Lecture 1 – NLP Preprocessing and resources

SD-TSIA214

Chloé Clavel

Introduction

NLP tasks

2 kind of tasks:

- **Classify documents by themes, opinions etc...**
  - Supervised learning
    - Ex : SVM (support vector machines), Naive Bayes, … ?
  - Unsupervised learning
    - Ex: Clustering

2 kind of tasks:

- **Detect particular expressions**
  - Ex: Named Entities
    -

[Localité d'Ukraine] menace les livraisons de gaz à l' UE
. affaire Madoff contient encore de nombreuses zones d
le l' UE sous l'il de Paris [Communes de France] . La
tionnisme de Nicolas Sarkozy [Chef d'État] . Avec l'
ement culturel . La Russie [Pays] a cessé de fournir
ent] n' a pas à craindre pour ses approvisionnements .
le de l' occupation américaine en Irak [Pays] . Le
curées entre jeunes et policiers . Des engins incendiaires

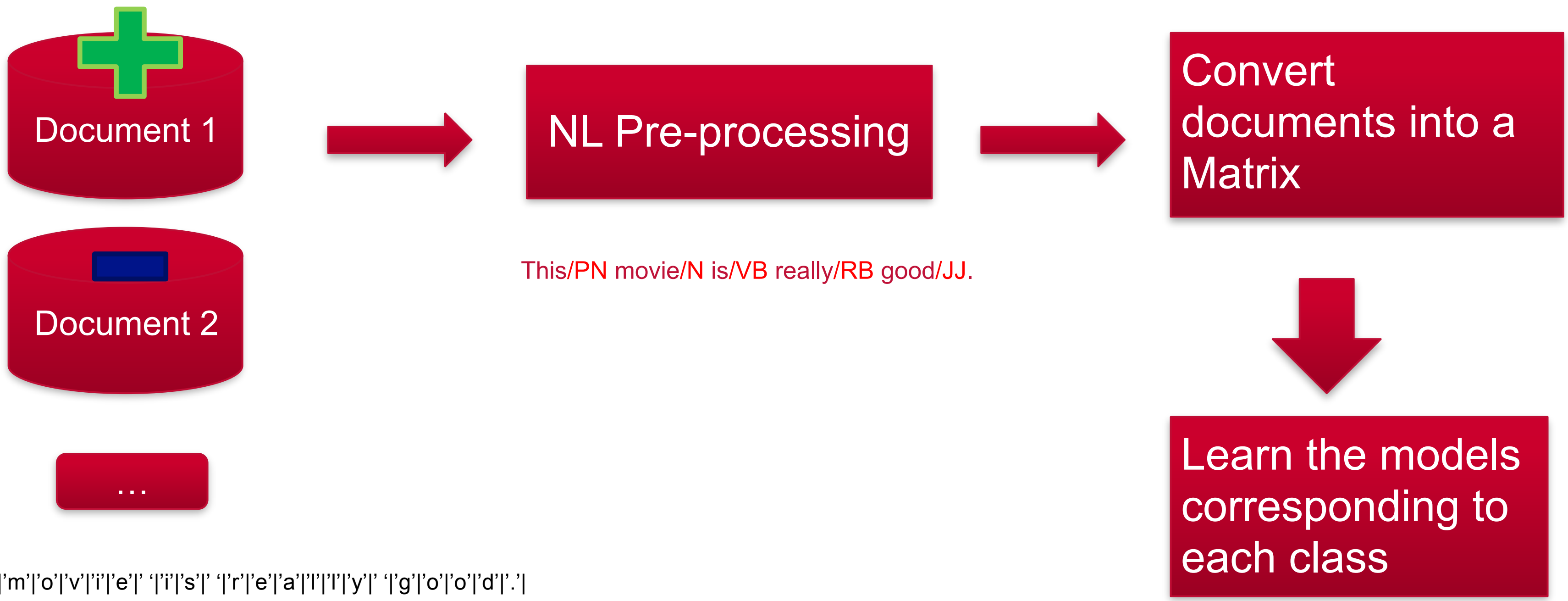
From http://www.tal.univ-paris3.fr/plurital/travaux-2009-2010/bao-2009-2010/MarjorieSeizou-AxelCourt/webservices.html

