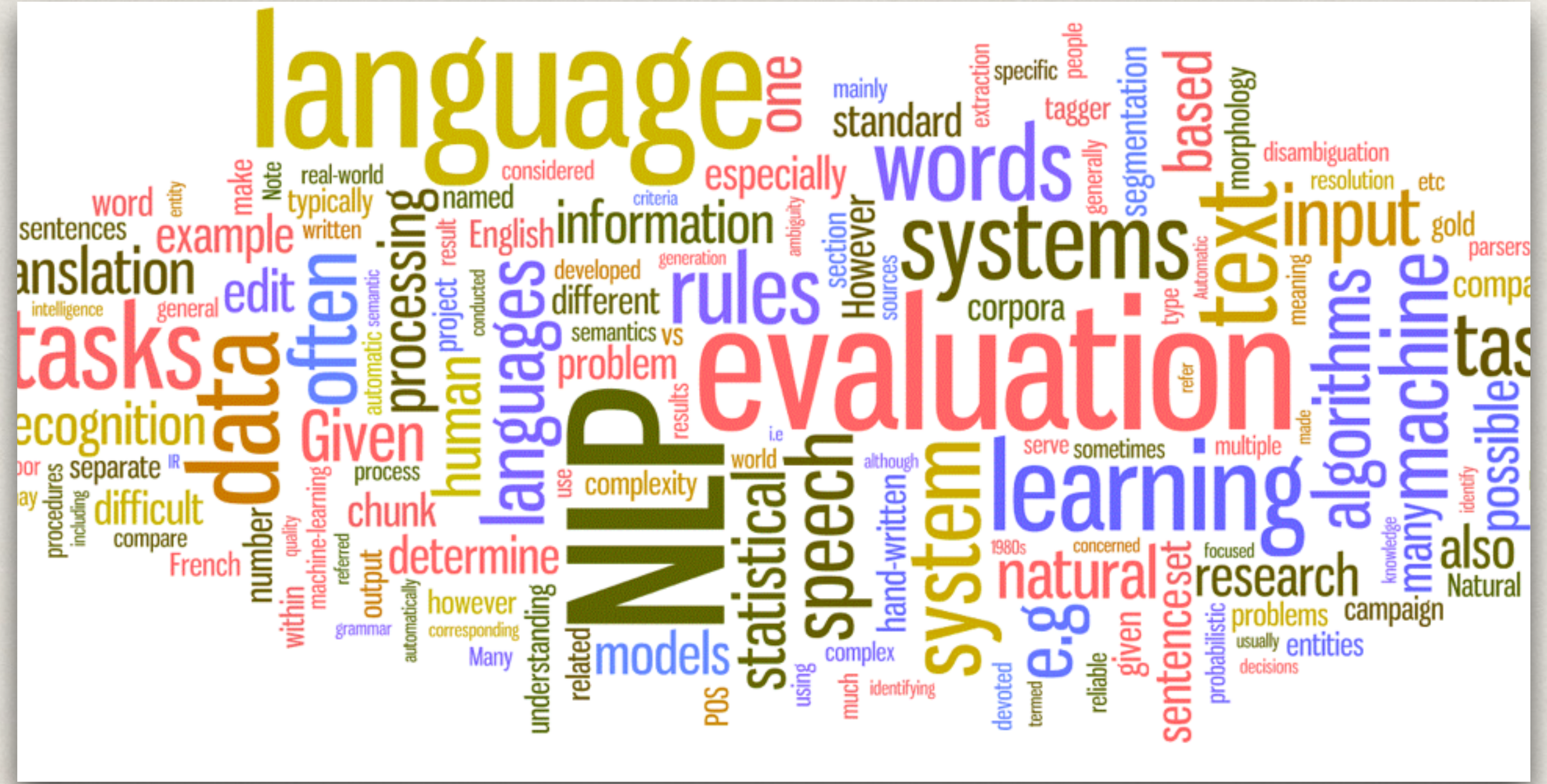


Natural Language Processing

Podstawy zagadnienia

Meetup, 28.03.2017

Agenda

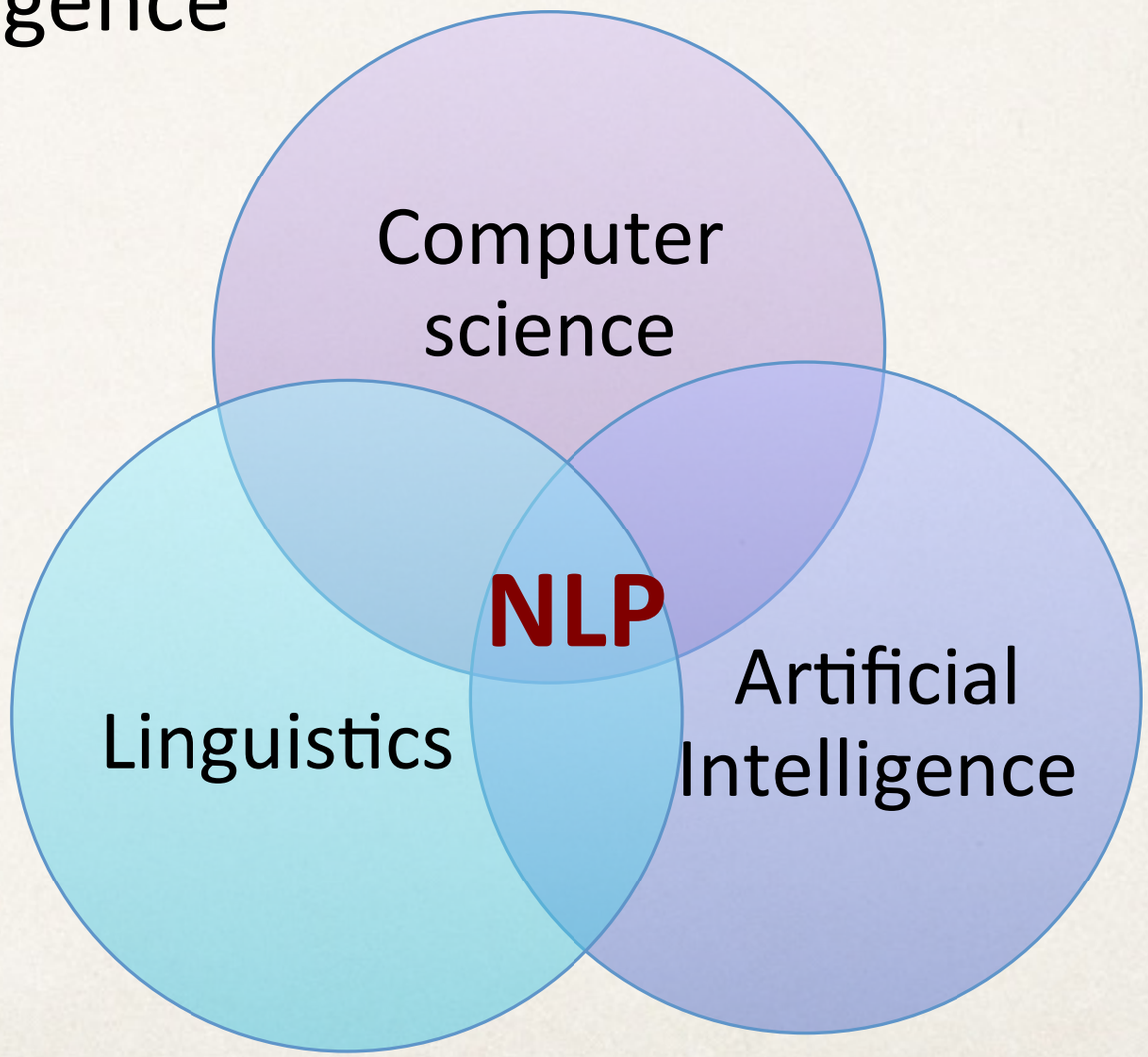
- Podstawy przetwarzania języka
- Dostępne narzędzia
- Trochę teorii
- Zastosowania
- InsightE

Natural Language Processing

- **Wiele nazw:**
  - **Natural Language Processing**
  - **Computational Linguistics**
  - **Natural Language Understanding**
  - **Language Engineering**
  - **Language Technology**
  - **Human Language Technology**

Czym jest NLP ?

