

Text mining - odkrywanie wiedzy z tekstowych zbiorów danych

Wykład 1

Wstęp

Lingwistyka

Przetwarzanie tekstu

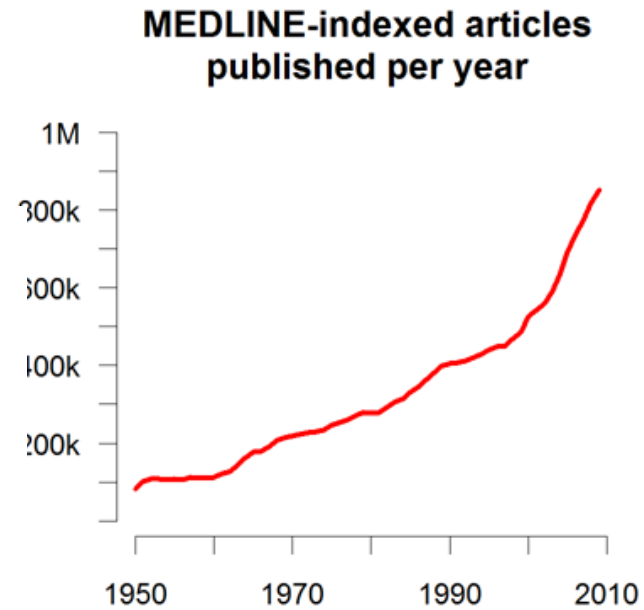
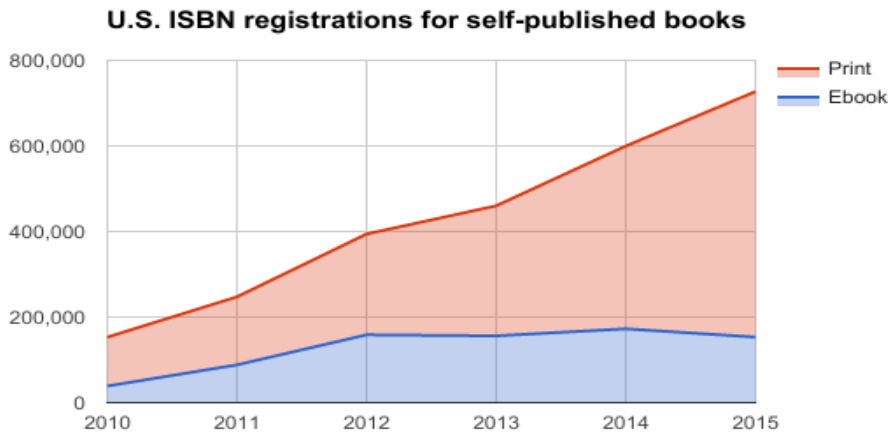
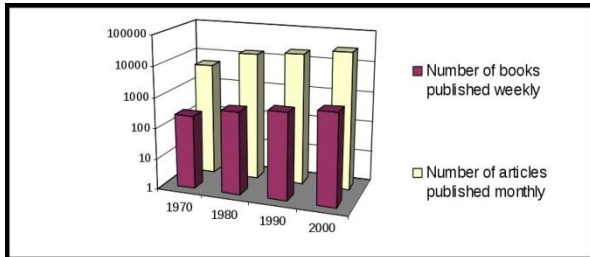
Regulamin

- Zaliczenie przedmiotu jest na podstawie ocen z zajęć laboratoryjnych – szczegóły na zajęciach laboratoryjnych.

Literatura i oprogramowanie

- C. Manning, H. Schütze, „**Foundations of Statistical Natural Language Processing**”, MIT Press, 1999, <http://nlp.stanford.edu/fsnlp/errata.html>
- C. Manning et al., “**Introduction to Information Retrieval**”, Cambridge Univ. Press, 2008, <http://www-nlp.stanford.edu/IR-book/>
- Dan Jurafsky et al. „**Speech and Language Processing**”, Prentice-Hall, 2008 <http://www.cs.colorado.edu/~martin/SLP/slp-errata.html>
- N. Indurkha, F. Damerau, „**Handbook of Natural Language Processing**”, Chapman and Hall, 2010
- M. Russel, "**Mining the Social Web**", OReilly, 2013
- **Python NLTK** (Natural Language Toolkit) <http://nltk.sourceforge.net>
- **Open NLP** <http://opennlp.sourceforge.net/>
- **GATE** <http://gate.ac.uk/>

Eksploracja informacyjna



- Zwiększające się znaczenie Internetu jako kanału dystrybucji informacji
- Minimalne koszty powielania informacji w formie elektronicznej
- Większość ludzkiej wiedzy zapisana jest w postaci dokumentów w języku naturalnym

Zadanie dziedziny

Komputery (systemy komputerowe) byłby o wiele bardziej użyteczne gdyby umiały np.

- gromadzić odpowiednie materiały na dany temat,
- generować streszczenia tekstów,
- procesować wiadomości mejlowe,
- rozmawiać z nami.

Jednak do tego potrzeba, aby komputery mogły posługiwać się językiem naturalnym.

Problem: jak nauczyć komputery posługiwania się językiem naturalnego.

Sztuczna inteligencja - argument chińskiego pokoju - 1984 J.R. Searle

Aksjomat 1: Mózgi są przyczynami umysłów.

Aksjomat 2: Syntaktyka nie wystarcza dla semantyki.

Aksjomat 3: Program komputerowy całkowicie określa jego formalna lub syntaktyczna struktura.