Classification

- **Phase 1 – learning**
  - Training corpus = set of labelled documents
    - Manual labeling : each document is assigned to a class :
      - Ex1. Movie reviews: the score attributed by a user (1 to 5)
      - Ex2. the topic of the document (sport, politics)
  - Goal : Learn from this corpus the specific features of each class

Phase 1 – learning

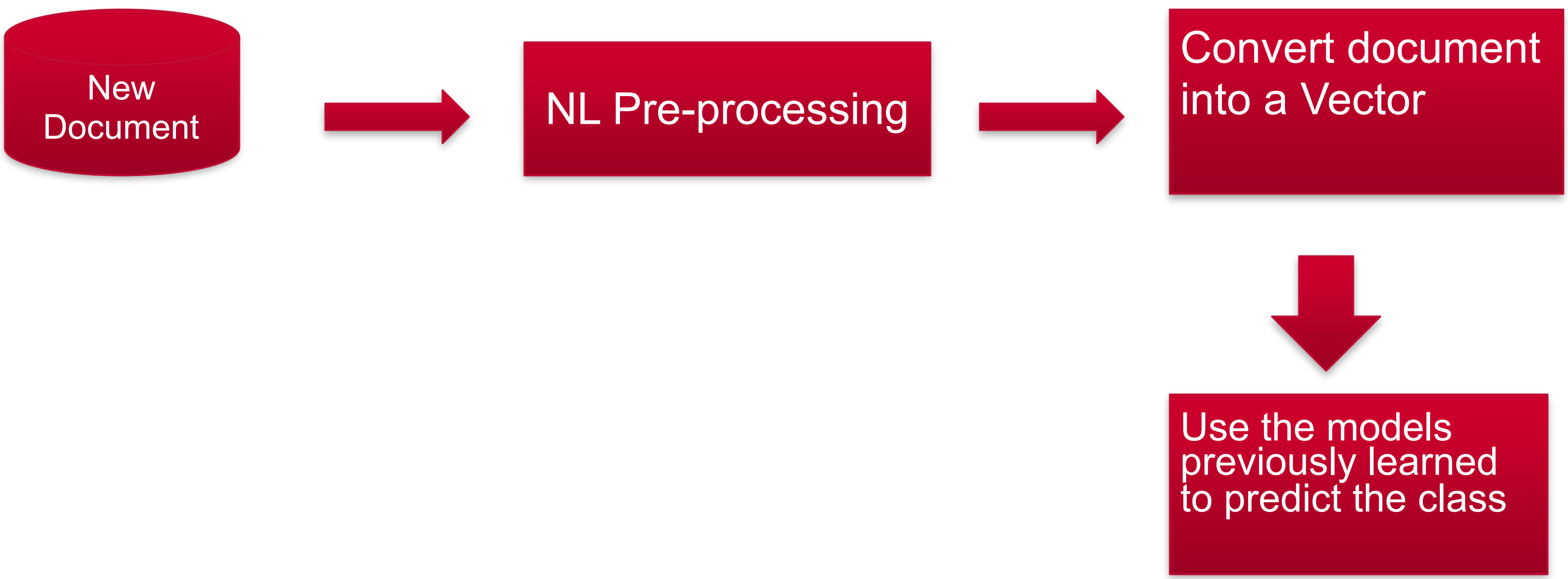
- **Learning the classes**



|’T’|’h’|’i’|’s’|’ ‘|’m’|’o’|’v’|’i’|’e’|’ ‘|’i’|’s’|’ ‘|’r’|’e’|’a’|’l’|’l’|’y’|’ ‘|’g’|’o’|’o’|’d’|’.’|

Phase 2 – classification

- **Phase 2 – classification**
  - Using the learned features, the system is able to assign a class to a new document



