P21 INTRODUCTION TEXTMINING

AGENDA

Natural Language Processing
Information extraction architecture
Topic modelling

Textmining toolstacks

- in python
  - NLTK
  - spaCy
  - notebook examples
- in R
  - tm package
  - tidytext

NATURAL LANGUAGE PROCESSING

The term Natural **Language Processing** encompasses a broad set of techniques **for automated generation, manipulation and analysis of natural or human languages**.

Although most NLP techniques inherit largely from Linguistics and Artificial Intelligence, they are also influenced by relatively newer areas such as Machine Learning, Computational Statistics and Cognitive Science.

CAN HUMANS PARSE NATURAL LANGUAGE?

Usually not !!! We make mistakes on complex parsing structures
We can't parse without world knowledge and lexical knowledge

- Need to know what we're talking about
- Need to know the words used

Garden Path Sentences (sentences usually not correctly parsed by humans)

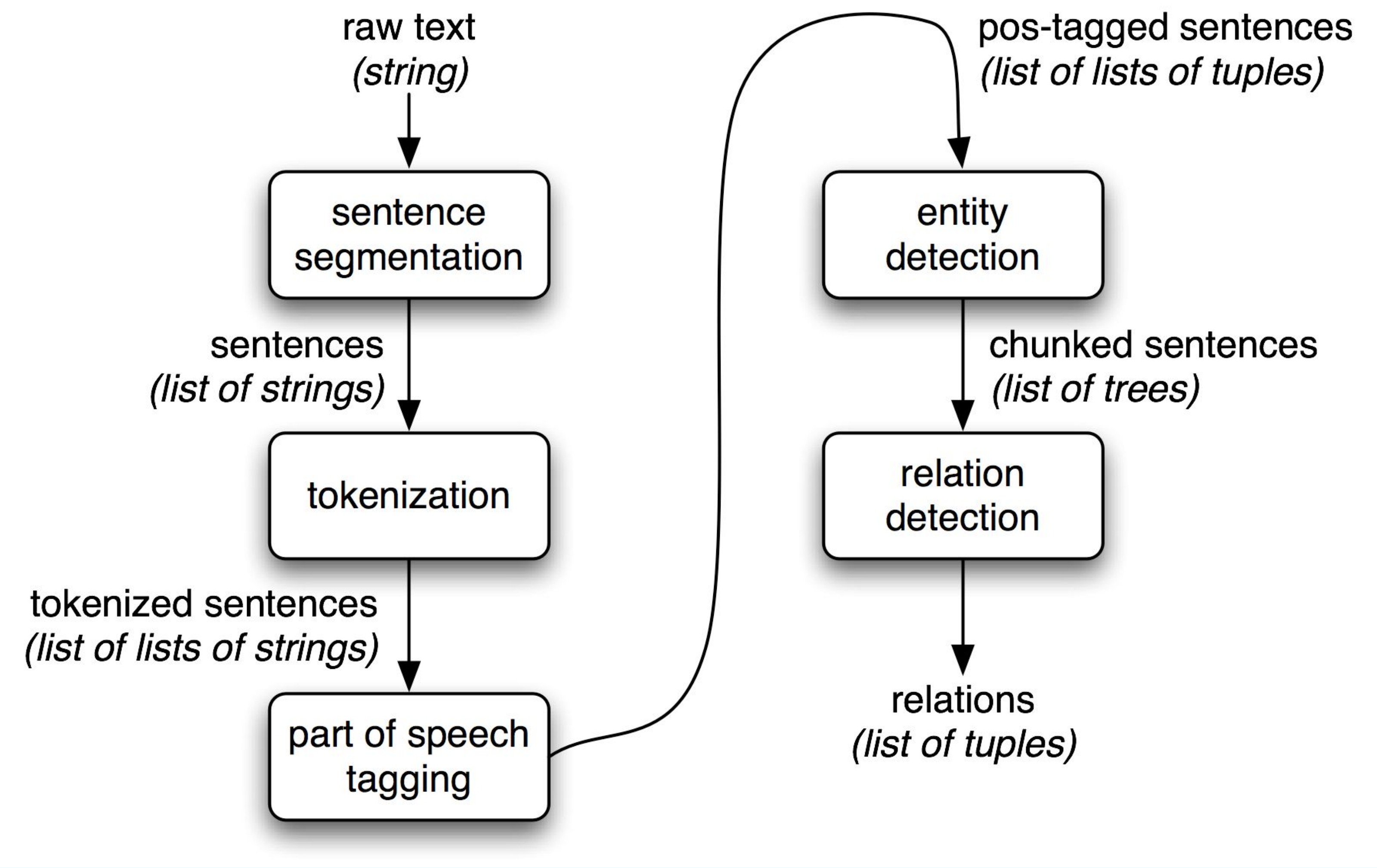
- While she hunted the deer ran into the woods.
- The woman who whistles tunes pianos. Confusing without context, sometimes even with