- Sub-field of Artificial Intelligence
- Inter disciplinary subject
- Based on linguistics



Zasoby i aplikacje

- **Dane:**
  - Korpusy językowe (np. Brown czy British National Corpus)
  - Internet (Wikipedia, blogi, social media, strony internetowe, dokumenty itd.)
- **Zadania:**
  - Machine translation
  - Text summarization / categorization
  - Question answering
  - Sentiment analysis
  - Word-sense disambiguation
  - Information extraction
- **Narzędzia:** taggery, stemmery, parsery, itd

Dwa podstawowe podejścia

- **Przetwarzanie języka naturalnego i badanie opinii**

- **Analiza statystyczna**
  - Etykietowane dane
  - Zliczanie
  - Uczenie maszynowe
- **Analiza formalna**
  - Bazy wiedzy
  - Drzewa leksykalne
  - Rozumienie tekstu

Projekty w Polsce

- **ZIL – Zespół Inżynierii Lingwistycznej IPI PAN**
  - Spory zbiór darmowych i otwartych narzędzi
  - Warto odwiedzić też strony CLIP i CLARIN-PL
  - Ten sam zespół rozwija wyszukiwarkę NEKST
- **G4.19** – grupa badawcza z Politechniki Wrocławskiej
- **PSI –** Pracownia Systemów Informacyjnych UAM

Trochę teorii

- **Fonetyka:** nauka zajmująca się systematyką dźwięków w językach mówionych
- **Morfologia:** dlaczego “impossible” jest poprawne a ”imred” nie?
- **Składnia:** co sprawia, że “I lecture on computational linguistics” jest poprawne a “I lecturing computational linguistics” już nie?
- **Semantyka:** dlaczego
  - “I borrowed it from Jim” jest poprawne a “I borrowed it to Jim” już nie?
  - czym się różnią “robbery by the lake” i “robbery by the fugitive”?
- **Pragmatyka:** dlaczego zdanie “taxes always go down” jest niepoprawne?

Podstawowe zagadnienia

- **Tokenizer** – dzieli tekst na słowa i znaki interpunkcyjne
  - Sporo pułapek, np. At 8 o’clock I didn’t feel good. =>
  - |At|8|o’clock|I|did|n’t|feel|good|.|
  - Wyrażenia regularne lub automaty
- **POS** – part-of-speech tagger, oznacza części mowy
  - Potrzebne do analizy składniowej („parsowania”) zdań
  - Przydatne do aplikacji Text-to-speech (OBject vs obJECT)
  - Hidden Markov Models albo rule-based
  - Czasami zamiast POS starczy stemming
- **N-grams** – n-tki słów opisane prawdopodobieństwem ich wystąpienia
  - Mogą służyć do podpowiadania słów lub generowania tekstów

Podstawowe zagadnienia

- **Parser** – sprawdza składnię, określa części zdania
  - W praktyce zależy nam stworzeniu drzewa (drzew?) składniowych
  - Dla języków naturalnych trudne zadanie
    - I saw a man on a hill with a telescope
  - Oparte na regułach lub statystyce (zliczaniu)

- **Semantyka** – najciekawszy i najtrudniejszy etap
  - Brak gotowych rozwiązań
  - Predykaty, sieci semantyczne, ontologie, taksonomie, statystyka
  - Techniki mocno zależne od zastosowania