Aksjomat 4: Umysły zawierają treści psychiczne, mówiąc dokładniej, treści semantyczne.

Wniosek 1: Nie ma takiego programu komputerowego, który sam w sobie wyposażyłby system w umysł. Mówiąc krótko, programy nie są umysłami ani same w sobie nie wystarczą dla powstania umysłu.

Wniosek 2: Czynności mózgu ograniczone tylko do realizowania programu komputerowego nie wystarczą, by funkcjonowanie mózgu doprowadziło do powstania umysłu.

Wniosek 3: Cokolwiek, co mogłoby być przyczyną umysłu, musiałoby mieć moc oddziaływania przyczynowego porównywalną z możliwościami mózgu.

Wniosek 4: Wyposażenie jakiegoś zbudowanego przez nas artefaktu w program komputerowy nie wystarcza, by miał on stany umysłowe porównywalne z ludzkimi. Artefakt taki powinien oczywiście mieć zdolność przyczynowego oddziaływania porównywalną z możliwością ludzkiego mózgu.

Zastosowanie NLP

- Korekta pisowni, sprawdzanie gramatyki
- Lepsze wyszukiwarki
- Automatyczne tłumaczenia
- Wydobywanie informacji
- Tworzenie dokumentów
- ...

Nowe interfejsy

- Rozpoznawanie mowy (czytanie tekstu)
- Systemy dialogowe (komputer pokładowy USS Enterprise)

NLP, NLU, NLG

NLP – Natural Language Processing

- Właściwie wszystko, co jest związane z przetwarzaniem informacji zapisanej w języku naturalnym
- Inne nazwy: Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)

NLU – Natural Language Understanding

- Dosłownie „rozumienie języka naturalnego”
- Co to jednak znaczy „rozumienie”?
- Semantyka i logika
- Rozumienie nie zawsze okazuje się niezbędne (*Chiński Pokój*)

NLG – Natural Language Generation

- To akurat jest proste (o ile nie mamy wygórowanych wymagań)

Proces komunikacji w językach naturalnych

- Tekst
- Budowa reprezentacji znaczenia (np. logika)
- Interpretacja w danym kontekście (sytuacja, wiedza, cel)
- Wybór działania - decyzja o tym jaka odpowiedź ma być sformułowana.
- Przygotowanie tekstu odpowiedzi w języku naturalnym.

Niestety NLU jest trudne (1)

- SMS/Email

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

- Segmentacja

the New York-New Haven
Railroad

- Idioms

dark horse

get cold feet

lose face

throw in the towel

- neologizmy

unfriend

Retweet

bromance

- semantyka

Mary and Sue are sisters.

Mary and Sue are mothers.

- nazwy własne

Where is *A Bug's Life*
playing ...

Let It Be was recorded ...

... a mutation on the *for*
gene ...

Niestety NLU jest trudne (2)

John **stopped** at the **donut store** on **his way home from work**.

He **thought** a **coffee** was good every few hours.

But **it** turned out **to be too expensive** there.

Przykład – J. Eisner

store where donuts shop? or is run by donuts?
or looks like a big donut? or made of donut?
or has an emptiness at its core?

I stopped smoking freshman year, but
John stopped at the donut store

Describes where the store is? Or when he stopped?

he stopped there from hunger and exhaustion, not just from work.

At that moment, or habitually? /Similarly: Mozart composed music./

That's how often he thought it?

But actually, a coffee only stays good for about 10 minutes before it gets cold.

Similarly: In America a woman has a baby every 15 minutes.
Our job is to find that woman and stop her.

the particular coffee that was good every few hours? the donut store? the situation?

...a to zaledwie trzy zdania.

Niestety NLU jest trudne (3)

