

Wprowadzenie i podstawowe techniki

Agnieszka Ławrynowicz

Wydział Informatyki Politechniki Poznańskiej

27 lutego 2018

Prowadzący

Wykłady

dr inż. Agnieszka Ławrynowicz
alawrynowicz@cs.put.poznan.pl

Laboratoria

mgr inż. Dawid Wiśniewski
dwisniewski@cs.put.poznan.pl

Podstawowa:

- ① Natural Language Processing with Python, Steven Bird, Ewan Klein, and Edward Loper, O'Reilly Media, 2009, dostępna online <http://www.nltk.org/book/>
- ② Speech and Language Processing (3rd ed. draft), Dan Jurafsky and James H. Martin. Draft chapters in progress, August 28, 2017, dostępna online <https://web.stanford.edu/~jurafsky/slp3/>

Wprowadzenie

Czym jest przetwarzanie języka naturalnego?

- Przetwarzanie języka naturalnego (ang. *Natural Language Processing (NLP)*) jest dziedziną na przecięciu:
 - informatyki,
 - sztucznej inteligencji,
 - lingwistyki.

Czym jest przetwarzanie języka naturalnego?

- **Cel:** aby komputery mogły przetwarzać lub “rozumieć” język naturalny w celu wykonywania przydatnych zadań, takich jak np. automatyczne odpowiadanie na pytania czy maszynowe tłumaczenie.
- Nie chodzi o pełne zrozumienie i reprezentację języka, bo zadanie perfekcyjnego zrozumienia języka należy do zadań tzw. silnej sztucznej inteligencji (*AI-complete*).

Postęp technologii przetwarzania języka

w większości rozwiązane

Detekcja spamu

Oznaczanie części mowy
(POS-tagging)

Rozpoznawanie jednostek
referencyjnych (NER)

zrobiono duże postępy

Analiza sentymentu

Analiza współwystępowania

Ujednoznaczniania znaczeń słów

Parsowanie

Maszynowe tłumaczenie

Ekstrakcja informacji

ciągle bardzo trudne

Odpowiadanie na pytania

Parafrazowanie

Streszczerwanie

Dialog

Odpowiadanie na pytania



WolframAlpha computational knowledge engine.

How much Calcium in skim milk?

Assuming skim milk | Use Ahold slim delux calcium-enriched fat free milk or [more](#) instead
Assuming any type of skim milk | Use skim milk, Great or [more](#) instead

Input interpretation:
 skim milk amount 1 cup calcium

Average result:
325 mg (milligrams)

Unit conversions:
0.33 grams
 3.3×10^{-4} kg (kilograms)

"Where is Poznań"

OK, here's Poznań:

MAPS

Poznań
10 km

YOU CAN ALSO TRY

"Wikipedia Poznan"
"Directions to the city Poznan"
"Find news about Poznan"

Ekstrakcja informacji

Temat: **spotkanie projektowe**

Data: Luty 27, 2018

Do: Jan Kowalski

Cześć Janie,

Zaplanowaliśmy kolejne spotkanie projektowe. Odbędzie się w sali 2.7.13 jutro w godzinach 13:30-14:30.

--

Ania

Ekstrakcja informacji

Temat: **spotkanie projektowe**

Data: Luty 27, 2018

Do: Jan Kowalski

Cześć Janie,

Zaplanowaliśmy kolejne spotkanie projektowe. Odbędzie się w sali
2.7.13 jutro w godzinach 13:30-14:30.

--

Ania



Ekstrakcja informacji

Temat: **spotkanie projektowe**

Data: Luty 27, 2018

Do: Jan Kowalski

Cześć Janie,

Zaplanowaliśmy kolejne spotkanie projektowe. Odbędzie się w sali
2.7.13 jutro w godzinach 13:30-14:30.