Word Sense Disambiguation

- Co wybrane słowo oznacza w danym kontekście?
- Słowa wieloznaczne, homonimy,
- Potrzebne zasoby:
  - Statystyki występowania słów
  - Słowniki
- Maksymalna trafność: zdolności ludzkie
- Minimalna: częstość najpopularniejszego znaczenia

Zastosowania WSD

1. **Podejście statystyczne:**
   - Bierzemy korpus etykietowanych tekstów
   - Dla wskazanego słowa określamy wszystkie jego konteksty w korpusie
   - Korzystamy z twierdzenia Bayesa i maksymalizujemy P(w|c)
2. **Podejście słownikowe:**
   - Porównujemy kontekst słowa z definicją słownikową
3. **Tezaurus**
   - Sprawdzamy bliskość każdego słowa w kontekście do znaczenia w słowosieci

Machine Translation

- **Chyba najtrudniejsze zastosowanie**
  - Celem jest automatyczne tłumaczenie tekstów
- **Przykłady działających systemów:**
  - BabelFish
  - Google Translate
  - Bing Translator
- **Podobne(?) zadanie: Natural Language Generation**

Text Summarization / Categorization

- Jak podsumować dłuższy tekst w kilku zdaniach?
- Trudne zadanie !
- Wyciąganie części zdań z tekstu (brak ciągłości)
- Podsumowywanie

Named Entity Recognition

- **Wykrywanie encji w tekście:**
  - Osoby
  - Miejsca
  - Organizacje
  - Określenia czasu
  - Wartości
- Dobre wyniki dla j. angielskiego (Stanford)
- Dla polskiego trochę gorzej…

Question Answering

- Automatyczne odpowiadanie na pytania
- Głównie związane z poszukiwaniem wiedzy i porad w Internecie
- Wyszukiwarki internetowe zaczynają to robić
- Bardzo dobre wyniki
- Wykorzystanie Wikipedii, słowosieci i baz wiedzy (ang. Knowledge bases):
  - Nell
  - Cyc

Sentiment Analysis

- Sentiment analysis (or opinion mining) - computational study of opinion, sentiment, appraisal, evaluation, and emotion:
  - Badanie opinii o produktach
  - Poszukiwanie opinii przed zakupem
  - W ostatnich latach bardzo popularne zadanie NLP
  - Ściśle powiązane z analizą mediów społecznościowych

„Sentyment” vs Opinia

- Bardziej związany z uczuciami

- Bardziej związana z przemyśleniami i spostrzeżeniami

Zastosowanie Sentiment Analysis

- Potrzebujemy dwóch definicji:
  - Czym jest pojedyncza opinia?
  - Jaka jest ogólna opinia (w populacji)?

- Opinia składa się z nastepujących 5 elementów:
  - **encja:** opisywany obiekt
  - **aspekt:** cecha opisywanego obiektu
  - **sentyment:** +, -, neu, ocena, liczba gwiazdek, emocja
  - **właściciel:** osoba wypowiadająca opinię
  - **czas:** moment kiedy opinia została wyrażona

Przykłady

Workflow


Data retrival
Data cleaning
Data processing
Data analysis

Typowe Elementy Workflow

Data cleaning

- tokenization
- standarization / unification

Data processing

- stemming / lemmatization
- POS (Parts of speech) tagging

Data processing

- Contextual analysis
- Sentiment analysis
- NER

Tokenization and Standarization (Text Pre-processing / Cleaning)