  Early semantic/pragmatic feedback in syntactic discrimination

Center Embedding
Leads to "stack overflow"

- The mouse ran.
- The mouse the cat chased ran.
- The mouse the cat the dog bit chased ran.
- The mouse the cat the dog the person petted bit chased ran

INFORMATION EXTRACTION ARCHITECTURE



SOME BASIC TERMINOLOGY

- **Token**: Before any real processing can be done on the input text, it needs to be segmented into linguistic units such as words, punctuation, numbers or alphanumerics. These units are known as tokens.
- **Sentence**: An ordered sequence of tokens.
- **Tokenization**: The process of splitting a sentence into its constituent tokens. For segmented languages such as English, the existence of whitespace makes tokenization relatively easier and uninteresting. However, for languages such as Chinese and Arabic, the task is more difficult since there are no explicit boundaries.
- **Corpus**: A body of text, usually containing a large number of sentences.
- **Part-of-speech (POS) Tag**: A word can be classified into one or more of a set of lexical or part-of-speech categories such as Nouns, Verbs, Adjectives and Articles, to name a few. A POS tag is a symbol representing such a lexical category - NN(Noun), VB(Verb), JJ(Adjective), AT(Article). One of the oldest and most commonly used tag sets is the Brown Corpus tag set.
- **Parse Tree**: A tree defined over a given sentence that represents the syntactic structure of the sentence as defined by a formal grammar.

TOKENIZATION

Tokenizers divide strings into lists of substrings.

For example, tokenizers can be used to find the list of sentences or words in a string.

STEMMING

Stemmers remove morphological affixes from words, leaving only the word stem online demo.

Simple stemmers:
Plural(meervoud)
Verbs(werkwoorden)

Different Stemming Algorithms:

- Paice/Husk Stemming Algorithm
- Porter Stemming Algorithm
- Lovins Stemming Algorithm
- Dawson Stemming Algorithm
- Krovetz Stemming Algorithm

PARTS OF SPEECH TAGGING (POS TAGGING)

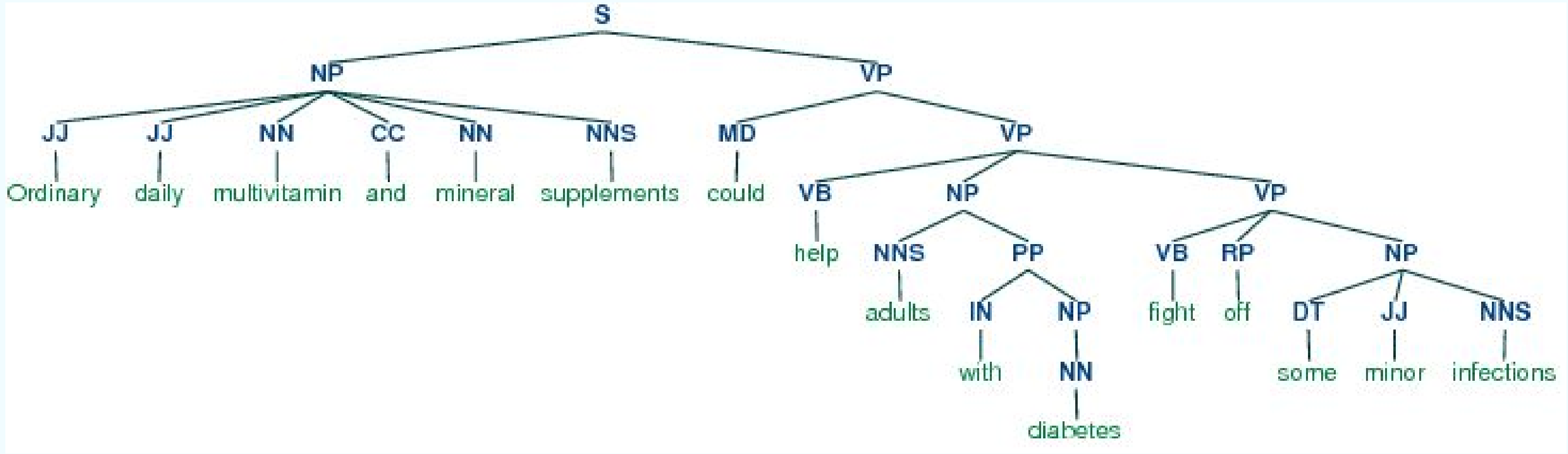
Parts of Speech Tagging (PoS tagging) is assigning Parts of Speech to the words in a text online demo.

```
Als vliegen vliegen vliegen vliegen vliegensvlug.
Als/CONJ vliegen/NN vliegen/VB vliegen/VB vliegen/NN vliegensvlug/ADV
```

