basics in Text Processing

Basics in Text Processing

MACQUARIE University

- Text Encoding
  - Computers work in binary code: 0s and 1s
  - Each company has its own encoding systems
    - ❑ Hampers communication across apps
  - A standard encoding: ASCII (American Standard Code for Information Interchange)
    - ❑ 8-bit for one letter or a punctuation mark

Character	Decimal Number	Binary Number		Character	Decimal Number	Binary Number
blank space	32	0010 0000		^	94	0101 1110
!	33	0010 0001		-	95	0101 1111
“	34	0010 0010		`	96	0110 0000
#	35	0010 0011		a	97	0110 0001
$	36	0010 0100		b	98	0110 0010

Basics in Text Processing

MACQUARIE University

- Text Encoding
  - Computers work in binary code: 0s and 1s
  - Each company has its own encoding systems
    - ❑ Hampers the communication across apps
  - A standard encoding: ASCII (American Standard Code for Information Interchange)
    - ❑ 8-bit for one letter or a punctuation mark
    - ❑ cat: 0110 0011, 0110 0001, 0111 0100

Basics in Text Processing

MACQUARIE University

- Text Encoding

IBM and Microsoft worked together, and defined:

Code Page 1252 Most of western Europe and parts of Africa, except for Dutch and Slovene which were hackily-supported.

Code Page 1253 Greek

Code Page 1251 Cyrillic (e.g. Russian)

In 1988 the International Standards Organisation establised ISO 8859-1 "Latin-1" encoding.

Basics in Text Processing

MACQUARIE University

- Text Encoding
- Microsoft: UTF-16

> "There are less than 65,536 different letters in the world. So we can encode everything in 16 bits."

- Chinese: GB18030-2022

> *The GB 18030-2022 standard defines 87,887 characters.*
> (More on GB 18030-2022 later)

Basics in Text Processing

MACQUARIE University

- Text Encoding
- Varying-width encoding

> Simple and common letters (e.g. ASCII characters) should take up fewer bytes than rarely-used symbols from exotic languages.
>
> UTF-16 is *now* a varying-width encoding, negating the whole point of using 16-bit code points.
>
> UTF-16 is the default on Windows systems (filenames, documents, network protocols etc.)
>
> Most JavaScript engines use UTF-16

Basics in Text Processing

MACQUARIE University

- Text Encoding
- UTF-8

1993: Ken Thompson and Rob Pike proposed a new encoding called UTF-8.

"If we have to have varying-width encodings anyway, why not be compatible with ASCII, and generally work in 8-bits?"

UTF-8 is the default on Linux, OSX and is an optional component to install on Windows.

Basics in Text Processing

MACQUARIE University

- Text Encoding

> The golden rule: a stream of bytes is useless without knowing the encoding.

ASCII	Simple, limited, upwardly compatible with UTF-8
EBCDIC	Ancient IBM technology.
CP1252	A *code page* for Western European languages
CPxxxx	Go and look up what the Microsoft code page is for your language: old PCs will still use it
ISO-8859-1/Latin-1	Closely related format
GB18030	Mandated Chinese format
UTF-16	Modern Windows default, encodes for Unicode
UTF-8	Where the world will end up, encodes for Unicode

Basics in Text Processing

MACQUARIE University

- Common NLP libraries

NLTK	The easiest to learn, good for teaching. We'll use this a lot. `www.nltk.org`
spaCy	What you are more likely to use in a job. `https://spacy.io`
scikit-learn	Has some text processing capabilities
keras_nlp	Deep learning and neural networks for text processing

Basics in Text Processing

MACQUARIE University

- What is NLTK?
  - ❑ Natural Language Toolkit
  - ❑ http://www.nltk.org
- Install NLTK
  - ❑ Pre-installed in Anaconda
  - ❑ pip install nltk
  - ❑ conda install nltk
- But, you need to use nltk.download() to fetch corpora and models.
  - ❑ punkt
  - ❑ wordnet
  - ❑ gutenberg

Basics in Text Processing

MACQUARIE University

- Gutenberg

  - Oldest digital library (1971)
  - 70,000 free books (HTML, EPUB)
  - Mostly books where copyright has expired