Objective of the lecture

- **Focus on Natural Language pre-processing**
  - for NLP tasks
- **Get familiar with:**
  - classical linguistic issues :
    - semantic disambiguisation, morphology, syntactic analysis, etc.
  - existing linguistic resources

Example of pre-processing steps for textual data analysis

1. **Segmentation / Tokenizing:**
   1. Words and sentences
2. **Lexical processing**
   1. determine the lexical information associated with each word in isolation (morphological rules and dictionary)
3. **Syntactic parsing:**
   1. Disambiguate according to the syntactic context, extract the grammatical relations that words and groups of words maintain between them
4. **Semantic parsing:**
   1. Word-sense disambiguation based on the context

Ex: "I saw a bat."

1. I/saw/a/bat/./
2. bat : - a noun referring to a flying mammal or a wooden club
3. bat : - object of the verb saw
4. bat : a flying mammal

Motivations for pre-processing

- **Speech synthesis**
  - Syntactic analysis
    - to define the pronunication
      - Couvent (sit on eggs) ou Couvent (convent)
    - to handle the prosody of the voice
      - Define where to put silent pauses

Tests under Acapella : les poules couvent au couvent.

Motivations for pre-processing

- **Pre-processing for the building of syntactic patterns for information extraction**
  - Ex : Patterns which call
    - syntactic categories (ex: *#PREP_DE, #NEG)*

(manque|~negation-patt|
*(il/#NEG/y/avoir/~negation-patt))/(#PREP_DE)?/ (conseil|contact|~services-lex)**

Examples of patterns will be given in Lecture 7 Opinion Analysis

Motivations for pre-processing

- **Pre-processing for text classification**
  - Reduce the representation space
    - Group inflected forms of words around lemmas
      - (ex: infinitive for a verb, masculine singular for a noun)

Details later in this lecture

Tokenization

Tokenization

I/saw/a/bat/./

- **"I saw a bat. "**
  - given a character sequence,
- tokenization is …
  - the task of chopping it up into pieces, called *tokens*

Tokenization

I/saw/a/bat/./

- Option1: consider all the tokens indifferently
  - Output : (I, saw, a, bat, .)
- Option 2: consider that token = word/term
  - throw away certain characters (such as punctuation)
    - Output : (I, saw, a, bat)
  - throw away words coming from a list of stop words (common words which would appear to be of little value for the NLP task)
    - Output : (saw, bat)

Tokenization

I/saw/a/bat/./

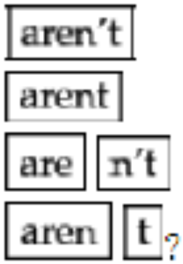
- **Simple Tokenization rule :**
  - chop on whitespace and throw away punctuation characters
- **Tricky cases**
  - Markers: dash (« - »), coma (« , »), tabulations (« »),
  - White space in « San Francisco »
  - End of sentence detection (find the dot (« . »)) : beware of acronyms E.N.S.T., numbers (3.14), and dates (02.05.2018)

Tests under Acapella
Nous sommes le 02.05.2018. Il y a quelques années le nom de l'école était l'ENST ou mieux l'E.N.S.T.

Tokenization

I/saw/a/bat/./

- **Tricky cases**
  - Markers: dash (« - »), coma (« , »), tabulations (« »),
  - End of sentence detection (find the dot (« . »))
  - uses of the apostrophe for possession and contractions



Tokenization

- **More involved methods for word segmentation :**
  - Heuristic-based :
    - Use of a large vocabulary
    - Take the longest vocabulary match
      - Ex : I went to San Francisco -> (I, went, San Francisco)
    - Use some heuristics for unknown words
  - Machine learning sequence models
    - trained over hand-segmented words
    - Ex: hidden Markov models
      - see Lecture 6 Hidden Markov Models - Laurence Likforman

such methods make mistakes sometimes, and so you are never guaranteed a consistent unique tokenization.

Syntactic analysis – Part of Speech tagging and chunking

Part-Of-Speech (POS) tagging

INPUT:
Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT:
Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ,/, as/P their/POSS CEO/N Alan/N Mulally/N announced/V first/ADJ quarter/N results/N ./.