Przykłady zapytań niemożliwych do wykonania bez wiedzy semantycznej

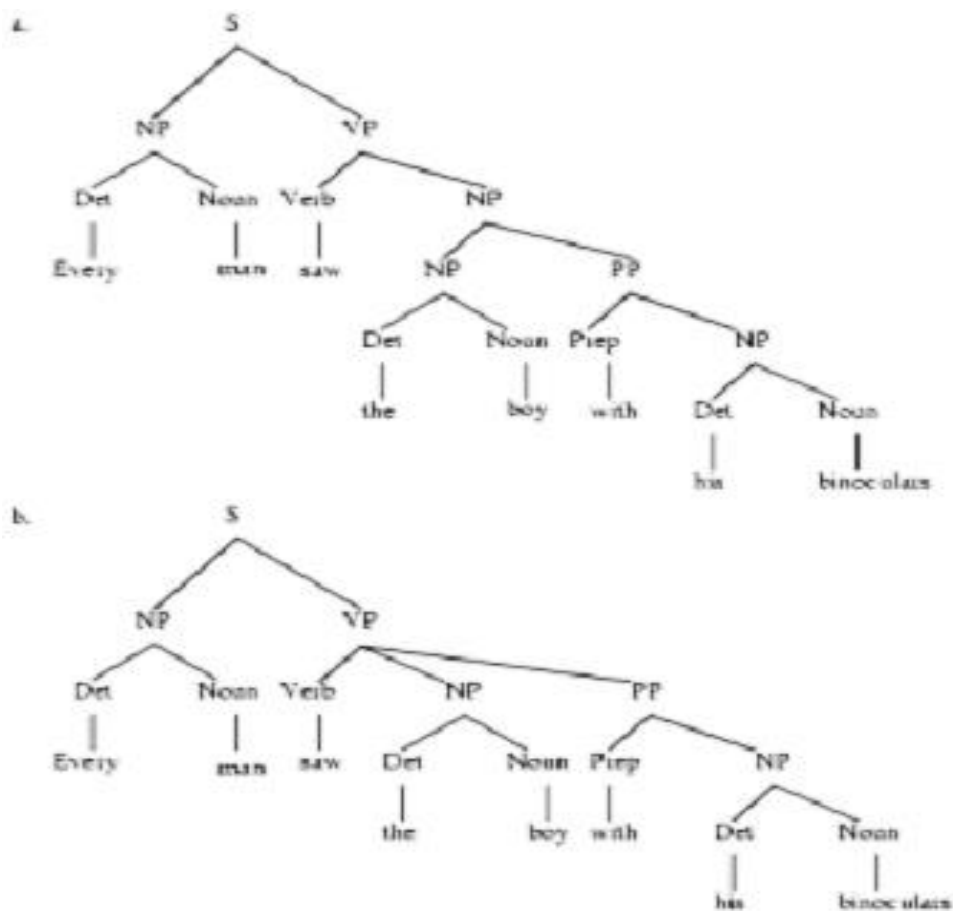
- Zarezerwuj mi coś w przyszły weekend, gdzieś gdzie jest ciepło, nie za daleko, gdzie mówią po francusku lub po angielsku.
- Znajdź informacje o zwierzętach, które wykorzystują sonar, ale nie są to nietoperze ani delfiny.
- Chciałbym wynająć mieszkanie z jedną sypialnią, w dobrym stanie, 5 km od centrum za około 800 euro miesięcznie.

Cechy języka naturalnego

Sama znajomość gramatyki nie jest wystarczająca:

Every man saw the boy with his binoculars

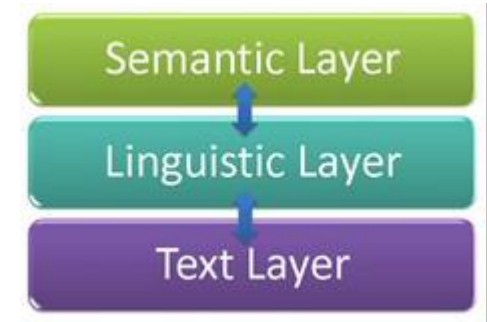
Rozbiór zdania:



Cechy języka naturalnego

Język naturalny

- **Nieprecyzyjny (na wszystkich poziomach)**
 - **Fonetyka**: dźwięki i wypowiedzane słowa
 - **Morfologia**: słowa (fleksja, formowanie słów)
 - **Składnia**: zdania i ich struktura
 - **Semantyka**: znaczenie treści wypowiedzi (relacje leksykalne, np. synonimy, antonimy, hiponimy, itd.) -> WordNet
 - **Frazeologia**: kontekst i typowe użycia słów: idiomy, zwroty, phrasal verbs
 - **Pragmatyka**: znaczenie samej wypowiedzi „w świecie”
- Skomplikowany (nawet jeśli uznać reguły gramatyczne)
- **Wymaga posiadania wiedzy o świecie**
- **Narzędzia**
 - Wiedza o języku
 - Wiedza o świecie
 - Sposób na ich połączenie