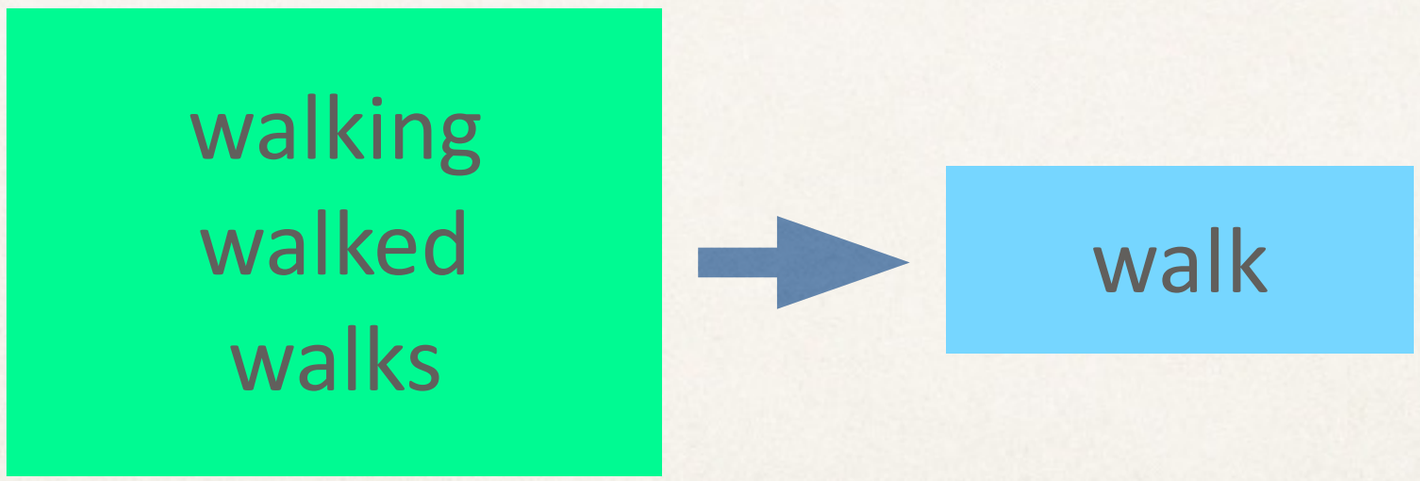
- text format preparation -> handle html, pdf, etc.
- tokenization
- standarization/unification
  - stopwords [ENG: the, hi, there, itd. PL: mnie, cię, już, ku itd.]
  - punctuation [.,;'!@#%^&*-_]
  - emoji
  - numbers
  - one-letter words
  - special characters
  - starting sentence seq
  - ending sentence seq
  - other methods...

Tokenization

- Standard Tokenizer - whitespace and punctuation as delimiters
- Keyword Tokenizer
- Letter Tokenizer
- Lower Case Tokenizer
- N-Gram Tokenizer
- Edge N-Gram Tokenizer
- ICU Tokenizer
- Path Hierarchy Tokenizer
- Regular Expression Pattern Tokenizer
- UAX29 URL Email Tokenizer
- White Space Tokenizer

Stemming

- temat wyrazu (część wyrazu, która nie podlega odmianie)
- Porter, Lancaster, Lovins, Snowball, Paice



Stemmers - porównanie

- **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.

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

- **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

- **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

Stemming vs Lemmatization

- **Stemming** - proces polegający na wydobyciu z wybranego wyrazu tzw. rdzenia, a więc tej jego części, która nie podlega odmianie

- **Lematyzacja** - proces podobny do powyższego, a oznacza sprowadzenie grupy wyrazów stanowiących odmianę danego zwrotu do wspólnej postaci, umożliwiającej traktowanie ich wszystkich jako te samo słowo.

Stemmers and Lemmatizers in NLTK

- from nltk.stem.api import StemmerI
- from nltk.stem.regexp import RegexpStemmer
- from nltk.stem.lancaster import LancasterStemmer
- from nltk.stem.isri import ISRIStemmer
- from nltk.stem.porter import PorterStemmer
- from nltk.stem.snowball import SnowballStemmer
- from nltk.stem.wordnet import WordNetLemmatizer
- from nltk.stem.rslp import RSLPStemmer