Basics in Text Processing

MACQUARIE University

- Using Gutenberg sample data

```
import nltk
nltk.download('gutenberg')
for id in nltk.corpus.gutenberg.fileids():
    print(id)
```

Output:

```
[nltk_data] Downloading package gutenberg to /home/gregb/nltk_data...
[nltk_data]   Unzipping corpora/gutenberg.zip.
austen-emma.txt
austen-persuasion.txt
austen-sense.txt
bible-kjv.txt
blake-poems.txt
```

Basics in Text Processing

MACQUARIE University

- Lexico-statistics

```
>>> import nltk
>>> emma = nltk.corpus.gutenberg.words('austen-emma.txt')
>>> len(emma)
192427
>>> emma[:10]
['[', 'Emma', 'by', 'Jane', 'Austen', '1816', ']', 'VOLUME'
>>> import collections
>>> emma_counter = collections.Counter(emma)
>>> emma_counter.most_common(10)
[(',', 11454), ('.', 6928), ('to', 5183), ('the', 4844),
('and', 4672), ('of', 4279), ('I', 3178), ('a', 3004),
('was', 2385), ('her', 2381)]
```

Basics in Text Processing

MACQUARIE University

- Lexico-statistics

Exercise

1. Find the most frequent word with length of at least 7 characters.
2. Find the words that are longer than 7 characters and occur more than 7 times.

Basics in Text Processing

MACQUARIE University

- Lexico-statistics
  - N-gram: a sequence of N words

❑ 1-gram (unigram), 2-gram (bigram), 3-gram (trigram), 4-gram (quadgram)

```
>>> list(nltk.bigrams([1,2,3,4,5,6]))
[(1, 2), (2, 3), (3, 4), (4, 5), (5, 6)]
>>> list(nltk.bigrams(emma))[:3]
[('[', 'Emma'), ('Emma', 'by'), ('by', 'Jane')]
```

Basics in Text Processing

MACQUARIE University

- Lexico-statistics
  - N-gram: a sequence of N words

❑ 1-gram (unigram), 2-gram (bigram), 3-gram (trigram), 4-gram (quadgram)

```
>>> list(nltk.ngrams(emma,4))[:5]
[('[', 'Emma', 'by', 'Jane'),
 ('Emma', 'by', 'Jane', 'Austen'),
 ('by', 'Jane', 'Austen', '1816'),
 ('Jane', 'Austen', '1816', ']'),
 ('Austen', '1816', ']', 'VOLUME')]
```

Basics in Text Processing

MACQUARIE University

- Lexico-statistics
  - N-gram: a sequence of N words

❑ 1-gram (unigram), 2-gram (bigram), 3-gram (trigram), 4-gram (quadgram)

Exercise

1. Find the most frequent bigram.
2. Find the most frequent bigram that begins with "the".

Basics in Text Processing

MACQUARIE University

- Sort
  - The function **sorted()** returns a sorted copy.
  - Sequences can be sorted in place with the sort() method.
  - Python 3 does not support sorting of lists with mixed contents.

```
>>> foo = [2,5,9,1,11]
>>> sorted(foo)
[1, 2, 5, 9, 11]
>>> foo
[2, 5, 9, 1, 11]
>>> foo.sort()
>>> foo
[1, 2, 5, 9, 11]
>>> foo2 = [2,5,9,1,'a']
>>> sorted(foo2)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: unorderable types: str() < int()
```

Basics in Text Processing

MACQUARIE University

- Sort

```
>>> l = ['a','abc','b','c','aa','bb','cc']
>>> sorted(l)
['a', 'aa', 'abc', 'b', 'bb', 'c', 'cc']
>>> sorted(l,key=len)
['a', 'b', 'c', 'aa', 'bb', 'cc', 'abc']
>>> sorted(l,key=len,reverse=True)
['abc', 'aa', 'bb', 'cc', 'a', 'b', 'c']
>>> sorted(l,key=lambda x: -len(x))
['abc', 'aa', 'bb', 'cc', 'a', 'b', 'c']
```

Basics in Text Processing

MACQUARIE University

- Sort

You're given data of the following form:

```
namedat = dict()
namedat['mc'] = ('Madonna', 45)
namedat['sc'] = ('Steve', 41)
```

1. How would you print a list ordered by name?

```
('Madonna', 45)
('Steve', 41)
```

2. How would you print out a list ordered by age?

```
('Steve', 41)
('Madonna', 45)
```

Basics in Text Processing

MACQUARIE University

- String in Python
  - String is a base datatype.
  - Strings are sequences and can use operations like:
    - foo = "A string"
    - **len**(foo)
    - foo[0]
    - foo[0:3]
    - multifoo = """*A multiline* string"""
  - In addition, there are some utility functions in the string module.