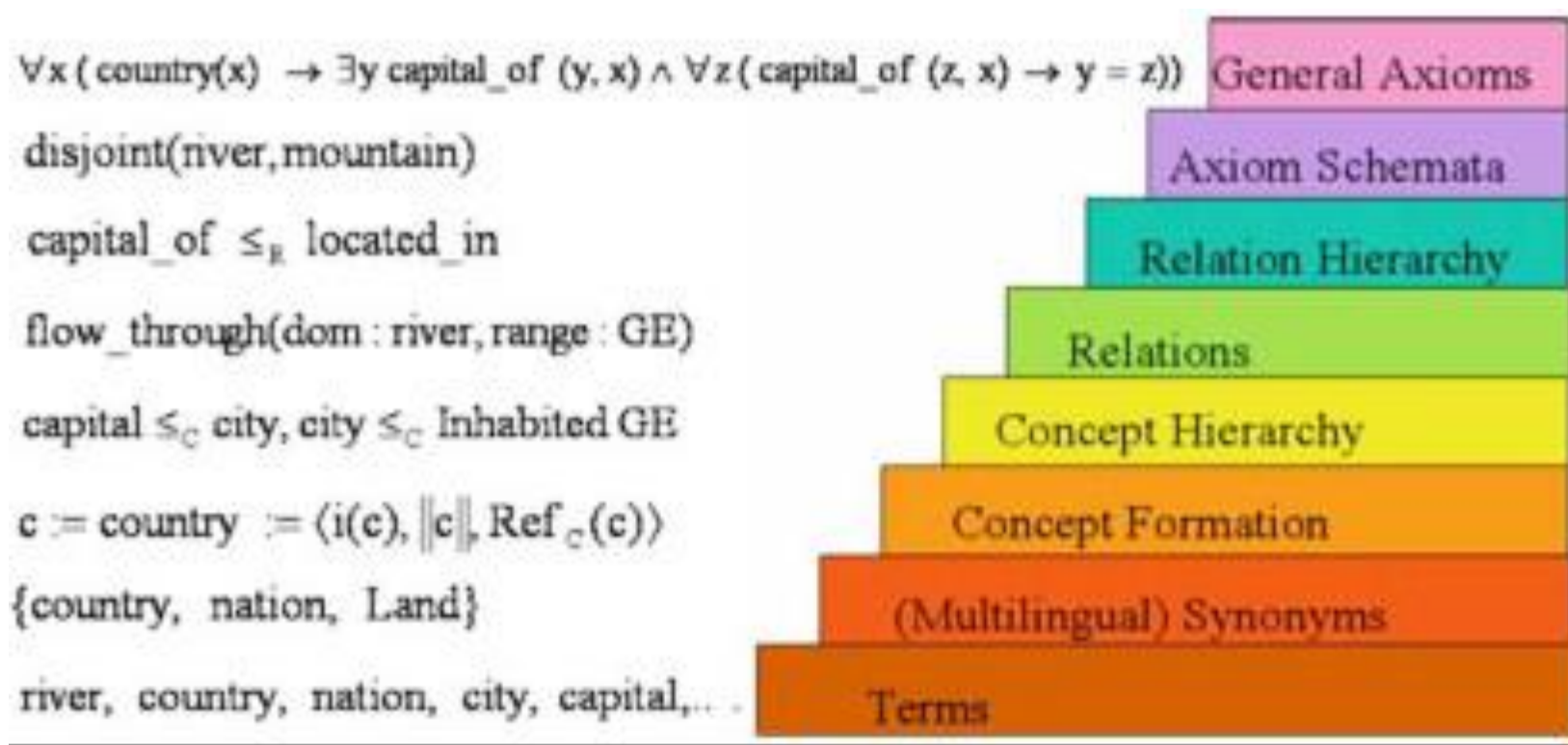


Ale czasami NLU jest łatwe – czasem (rzadko...) same informacje ilościowe o tekście wystarczają

Ostatnio modny trend - modele probabilistyczne

- $P(\text{"maison"} \rightarrow \text{"house"})$ **wysoko prawdopodobne**
- $P(\text{"L'avocat general"} \rightarrow \text{"the general avocado"})$ **nisko**

Wiedza o świecie → NLU



Ontology learning layer cake


In: Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text. Paul Buitelaar, Philipp Cimiano (Eds.), IOS Press, p. 225-249, January 2008.

Dwa podejścia w NLP

„Gramatyczne”

- Język naturalny można opisać wykorzystując aparat logiki matematycznej
- Lingwistyka porównawcza – Jakob Grimm, Rasmus Rask
- Noam Chomsky – I-Language i E-language
- Argument „*poverty of stimulus*”

„Statystyczne”

- Przekonanie, iż struktura i reguły użycia słów w języku naturalnym można odkryć, analizując rzeczywiste wypowiedzi
 - Najlepiej analizować dużo wypowiedzi...
 - Bardzo dużo wypowiedzi...
- 
- Statystyka
 - Pierwsze próby – Markow /łańcuchy Markowa/, Shannon

Przykład metody statystycznej

Ujednoznacznianie znaczenia słów

(ang. word sense disambiguation - WSD):

They put the money in the **bank**

*River **bank**?*

*Savings **bank**?*

Potrzebny jest korpus poprawnych tekstów w języku angielskim. Na jego podstawie należy obliczyć prawdopodobieństwa:

P_1 – współwystępowanie <money, savings>

P_2 – współwystępowanie <money, river>

$$P_1 > P_2$$

Nieco historii

1900 – początki

- eksperymenty w logice matematycznej, automatyczne dowodzenie twierdzeń (to jeszcze plan Hilberta), *formalna teoria języka* – Tarski, Russel, Wittgenstein
- łańcuchy Markowa, rozwój statystyki

1940-1950 – lingwistyka „empiryczna” (Harris, Firth)

- „*You shall know a word by a company it keeps*” – Firth
- *Model kanału transmisyjnego* (Shannon)