N	=	Noun
V	=	Verb
P	=	Preposition
Adv	=	Adverb
Adj	=	Adjective

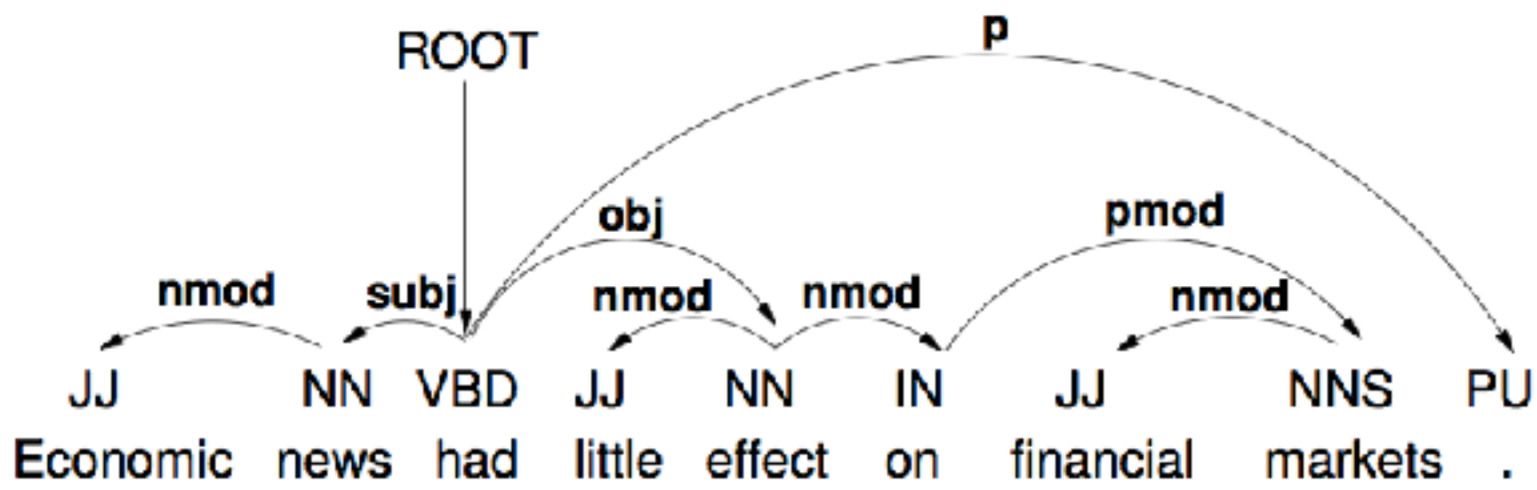
...

Chunking task

[John] [talked] [to the children][about drugs]

- **Detection of syntactic components :**
  - Noun phrase (groupe nominal), verbal phrase (noyau verbal),etc.
  - Borderline detection
  - Phrase labelling

Dependency parsing



Methods and challenges

- **2 types of methods**
  - Based on linguistic expertise

Lecture 3 Syntax and Parsing Jean-Louis Dessalles

  - Machine learning:
    - From a BIG labelled corpus/database
    - Learn the probabilities for the different transitions between syntactic categories

Modelling linguistic expertise for POS tagging/chunking/dependency parsing

- **More details in**
  - *Lecture 3 Syntax and Parsing Jean-Louis Dessalles*
- **Example :**
  - DET/PRO V -> PRO V
  - NP (Noun Phrase) : DET ADJ* NN ADJ*
- **Strengths :**
  - Readable rules,
  - Errors are easier to understand
- **Weaknesses**
  - Not robust to noisy inputs and out of vocabulary words
  - Time-consuming

Machine learning for POS tagging

- **Problem formulation for Hidden Markov Models**

$$
\begin{array}{lcccccccccc}
\text{M} = & \cdot & \cdot & \cdot & w_{i-2} & w_{i-1} & w_i & \cdot & \cdot & \text{mots} \\
\text{E} = & \cdot & \cdot & \cdot & e_{i-2} & e_{i-1} & e_i & \cdot & \cdot & \text{étiquettes}
\end{array}
$$