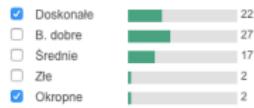
--

Ania

Wprowadź nowe
wydarzenie do
kalendarza

Wydarzenie: spotkanie projektowe
Data: 28-02-2018
Początek: 13:30
Koniec: 14:30
Miejsce: 2.7.13

Analiza sentymantu



Zobacz opinie podróżnych:



 Wyszukaj recenzje

1 z 10 wśród 24 recenzji



Zrecenzowano wczoraj

Rai dla narciarzy

Duża ilość wyciągów, różny stopień trudności. Minusem jest to, że chodniki dla najmłodszych są tylko do nauki z instruktorem i samemu nie można tam dziecka uczyć. Dość wysoki koszt lekcji w porównaniu z innymi szkołami. Duże parkinie.



edytakrakow
Krakow,
Poland

5 / 1

Zadaj pytanie dotyczące obiektu Czarna Góra- Narty

 1 Dziękujemy, edytatorki!

Ta recenzja jest prywatną opinią członka witryny TripAdvisor i nie stanowi opinii firmy TripAdvisor LLC.



Zrecenzowane 1 tydzień temu

Walka o przetrwanie, bezmyślność, nigdy wiecej tam nie pojade

To nie był mity dzień na stoku, ale walka o przetrwanie... Ponad połowa narciarzy (nawet na czarnej trasie) nie wie, jak należy zachować się na stoku i nigdy nie miała nawet godzinnej lekcji z instruktorem. W kompleksie nie zauważalam ani jednej widocznej informacji (dla tych... [\[Wiecej\]](#)



Zadaj pytanie dotyczące obiektu Czarna Góra- Narty

 1 Dziękujemy Ewa B.

Maszynowe tłumaczenie

Tłumacz

Wyłącz szybkie tłumaczenie



polski angielski niemiecki Wykryj język ▾



angielski polski niemiecki ▾

Przetłumacz

Nic dwa razy się nie zdarza i nie zdarzy. Z tej przyczyny zrodziliśmy się bez wprawy i pomrzemy bez rutyny.



Nothing happens twice and it will not happen. Of this reason we were born without practice and we will die without routine.



108/5000

Detekcja "Fake news"

HOW TO SPOT FAKE NEWS

CONSIDER THE SOURCE
Click away from the story to investigate the site, its mission and its contact info.

READ BEYOND
Headlines can be outrageous in an effort to get clicks. What's the whole story?

CHECK THE AUTHOR
Do a quick search on the author. Are they credible? Are they real?

SUPPORTING SOURCES?
Click on those links. Determine if the info given actually supports the story.

CHECK THE DATE
Reposting old news stories doesn't mean they're relevant to current events.

IS IT A JOKE?
If it is too outlandish, it might be satire. Research the site and author to be sure.

CHECK YOUR BIASES
Consider if your own beliefs could affect your judgement.

ASK THE EXPERTS
Ask a librarian, or consult a fact-checking site.

International Federation of Library Associations and Institutions



Niejednoznaczność powoduje trudność przetwarzania

Teacher Strikes Idle Kids

One morning I shot an elephant in my pajamas

The burglar threatened the student with the knife

We saw her duck

Inne powody trudności przetwarzania języka naturalnego

- niestandardowy język (np. na Twitterze)
- neologizmy (np. retweet)
- idiomy (np. “urwanie głowy”)
- brak zakodowanej wiedzy dziedzinowej

Poziomy przetwarzania języka naturalnego



Wyrażenia regularne

Wyrażenia regularne

Wyrażenia regularne

Formalny język, wzorce do opisułańcuchów symboli.

Jak wyszukać każdego z tych słów?

kot

Kot

koty

Koty



Wyrażenia regularne: alternatywa

Metaznak	Znaczenie	Przykład	Łańcuch zgodny z wyrażeniem
[...]	dowolny z podanych znaków	[kK]o[st]	kot , Kos
[^...]	dowolny poza podanymi znakami (jeśli na początku wyrażenia)	[^kK]ot	bot , lot
[m – n]	zakres	[a^b] [a-z][a-z0-9][a-z0-9]	a , ^ , b mp3 , p2p

Wyrażenia regularne: alternatywa c.d.

Metaznak	Znaczenie	Przykład	Łańcuch zgodny z wyrażeniem
	rurka, jedno z dwóch słów	kot pies	<p>kot pies</p> <p>= [xyz]</p>

Wyrażenia regularne: ? * + .