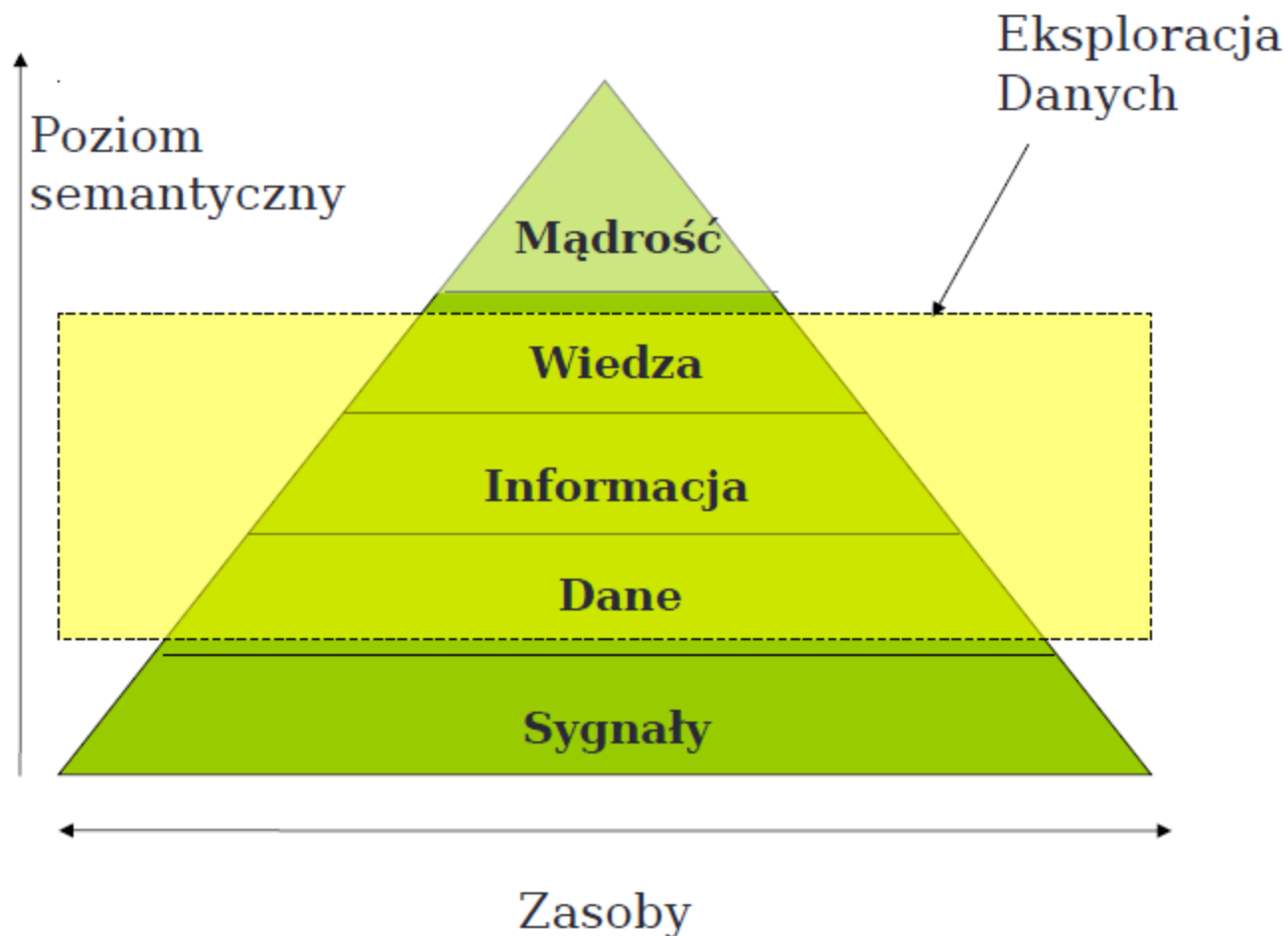
1950-1980 – statystyka uznana za „nieśluszną”

- Chomsky, pojęcie „gramatyczności”
- „*Every day I fire a linguist my efficiency goes up*”
- *Gramatyka symboliczna* (Prolog)

od 1980 – powrót do metod statystycznych

- *Rozwój metod eksploracji danych opartych na statystyce*
- *Wpływ badań nad rozpoznawaniem mowy (IBM)*

Piramida wiedzy



Piramida wiedzy w TM

- **Dane** (ang. data) - poziom składni, słowa, liczby, wyrażenia.
- **Informacje** - poziom semantyczny; reprezentacja faktów, operacje, projekty, zasoby, itp.
- **Wiedza** - poziom ontologiczny - informacja skategoryzowana w celu opisu zasobów informacyjnych.
- **Mądrość** – jak najlepiej wykorzystać wiedzę, podejmowanie właściwych działań

Eksploracja danych tekstowych (ang. text mining TM)

Eksploracja danych tekstowych - proces automatycznego odkrywania obowiązującej, dotychczas nieznanej, użytecznej oraz zrozumiałej dla użytkownika informacji/wzorców/wiedzy z repozytorium dokumentów tekstowych.

Eksploracja danych tekstowych
=
Metody eksploracji danych
+
Klasyczne NLP

Obecna sytuacja (metody)

Zarówno statystyka jak i podejście „gramatyczne”

- „czyste” modele oparte na regułach można wzbogacić o elementy probabilistyczne (np. PCFG)
- metody statystyczne można wzbogacić poprzez wykorzystanie znanych reguł i źródeł „twardej” wiedzy (np. ujednoznacznianie znaczenia słów + słowniki + ontologie)

Dodatkowo znaczenie zyskują źródła informacji nietekstowej, związane m.in. z WWW:

- Analiza grafów hiperpołączeń
- Analiza formatowania tekstu
- Analiza ruchu w sieci Internet
- ...

Zadania eksploracji danych tekstowych

- Klasyfikacja
- Grupowanie
- Wykrywanie różnego rodzaju związków
 - taksonomicznych (relacja „is-a”),
 - nietaksonomicznych (synonimy, homonimy, kolokacje).
- Ujednoznacznianie znaczenia słów
- ...

Lingwistyka

- **Sposób opisu języka**
- **Gramatyka – zbiór reguł** opisujących **formy słów** i ich **współwystępowanie** dopuszczalne w danym języku
- Gramatyka klasyczna
 - Przeznaczona dla ludzi (najlepiej znających dany język)
 - Reguły zwykle oparte na przykładach, także wyjątki od reguł
 - Zwykle nie jest sformalizowana, nie istnieją narzędzia (matematyczne, IT) które ją rozumieją
- Gramatyki formalne
 - (CFG, LFG, GPSG, HPSG, ...)
 - Opis formalny
 - Sprawdzalne na danych (korpusach tekstowych)

Poziom opisu języka (1)

- Poziomy opisu języka
 - Fonetyka
 - Fonologia
 - Morfologia
 - Składnia
 - Semantyka
 - Pragmatyka
- Każdy z poziomów możemy interpretować jako filtr, posiadający wejście (od poziomu niższego) i wyjście (do poziomu wyższego)
 - *Oczywiście nie zawsze interesuje nas przejście od fonetyki do pragmatyki*