- Labelled corpus :
  - sequences of pairs (word, syntactic category)
- Training :
  - learn the probabilities for the different transitions between syntactic categories

See Lecture 6 Hidden Markov Models - Laurence Likforman

Machine learning for POS tagging

- **Problem formulation for Hidden Markov Models**

$$\begin{array}{lccccccccl} M = & \cdot & \cdot & \cdot & w_{i-2} & w_{i-1} & w_i & \cdot & \cdot & \text{mots} \\ E = & \cdot & \cdot & \cdot & e_{i-2} & e_{i-1} & e_i & \cdot & \cdot & \text{étiquettes} \end{array}$$

N: number of distinct observations (vocabulary size)
C: number of grammatical categories

- Training : Learning the model λ = (**A, B, Π**) from a labeled corpus
  - A : CxC transition state matrix
    - e.g. probability to have a VERB after a DET
  - B : NxC matrix of the probabilities of the observation i in state j
    - e.g. probability to generate « like » if the state is a verb
  - Distribution Π of the initial state : vector of length C
    - E.g. probability to begin with a DET

See Lecture 6 Hidden Markov Models - Laurence Likforman

Machine learning for POS tagging

- **Problem formulation for Hidden Markov Models**

$$\begin{array}{lcccccccccc} M = & \cdot & \cdot & \cdot & w_{i-2} & w_{i-1} & w_i & \cdot & \cdot & \text{mots} \\ E = & \cdot & \cdot & \cdot & e_{i-2} & e_{i-1} & e_i & \cdot & \cdot & \text{étiquettes} \end{array}$$

- Training :
  - Learning the model λ = (**A, B, Π**) from a labeled corpus
- Decision:
  - find the best sequence E that maximizes the model for the sequence of words M

See Lecture 6 Hidden Markov Models - Laurence Likforman

Machine learning for POS tagging

- Simplifying hypothesis:
  - Markov assumption : first-order markov chain (the probability of a particular state depends only on the previous state)

$$P(q_i|q_1 \ldots q_{i-1}) = P(q_i|q_{i-1})$$

  - Conditionnally to the labels, words are independant :

$$p(w_i|e_1 \ldots e_i, w_1 \ldots w_i) = p(w_i|e_i)$$

See Lecture 6 Hidden Markov Models - Laurence Likforman

Syntactic analysis - challenges

- **Challenges**
  - Disambiguation : « La petite brise la glace »
  - Capable of handling mistakes and typos

Filtering words

To reduce vocabulary size for NLP tasks

Reduce vocabulary size

- **According to the NLP task, filtering …**
  - Punctuation (??,!!, .)
    - NB: useful for opinion mining
  - Dates
    - NB: useful for Named Entity Recognition
  - stop words using a predefined list (e.g. linking words)
    - NB: linking words are useful for argument mining
  - Hapax
    - Marginal terms (occuring once or twice) in the corpus
      - often corresponds to mispelling words

Reduce vocabulary size

- **Gather inflectional forms and derivationally related forms**
  - Inflectional forms : a change in or addition to the form of a word that shows a change in the way it is used in sentences
  - Morphological derivation, : the process of forming a new word from an existing word, often by adding a prefix or suffix, such as -ness or un-

PRACTICE :

ENGLISH :

propose inflectional forms of « dog », « sit »

propose derivational form of happy

FRENCH : donner les flexions du verbe « jouer »

Reduce vocabulary size

- **Gather inflectional forms and derivationally related forms of a word around …**
  - Their stems -> stemming
    - « cherchons » -> « cherch »
  - Their lemmas -> lemmatization
    - am, are, is => be
    - car, cars, car's, cars' => car

- **Stemming and lemmatization are based on**
  - a morphological analysis of the words

Morphological analysis : an analysis of word internal structure
Morpheme : minimal linguistic unit carrying a sense (abstract unit)
Morphologic processes : flection, declension, conjugation, derivation (anti-constitu-tionn-elle-ment)