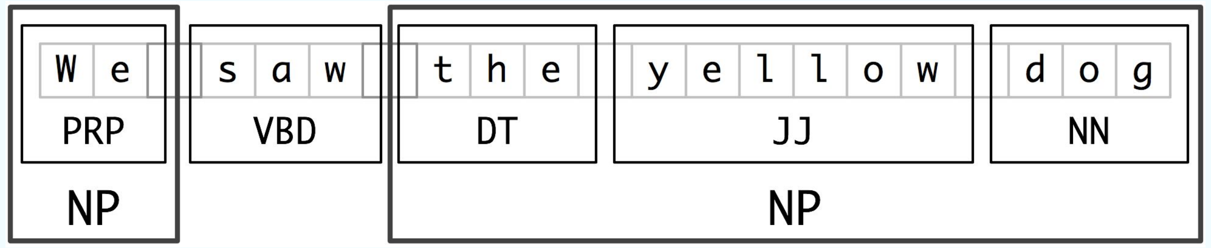
PoS tagging is a kind of word sense disambiguation: the PoS tag gives some information about the sense of the word in the context of use. It is a non-trivial task:

- Some words (at least in a sense of this word) that occur in the lexicon or dictionary have more than one possible Part of Speech. Like: "vliegen", it can be a noun as well as a verb.
  Note that even if we restrict to verbs the word “vliegen” has several senses: "Een vogel kan vliegen", "Als de bom valt vliegen de mensen uiteen."
- Some words are unknown.
- Tags are not well-defined. In "Wat fietsen" is “fietsen” a Noun or a Verb ?

PARSE TREE EXAMPLE



CHUNKING

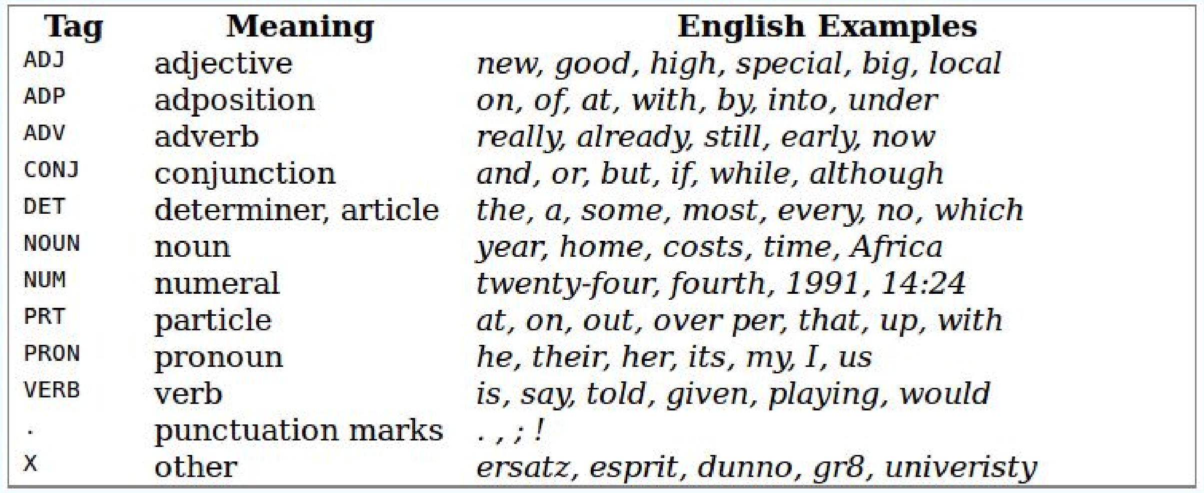


We	saw	the	yellow	dog
PRP	VBD	DT	JJ	NN
NP		NP		

The basic technique we will use for entity detection is chunking, which segments and labels multi-token sequences as illustrated above.