Poziom opisu języka (2)

Fonetyka – badanie dźwięków mowy ludzkiej (głosek): artykulacja, cechy akustyczne, odbiór, reakcje psychiczne jakie wywołują

Fonologia – badanie funkcjonowania dźwięków w języku i systemu, który tworzą

Wejście: sygnał mowy

Wyjście: ciąg głosek, ciąg liter

Problemy

- Głos każdego człowieka daje nieco inny sygnał, wydzielenie sygnału mowy z szumu (który może zawierać inne rozmowy), intonacja itp.
- Klasyfikacja głosek – samogłoski, spółgłoski
- W wielu językach - trudna reprezentacja tekstowa głosek
- Błędy w wypowiedzi, gwary języka itp.
- Konieczność określenia przerw pomiędzy wyrazami

Poziom opisu języka (3)

- **Morfologia** - formy odmiennej części mowy (fleksja) oraz słowotwórstwo
- **Składnia, syntaktyka** - budowa wypowiedzi: funkcje wyrazów w zdaniu, zależności między wyrazami w zdaniu
- **Semantyka** - znaczenie w języku, relacje między znaczeniem podstawowym wyrazu, a jego znaczeniem w konkretnym wypowiedzeniu
- **Pragmatyka** - sposoby posługiwania się mową przez ludzi, rozumienie i interpretowanie wypowiedzi w zależności od kontekstu

Morfologia

- **Morfologia** - formy odmiennej części mowy (fleksja) oraz słowotwórstwo
 - **Wejście**: Sekwencja głosek (słowo)
 - **Wyjście**: Sekwencja oznakowanych morfemów
- **Morfem – niepodzielna część znaczeniowa wyrazu**, nie można więc podzielić na mniejsze jednostki znaczeniowe, elementarna jednostka morfologii
- **Morfemy leksykalne (rdzenne)** – **rdzenie wyrazów**: samodzielne lub związane, obecne w każdym **leksemie**
- **Morfemy poboczne**:
 - **Morfemy słowotwórcze** - **prefiksy** (np. *za-* w *zanieść*), **interfiksy** (np. *-o-* w *parowóz*), **sufiksy** (np. *-ek* w *kotek*))
 - **Morfemy fleksyjne (gramatyczne)** służące do reprezentacji odmiany słowa (np. *-em* w *kotem*)

Morfologia (2)

- **Rodzina wyrazów** – **grupa wyrazów** mających **wspólne pochodzenie**, czyli wywodzących się od jednego wyrazu podstawowego
np. *dom* (rdzeń), *domowy*, *domownik*, *przydomowy*, *udomowiony*, *domek*, *domeczek*, *domostwo*, *bezdomny*, *podomka*, *domofon*
- **Leksem** – oznaczenie wszystkich form fleksyjnych danego słowa (W potocznym sensie na oznaczenie leksemu używa się nazwy słowo.)
np. *czytać*, *czytam*, *czytali*, *przeczytasz* – formy tego samego leksemu
- **Lemma** - kanoniczna, podstawowa forma leksemu, którą najczęściej podaje się w słownikach.

Morfologia (3)

W języku mamy do czynienia z **procesem morfologicznym**, który **tworzy** nam nowe **słowa** i nowe **formy słów**

To co zwykle chcemy uzyskać to informacja, jaką **część mowy** stanowi dany wyraz.

Fleksja

Fleksja – nadanie znaczenia rdzeniowi wyrazu za pomocą przyrostków i przedrostków zmieniająca liczbę, rodzaj, przypadek (w tych językach, w których występują przypadki) itd. ale niezmieniająca części mowy

np. dog → dog-s, chodz-ić → chodz-ę

Fleksja (2)

Angielska fleksja jest bardzo prosta...

- **rzeczowniki** – liczba mnoga, possessive
- **czasowniki** – w zależności od używanego czasu

ale niektóre słowa są nieregularne np.

- regularne: np. walk, walks, walking, walked, walked
- nieregularne
 - Eat, eats, eating, **ate**, **eaten**
 - Catch, catches, catching, **caught**, **caught**
 - Cut, cuts, cutting, **cut**, **cut**

także rzeczowniki: mouse/mice, goose/geese, ox/oxen

Powyższe problemy komplikują zastosowania takie jak wyszukiwanie informacji – nie można zastosować zwykłego dopasowywania wzorców, ani wyrażen regularnych.

Polska fleksja natomiast jest...

Np. czasownik JEŚĆ

Ja jem – od jajo?

Ty jesz – od tyć?

On je – je to biernik liczby mnogiej zaimka ona, hmmm...???

My jemy – od myć?

Wy jecie – od wyć?

Oni jedzą – od jechać, bo:

My jedziemy

Wy jedziecie

Oni jedzą

Źródło: Julian Tuwim, Cicer cum Caule czyli groch z kapustą, Panopticum i Archiwum Kultury

Przetwarzanie tekstów

- Najczęściej pierwotny tekst nie nadaje się do automatycznej analizy.
- Przed właściwą analizą konieczne jest przetworzenie tekstu.
- W przypadku analizy dokumentów tekstowych często zachodzi potrzeba utworzenia ich reprezentacji.