Basics in Text Processing

MACQUARIE University

- String in Python

```
>>> "my string".capitalize()
'My string'
>>> "my string".upper()
'MY STRING'
>>> "My String".lower()
'my string'
>>> a = "my string with my other text"
>>> a.count('my')
2
>>> a.find('with')
10
>>> a.find('nothing')
-1
```

Basics in Text Processing

MACQUARIE University

- String in Python
  - split (sep) is a central string operation.
  - It splits a string wherever sep occurs (blank space by default).
  - It is either a function in the string module or a method of string objects.

```
>>> foo="one :: two :: three"
>>> foo.split()
['one', '::', 'two', '::', 'three']
>>> foo.split('::')
['one ', ' two ', ' three']
>>> import string
>>> string.split("this is a test")
['this', 'is', 'a', 'test']
```

Basics in Text Processing

MACQUARIE University

- String in Python
  - Join is another useful function/method in the string module.
  - It takes a list and joins the elements using some delimiter.

```
>>> text="this is some text to analyse"
>>> words=text.split()
>>> words.sort()
>>> print(",".join(words))
analyse,is,some,text,this,to
>>> print("".join(words))
analyseissometextthisto
```

Basics in Text Processing

MACQUARIE University

- String in Python

```
def censor(text):
    'replace bad words in a text with XXX'
    badwords = ['poo', 'bottom']
    for b in badwords:
        text = text.replace(b, 'XXX')
    return text
```

Basics in Text Processing

MACQUARIE University

- NLTK tools

Some NLTK tools that are useful for text pre-processing are:

- word_tokenize(text)
- sent_tokenize(text)
- pos_tag(tokens)
- pos_tag_sents(sentences)
- PorterStemmer()

Basics in Text Processing

MACQUARIE University

- NLTK tools

  - ❑ Tokenization

    - NLTK can split text into sentences and words.
      - Sentence segmentation splits text into a list of sentences.
      - Word tokenisation splits text into a list of words (tokens).
    - NLTK's default word tokeniser works best after splitting the text into sentences.

Basics in Text Processing

MACQUARIE University

- NLTK tools

```
In [1]: import nltk

In [2]: text = "This is a sentence. This is another sentence

In [3]: nltk.sent_tokenize(text)
Out[3]: ['This is a sentence.', 'This is another sentence.']

In [4]: for s in nltk.sent_tokenize(text):
   ...:     for w in nltk.word_tokenize(s):
   ...:         print(w)
   ...:     print()
```

Basics in Text Processing

MACQUARIE University

- NLTK tools
  - ❑ Part of Speech Tagging

    - Often it is useful to know whether a word is a noun, or an adjective, etc. These are called parts of speech.
    - NLTK has a part of speech tagger that tags a list of tokens.
    - The default list of parts of speech is fairly detailed but we can set a simplified version (called universal by NLTK).

Basics in Text Processing

MACQUARIE University

- NLTK tools
  - ❑ Part of Speech Tagging

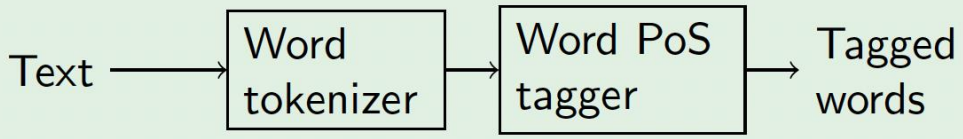
Tag	Meaning	English Examples
ADJ	adjective	new, good, high, special, big, local
ADP	adposition	on, of, at, with, by, into, under
ADV	adverb	really, already, still, early, now
CONJ	conjunction	and, or, but, if, while, although
DET	determiner, article	the, a, some, most, every, no, which
NOUN	noun	year, home, costs, time, Africa
NUM	numeral	twenty-four, fourth, 1991, 14:24
PRT	particle	at, on, out, over per, that, up, with
PRON	pronoun	he, their, her, its, my, I, us
VERB	verb	is, say, told, given, playing, would
.	punctuation marks	. , ; !
X	other	ersatz, esprit, dunno, gr8, univeristy