Metaznak	Znaczenie	Przykład	Łańcuch zgodny z wyrażeniem
?	0 lub 1 wystąpienie poprzedzającego znaku	koty?	koty
*	0 lub wiele wystąpień poprzedzającego znaku	wo*w!	kot woow! woooow!
+	1 lub wiele wystąpień poprzedzającego znaku	a+	aa aaa
.	zastąpienie 1 dowolnego znaku	k.t wa..a	kot wanna wart

Wyrażenia regularne: kotwice ^ \$

Metaznak	Znaczenie	Przykład	Łańcuch zgodny z wyrażeniem
^	rozpoczyna się od	^[A-Z] ^ko	Poznań kot kosa
\$	kończy się na	niec\$	koniec wieniec

Przykład: znajdź instancje wyrazu kos

kos

[kK]os[^a-zA-Z]
[kK]os

nie znajduje instancji zaczynających się od dużej litery

nieprawidłowo zwraca wyrazy takie jak kosa

ok



Typowe błędy w przetwarzaniu języka naturalnego

Dwa rodzaje:

- pasujące łańcuchy, których nie powinniśmy dopasować (kosa)
 - Wyniki **fałszywie pozytywne**
 - Minimalizujemy je poprzez zwiększenie trafności lub precyzji
- Nie pasujące łańcuchy, które powinniśmy dopasować (Kos)
 - Wyniki **fałszywie negatywne**
 - Minimalizujemy je poprzez zwiększenie pokrycia lub czułości

Zmniejszenie błędu w konkretnej aplikacji często wiąże się więc z dwoma przeciwnymi celami.

Wyrażenia regularne: podsumowanie

- Zaawansowane wyrażenia regularne są często testowane jako pierwszy model w celu przetworzenia dowolnego tekstu
- Do trudnych zadań używamy uczenia maszynowego (np. klasyfikatorów). Wtedy wyrażenia regularne mogą być używane jako cechy.

Korpusy

Korpus

Korpus

Odpowiednio duży i uporządkowany **zbiór tekstów**, zazwyczaj zdatny do elektronicznego przetwarzania, który może służyć analizom lingwistycznym (np. badaniu częstości występowania słów, konstrukcji składniowych itp.).

Lista rangowa

Lista rangowa

Lista wyrazów posortowanych w malejący sposób względem liczby wystąpień danego wyrazu w korpusie (tekście).

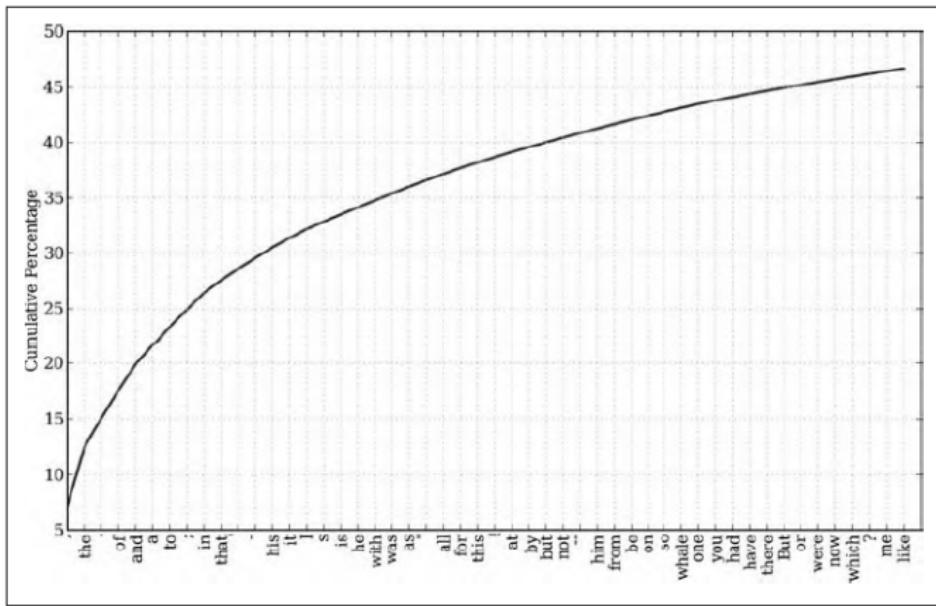
Częstość

Liczba wystąpień wyrazu w korpusie (tekście).

Ranga

Liczba porządkowa danego wyrazu na liście rangowej.