Wyrażenia regularne (ang. regular expressions)

- **Język formalny specyfikujący ciągi znaków**
- **Są w wielu miejscach**
 - emacs, vi, perl, python, grep, sed, awk,...
- **Elementy wyrażen regularnych**
 - Ciągi znaków
 - Kleene star
 - Zbiór znaków, dopełnienie zbioru
 - Kotwice
 - Zakres
 - Alternatywa
 - Grupowanie

<https://www.regular-expressions.info/>

Wyrażenia regularne (1)

Wielkość liter ma zwykle znaczenie (*case sensitive*)

Ciągi

[wW]oodchuck

Woodchuck lub
woodchuck

Woodchuck

[abc]

a, b lub c

In uomini, in soldati

[1234567890]

Dowolna cyfra

Plenty of 7 to 5

Zakres

[A-Z]

Wielka litera

we call it „A great

[a-z]

Mała litera

my dear

[0-9]

Dowolna cyfra

Chapter 1: in

[A-Za-z]

litera

Wyrażenia regularne (2)

Dopełnienie

[^A-Z]

Nie wielka litera

Woodchuck

[e^]

e lub ^

Look up ^ now

a^b

Ciąg a^b

Look up a^b now

Znaki opcjonalne

woodchucks?

woodchuck lub
woodchucks

woodchuck

colou?r

color lub colour

colour

Kleene *

Zero lub więcej powtórzeń poprzedzającej sekwencji

[ab]* - aaaa, bbbb, abababbba, bbabaaab

Wyrażenia regularne (3)

Alternatywa i grupowanie

`cat | dog`

cat lub dog

cat

`gupp(y | ies)`

guppy lub guppies

guppy

`(Column_[0-9]+_*)*`

Column 1 Column 2 itd.

Kotwice

- `^` - początek ciągu
- `$` - koniec ciągu
- `\b` - granica słowa
- `\B` - środek słowa

Kleene +

- Przynajmniej jedno wystąpienia sekwencji
- `[0-9]+` - liczba całkowita

Wyrażenia regularne - Hierarchia operatorów

- | | |
|----------------|---------------|
| 1. Grupowanie | () |
| 2. Liczniki | * + ? {} |
| 3. Kotwice | the ^my end\$ |
| 4. Alternatywa | |

{n} – n wystąpień sekwencji

{n,m} – od n do m wystąpień

{n, } - przynajmniej n wystąpień

Character escaping – np. *, \. itd.

beg.n

W miejscu kropki -
dowolny znak

begin, begun, beg3n

Wyrażenie regularne - przykład

- Znaleźć wystąpienie rodzajnika „the”

The recent attempt by **the** police to retain their current rates of pay has not gathered much favor with **the** southern factions.

1. Wyrażenie: the

The recent attempt by **the** police to retain **their** current rates of pay has not gathered much favor with **the** southern factions.

2. Wyrażenie: [Tt]he

The recent attempt by **the** police to retain **their** current rates of pay has not gathered much favor with **the** southern factions.

3. Wyrażenie `\b[Tt]he\b` lub `^[^A-Za-z][Tt]he[^A-Za-z]`

The recent attempt by **the** police to retain their current rates of pay has not gathered much favor with **the** southern factions.

Wielkość liter

Zwykle dla dalszego przetwarzania NLP wielkość liter nie ma znaczenia:

THE = The = the

Co jednak z wielkimi literami w nazwach własnych, na początku zdań?
w zasadzie wypadałoby oznaczać wystąpienie wszystkich nazw własnych – to jednak wymaga posiadania ich słownika, aby było 100% dokładne.

Prosta heurystyka – zamieniamy na małe litery początki zdań oraz słowa pisane wyłącznie wielkimi literami – w ten sposób pozostawiamy wielkość liter w nazwach własnych, ale:

- potrzebujemy algorytmu wykrywania końców zdań (co może nie być łatwe),
- nie zawsze słowa pisane wielką literą będą nazwami własnymi (np. „*Ogon to dla nich nie ogon, tylko Mały Dodatkowy Kawałeczek przyczepiony z tyłu.*” – Kłapouchy „Praca porusza tematykę związaną z obliczaniem stałej Gardentoffa. Zastosowaną metodę Autor ocenia jako najlepszą z dotychczas zaproponowanych.”)

Podział tekstu na słowa (1)

podejście naiwne – słowa są ciągami znaków alfabetycznych, oddzielonych od innych słów białymi znakami, mogą zawierać także apostrofy i myślniki – Kucera, Francis

- nie działa np. dla *Micro\$oft*, *C/net*, *23.13\$*, itd.

kropka – słowa nie zawsze są oddzielone białymi znakami, czasami po słowie występuje kropka:

- skróty (ale uwaga – wewnątrz skrótu może być więcej kropek) – *Inż.*, *itd.*, *U.S.*
- kropki zwykle pojawiają się na końcu zdań

Podział tekstu na słowa - apostrof

apostrof – szczególne problemy w języku angielskim, apostrof może mieć znaczenie gramatyczne

- *I'll* -> *I will* – to muszą być dwa oddzielne słowa, morfologicznie nie jest bowiem możliwe złączenie czasownika i zaimka
- forma dzierżawcza – *Peter's, boys'*
- zwykle przyjmuje się iż apostrof jest formą słowa, wtedy:
 - *I'll* -> *I + ' + ll*
- niestety apostrofy mogą się także pojawić jako znaczniki cytowania
 - trzeba odróżnić *boys'* od *she said 'hello boys'*

Podział tekstu na słowa - myślnik

myślnik – zwykle ma jedną z trzech głównych funkcji

- dzielenie słów przy formatowaniu – występują, gdy pozyskujemy tekst do korpusu z materiału drukowanego, mogą wtedy mylić się z pozostałymi dwoma formami;
- oddzielenie poszczególnych morfemów w obrębie leksemu np. *co-operative, e-mail, pro-Arab*;
- jako łączniki oddzielnych słów tworzących związek frazeologiczny np. *once-in-a-lifetime, text-based, 26-year-old* (przykład zdania: *the once-quiet study of superconductivity...*

Nawet w języku literackim nie ma stałych reguł dotyczących użycia myślników:

- wszystkie formy: *database, data base* oraz *data-base* są poprawne – czy stanowią różne sposoby zapisania pojedynczego leksemu?