Basics in Text Processing

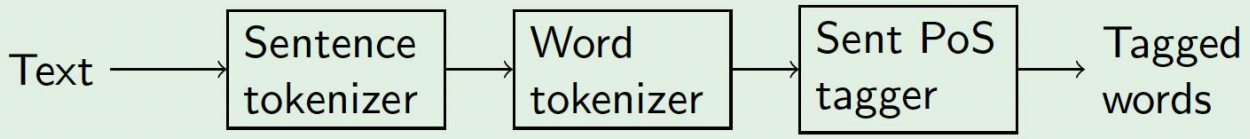
MACQUARIE University

- NLTK tools
  - ❑ Part of Speech Tagging

NLP pipeline for Word PoS tagging in NLTK



NLP pipeline for Sentence PoS tagging in NLTK



Note that the above pipelines may differ in other environments.

Basics in Text Processing

MACQUARIE University

- NLTK tools
  - ❑ Part of Speech Tagging

```
In [28]: nltk.pos_tag(["this", "is", "a", "test"])
Out[28]: [('this', 'DT'), ('is', 'VBZ'), ('a', 'DT'), ('test', 'NN')]

In [29]: nltk.pos_tag(["this", "is", "a", "test"], tagset="universal")
Out[29]: [('this', 'DET'), ('is', 'VERB'), ('a', 'DET'), ('test', 'NOUN')]

In [30]: nltk.pos_tag(nltk.word_tokenize("this is a test"), tagset="universal")
Out[30]: [('this', 'DET'), ('is', 'VERB'), ('a', 'DET'), ('test', 'NOUN')]
```

Basics in Text Processing

MACQUARIE University

```
Out[31]: 'This is a sentence. This is another sentence.'

In [34]: text_sent_tokens = [nltk.word_tokenize(s) for s in nltk.sent_tokenize(t
    ...: ext)]

In [35]: text_sent_tokens
Out[35]:
[['This', 'is', 'a', 'sentence', '.'],
 ['This', 'is', 'another', 'sentence', '.']]

In [38]: nltk.pos_tag_sents(text_sent_tokens, tagset="universal")
Out[38]:
[[('This', 'DET'),
  ('is', 'VERB'),
  ('a', 'DET'),
  ('sentence', 'NOUN'),
  ('.', '.')],
 [('This', 'DET'),
  ('is', 'VERB'),
```

Basics in Text Processing

MACQUARIE University

- NLTK tools

  ❑ Stemming

  What is stemming?

```
In [46]: s = nltk.PorterStemmer()

In [47]: s.stem("books")
Out[47]: 'book'

In [48]: s.stem("is")
Out[48]: 'is'

In [50]: s.stem("runs")
Out[50]: 'run'

In [51]: s.stem("running")
Out[51]: 'run'

In [52]: s.stem("run")
Out[52]: 'run'

In [53]: s.stem("goes")
Out[53]: 'goe'
```

Basics in Text Processing

MACQUARIE University

- NLTK tools
  - ❑ Stemming
    - Often it is useful to remove information such as verb form, or the difference between singular and plural.
    - NLTK offers stemming, which removes suffixes.
      - The Porter stemmer is a popular stemmer.

Basics in Text Processing

MACQUARIE University

- Exercise

1. What is the sentence with the largest number of tokens in Austen's "Emma"?
2. What is the most frequent part of speech in Austen's "Emma"?
3. What is the number of distinct stems in Austen's "Emma"?
4. What is the most ambiguous stem in Austen's "Emma"? (meaning, which stem in Austen's "Emma" maps to the largest number of distinct tokens?)

Before you go

- **The practicals will help your assignments**
-

Complete the post-lecture survey on iLearn

- What was the most important thing you learned in today's lecture?
- What question remains uppermost in your mind at the end of today's lecture?
- What was the "muddiest point" in today's lecture?

Basics in Text Processing

- Acknowledgments
  - Diego
  - Greg
  - Diyi Yang