The smaller boxes show the word-level tokenization and part-of-speech tagging, while the large boxes show higher-level chunking. Each of these larger boxes is called a chunk.

UNIVERSAL PART-OF-SPEECH TAGSET



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*

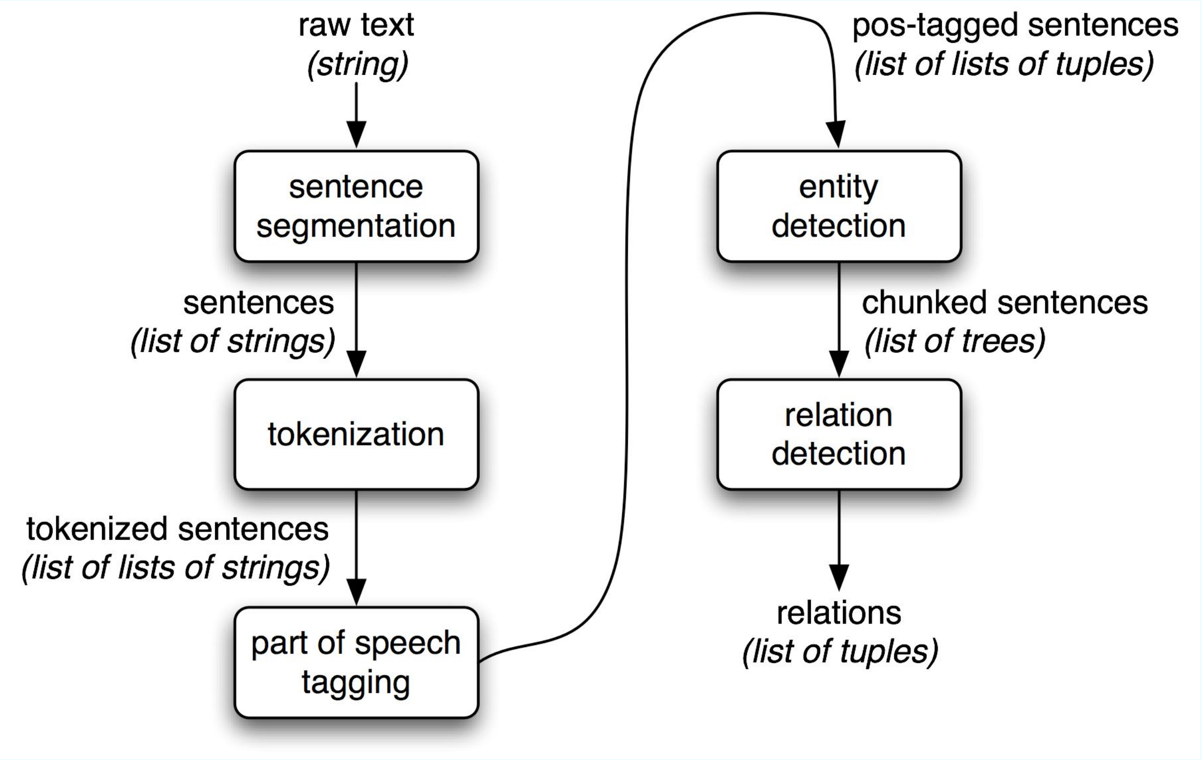
NAMED ENTITY RECOGNITION

Er zijn websites en API ‘s die dit process voor je kunnen doen.

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

RECAP: NATURAL LANGUAGE PROCESSING

- Tokenise
- Stemming
- Tagging
- Chunking
- Entity Recognition



FROG

For Dutch language : frog

TEXTMINING IN PYTHON

Textmining and natural language processing has become a huge field of research...