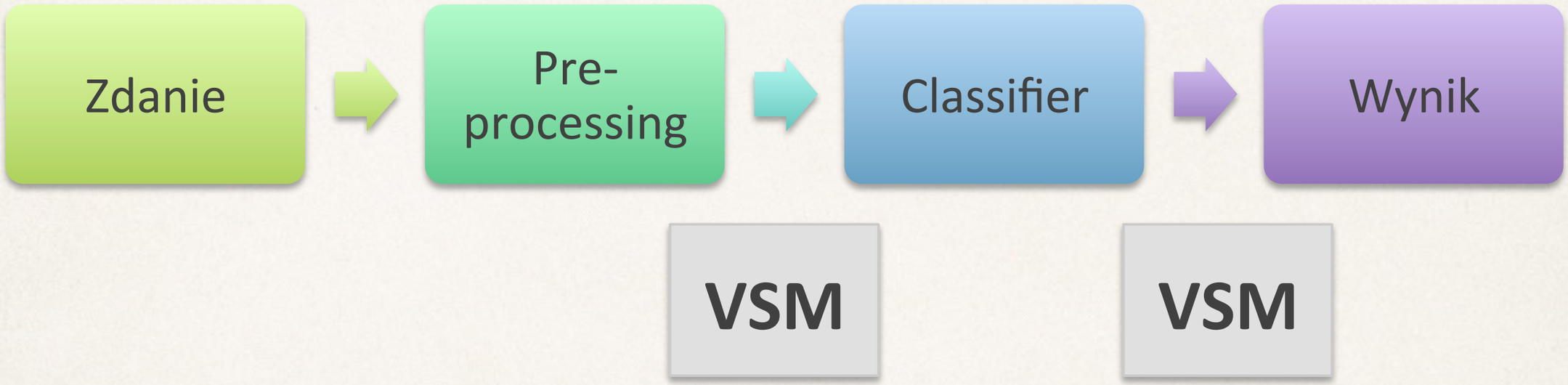
POS tagging

1. CC Coordinating conjunction
2. CD Cardinal number
3. DT Determiner
4. EX Existential there
5. FW Foreign word
6. IN Preposition or subordinating conjunction
7. JJ Adjective
8. JJR Adjective, comparative
9. JJS Adjective, superlativ
10. LS List item marker
11. MD Modal
12. NN Noun, singular or mass
13. NNS Noun, plural
14. NNP Proper noun, singular
15. NNPS Proper noun, plural
16. PDT Predeterminer
17. POS Possessive ending
18. PRP Personal pronoun
19. PRP$ Possessive pronoun
20. RB Adverb
21. RBR Adverb, comparative
22. RBS Adverb, superlative
23. RP Particle
24. SYM Symbol
25. TO to
26. UH Interjection
27. VB Verb, base form
28. VBD Verb, past tense
29. VBG Verb, gerund or present participle
30. VBN Verb, past participle
31. VBP Verb, non-3rd person singular present
32. VBZ Verb, 3rd person singular present
33. WDT Wh-determiner
34. WP Wh-pronoun
35. WP$ Possessive wh-pronoun
36. WRB Wh-adverb

Not covered here

- N-grams
- Vector Space Models

Sentiment Analysis



- Supervised Learning Algos
  - Naive Bayes
  - MaxEnt
  - SVM
  - NN

Confusion matrix

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Typy Błędów

		Predicted	
Actual	Positive	Positive	Negative
	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

- The recall is the ratio **tp / (tp + fn)**

Type II error

- The precision is the ratio **tp / (tp + fp)**

Type I error

Advanced NLP

- Concept tagging
- Insights
- Emotions retrieval
- Summarization
- Classification
- Ontologies
- Language dependence

Dziękuję

Najpopularniejsze biblioteki

TextBlob: https://textblob.readthedocs.io/en/dev/

NLTK: http://www.nltk.org

Gensim https://radimrehurek.com/gensim/

Pattern: http://www.clips.ua.ac.be/pattern

Polyglot: https://pypi.python.org/pypi/polyglot

Scikit-learn: http://scikit-learn.org