Myślniki mogą być używane zamiast białych znaków do oddzielenia części zdania
np.: *I am happy-Bill is not.*

Podział tekstu na słowa – białe znaki

- **nie zawsze są używane do podziału zdań na słowa np.**

język chiński – zupełny brak podziału na słowa

Waterloo znajduje się na południe od Kanady

滑铁卢位於加拿大南部。

język niemiecki – niektóre rzeczowniki zapisywane bez spacji

Lebensversicherungsgesellschaftsangestellter – pracownik firmy ubezpieczeniowej

- **czasami pojawiają się w środku słowa (leksemu)**
 - nazwiska, skróty: *Mr. John Smith, New York, U. S. A.*
 - idiomy: *work out, make up*
 - numery telefonów: *+48 (22) 67728911*
 - nazwy własne: *Politechnika Warszawska*
- **problem formatów danych jest ogólniejszy**

np. jak rozpoznać zapis liczby?

Angielski 123,456.78

[0-9]((([0-9]+[,]))([.][0-9]+)*

Francuski 123 456,78

[0-9]((([0-9]+[])) ([,][0-9]+)*

wyrażenia regularne

Podział na zdania (1)

(ang. sentence boundary detection)

podejście naiwne:

zdanie to ciąg znaków zakończony '.', '!', lub '?',

ale...

- kropki występują także w skrótach;
- zdania złożone zawierają także '-', ';', ':' itp.;
- zdania mogą mieć strukturę hierarchiczną np.
„You remind me”, she remarked, „of your mother”.

to są tak naprawdę dwa zdania



Podział na zdania (2)

podejście lepsze – heurystyka:

- 1) wstępne podziały zdań po . ? !.
- 2) uwzględnienie cudzysłowów występujących po powyższych zdaniach - przesunąć granicę po cudzysłowach.
- 3) skasowanie podziału -jeśli:
 - a) jeśli jest poprzedzony znanym skrótem, po którym występuje zwykle nazwa własna – np. Prof. lub vs.
 - b) jeśli jest poprzedzony znanym skrótem, po którym nie występuje słowo rozpoczęte wielką literą
 - c) jeśli podział zdania wynikał z wystąpienia ‘!’ lub ‘?’ oraz następuje po nim mała litera

Podział na zdania (3)

Jeszcze lepsze podejście - klasyfikatory

- Drzewa decyzyjne (Riley, 1989) - analiza częstości występowania słów przed i po końcach zdań a także długość i wielkość liter słów.
- Sieci neuronowe (Hearst, 1997) – analiza występowania części mowy słów przed i po końcach zdań.
- Obecnie jakość wykrywania zdań (rozumianego jako klasyfikacja) ~ 99% dla języka angielskiego.

Podział na zdania (4)

- Czasami trudno określić co jest pojedynczym zdaniem (w kontekście analizy tekstu):
 - cytowanie całych zdań
 - wyliczenia

Podział na paragrafy

Założenie: tekst jest podzielony na linie

Heurystyka:

- Linia zawierająca mniej znaków niż podany próg kończy paragraf.
- Linia występująca po linii końca paragrafu lub linii pustej rozpoczyna paragraf.

Zastosowanie

- zwiększenie kontekstu
- utworzenie nowych dokumentów

Błędy ortograficzne (1)

Tekst, który analizujemy w NLP, nie jest zwykle generowany przez maszynę (*natural* language) – może zawierać błędy.

Błędy ortograficzne – zwykle drobne

- 80% wszystkich błędów ortograficznych dotyczy pojedynczej litery (Damerau, 1964)
 - **Wstawienie (ang. *insertion*)** – the -> ther
 - **Skasowanie (ang. *deletion*)** – the -> th
 - **Podstawienie (ang. *substitution*)** – the -> thw
 - **Transpozycja (ang. *transposition*)** – the -> hte

Błędy ortograficzne (2)

Wiele zależy od źródła danych – wpływ układu klawiatury, gdy tekst wpisywany ręcznie, wpływ wyglądu liter, gdy OCR

Rodzaje błędów

- **Non-words:** giraffe -> graffe
- **Isolated errors:** bez kontekstu
- **Real-words:** piece of cake -> peace of cake

Błędy ortograficzne – poprawianie słów

- **Metody probabilistyczne** - bazujące na korpusach tekstów (np. Bayesowska)

t – błędny (obserwowany) wyraz, c – poprawiony wyraz

$$\hat{c} = \operatorname{argmax}_{c \in V} P(c|t) = \operatorname{argmax}_{c \in V} P(t|c)P(c)$$

- **Metody słownikowe** – oparte o odległość edycyjną (np. odległość Levenshteina)