- NLTK
- spaCy
- Gensim

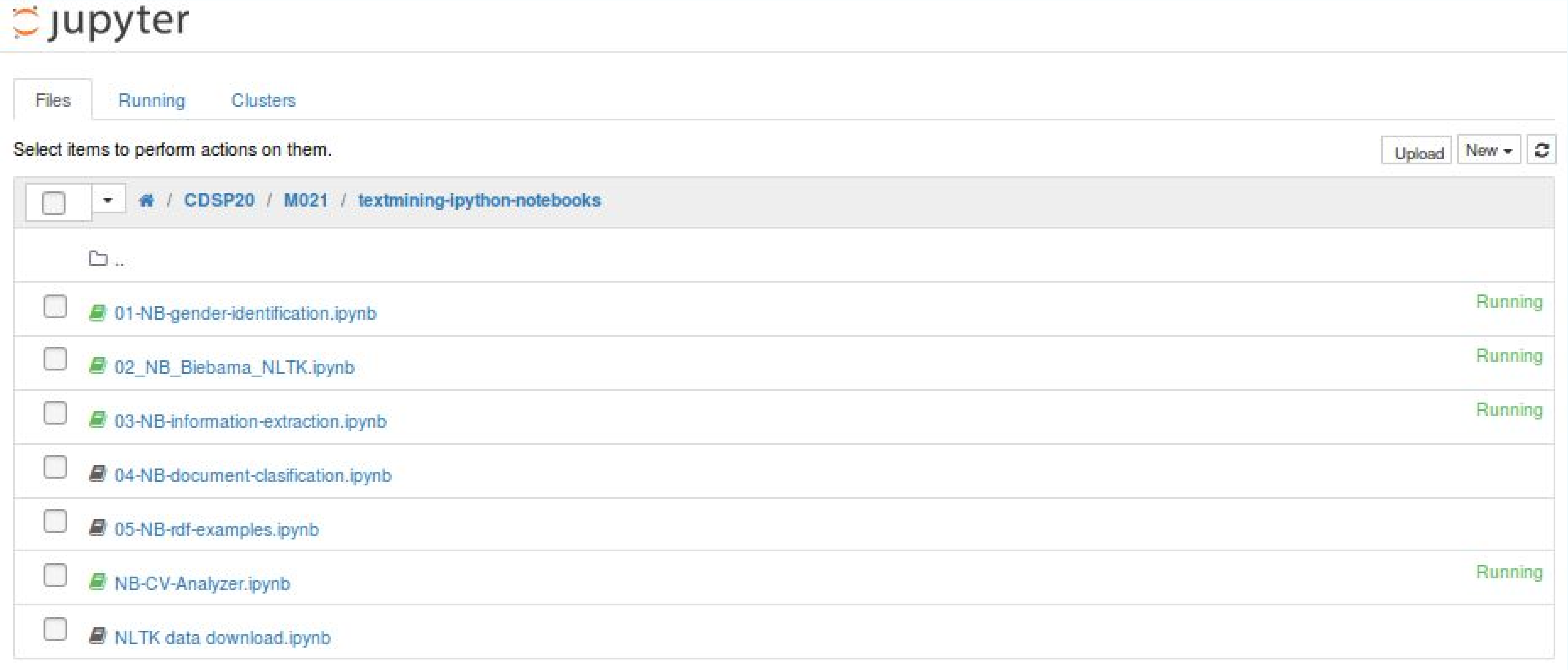
Deeplearning advancements in natural language processing are enormous, see huggingface

- word embedding (tensorflow embedding projector demo)
- transformers
- automatic translation
- automatic summary
- automatic Q&A chatbots

JUPYTER NOTEBOOK BASIC EXAMPLES

Start jupyter notebook server and play around with some examples:

```
jupyter notebook
```



Select items to perform actions on them.

/ CDSP20 / M021 / textmining-ipython-notebooks

..

□ 01-NB-gender-identification.ipynb	Running
□ 02_NB_Biebama_NLTK.ipynb	Running
□ 03-NB-information-extraction.ipynb	Running
□ 04-NB-document-clasification.ipynb	
□ 05-NB-rdf-examples.ipynb	
□ NB-CV-Analyzer.ipynb	Running
□ NLTK data download.ipynb	

DIKW
ACADEMY

QUESTIONS?