Reduce vocabulary size

- **Stemming and lemmatization are based on**
  - a morphological analysis of the words
- **What is morphological analysis?**
  - An analysis of word internal structure
  - Morpheme : minimal linguistic unit carrying a sense (abstract unit)
  - Morphologic processes : flection, declension, conjugation, derivation (anti-constitu-tionn-elle-ment)

Reduce vocabulary size

- Stemming :
  - a crude heuristic process that chops off the ends of words (removal of derivational affixes)
  - How? Ex: Porter’s algorithm based on morphological rules [Porter, 1980]

(1) | Rule | | | Example | |
---|---|---|---|---|---|---
 | SSES | → | SS | caresses | → | caress
 | IES | → | I | ponies | → | poni
 | SS | → | SS | caress | → | caress
 | S | → | | cats | → | cat

PRACTICE:

What is the stem of the word « frontal » in French?

Reduce vocabulary size

– Example of stemmer outputs

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Reduce vocabulary size

- **Lemmatization :**

“the boy's cars are different colors” =>
“the boy car be differ color”

- use of a vocabulary and morphological analysis of words,
- to return the *lemma:*
  - the base or dictionary form of a word: .
    - am, are, is => be
    - car, cars, car's, cars' => car

- NB: syntactic analysis can help to disambiguate some cases
    - Ex: « les poules du couvent couvent »
      - Couvent -> couvent (noun) ou couver (verb)

PRACTICE:
What are the stem and the lemma of the word « saw » in English?

Reduce vocabulary size

- **Lemmatization – existing tools**
  - For French
    - Treetagger
    - Xerox, Brill [Brill, 1995]
    - LIA_Tag, macaon http://macaon.lif.univ-mrs.fr/index.php?page=home-en
  - For English:
    - NLTK : http://www.nltk.org/
    - Treetagger

Xerox

[10/1996, 10/1997]

La	le	+DET_SG
petite	petit	+ADJ2_SG
ferme	ferme	+NOUN_SG
du	de=le	+PREP_DE
père	père	+NOUN_SG
Fouchard	Fouchard	+NOUN_INV
se	se	+PC
trouvait	trouver	+VERB_P3SG
au sortir du	au sortir de=le	+PREP
défilé	défilé	+NOUN_SG
.	.	+SENT

Lemma

Part of Speech

Reduce vocabulary size

- **Stemming vs. Lemmatization**
  - What is the best choice?
    - It depends on the language
      - Ex: stemming works well in German

Resources

Resources

- **Wordnet : lexical database**
  - Retrieve information on word meaning/sense
  - Core idea :
    - A word can have several meanings (ex: « bat »)
    - groups English words into *synsets*
    - Synsets : set of synonyms

PRACTICE :
Let's search the word « estimable » on Wordnet website for English
http://wordnetweb.princeton.edu/perl/webwn

Q1 : how many senses are existing for this word?
Q2 : what is the size of the synset of **estimable#2?**

Resources

- **Wordnet : lexical database**
  - Synsets : set of synonyms
    - Synonyms : words that are interchangeable in some context without changing the truth value of the proposition
    - Synsets include simplex words as well as collocations like "eat out" and "car pool.“
    - The meaning of a synset is further clarified with a short definition and one or more usage examples

Example :

good, right, ripe – (most suitable or right for a particular purpose; "a good time to plant tomatoes"; "the right time to act"; "the time is ripe for great sociological changes")

Resources

- **Wordnet : lexical database**
  - All synsets are connected to other synsets by means of semantic relations:
    - Ex: canine is a hypernym of dog
    - *window* is a meronym of *building*

PRACTICE
On wordnet
To see the semantic relation click on the S

Version française : Wordnet Libre du Français (WOLF) : http://alpage.inria.fr/~sagot/wolf.html

References

- https://nlp.stanford.edu/IR-book
- [PORTER, 1980]
  - M.F. Porter, 1980, An algorithm for suffix stripping, *Program*, 14(3) pp 130−137.