Lista rangowa: przykład



Rysunek: Skumulowany wykres częstości dla 50 najczęściej używanych wyrazów w powieści "Moby Dick", które stanowią prawie połowę tokenów (elementarnych wyrazów)¹

¹ źródło: Natural Language Processing with Python, S. Bird, E. Klein, E. Loper, O'Reilly Media, 2009.

Prawo Zipfa

Prawo Zipfa

Częstość wyrazu jest odwrotnie proporcjonalna do jego rangi.

Prawo Zipfa

Przyjmijmy oznaczenia:

f -częstość

r -ranga

Powinna istnieć taka stała k , że: $f \cong \frac{k}{r}$

Konsekwencje prawa Zipfa

- najczęściej występujące wyrazy pojawiają się w większości tekstów
- niektóre z wyrazów występują w tekście tylko raz ("długie ogony")
- znaczenie tekstu reprezentowane jest poprzez średnio często i rzadziej występujące wyrazy

Tokenizacja

Tokenizacja

Jeden z początkowych etapów w procesie przetwarzania języka naturalnego, polegający na podziale tekstu na **tokeny** (ciągi znaków oddzielone znakami zdefiniowanymi jako separatorzy).

Tokeny są w jakimś sensie **elementarnymi** "wyrazami".

Typ wyrazu

Typ wyrazu

Forma lub pisownia wyrazu niezależnie od jego konkretnych wystąpień w tekście, tzn. wyraz uważany za unikalny element słownika.

Token i typ: podsumowanie

Token: każdy wyraz pojawiający się w tekście/korpusie

Typ: unikalne wyrazy

Przykład

Korpus	Tokeny	Typy
Shakespeare	884 tys	29 tys
Alice in Wonderland	36 tys	2,5 tys
Switchboard (rozmowy telefoniczne)	2,4 mln	20 tys

Ile wyrazów?

- Kot-Dziwak → Kot; Dziwak / Kot-Dziwak?
- Poland's flag → Poland; 's; flag/Poland; ';' s; flag/Poland's; flag?
- Zielona Góra → Zielona; Góra/Zielona Góra?

Odległość edycyjna

Na ile podobne są dwa łańcuchy znakowe?

Korekta błędów pisowni: Na ile dane, błędnie zapisane słowo jest podobne do zamierzonego?

Typy błędów:

- **Dodanie:** do słowa dodano znak
 - rakijeta zamiast rakieta
- **Usunięcie:** w słowie pominięto znak
 - psychologia zamiast psychology
- **Zamiana:** jeden znak został zastąpiony innym
 - Lion zamiast Lyon
- **Transpozycja:** Dwie sąsiednie litery zostały zamienione miejscami
 - ocsypek zamiast oscypek

Odległość edycyjna (Levenshteina)

Minimalna odległość edycyjna

Minimalna liczba prostych operacji edycji potrzebnych do przekształcenia słowa1 na słowo2.

Proste operacje edycji:

- dodanie znaku
- usunięcie znaku
- zamiana znaku

Odległość edycyjna: przykłady

$$LD(\text{kot}, \text{kot}) = 0$$

$$LD(\text{kot}, \text{koc}) = 1$$

$$LD(\text{informatyk}, \text{informator}) = 2$$

Odległość Levenshteina

- dla dwóch łańcuchów znakowych:
 - X o długości n
 - Y o długości m
- definiujemy $LD(i, j)$:
 - odległość edycyjną pomiędzy $X[1 \dots i]$ a $Y[1 \dots j]$ (pierwsze i znaków X i pierwsze j znaków Y)
 - odległość edycyjna pomiędzy X i Y to $LD(n, m)$

Odległość Levenshteina

- Inicjalizacja

```
1:  $LD(i, 0) = i$ 
2:  $LD(0, j) = j$ 
```

- Równanie rekurencyjne

```
1: for  $i = 1$  to  $M$  do
2:   for  $j = 1$  to  $N$  do
3:      $LD(i, j) =$ 
         $\min \begin{cases} LD(i - 1, j) + 1 \\ LD(i, j - 1) + 1 \\ LD(i - 1, j - 1) + 1_{X_i \neq Y_j} \end{cases}$ 
         $1_{X_i \neq Y_j} = \begin{cases} 1 & \text{if } X(i) \neq Y(j) \\ 0 & \text{if } X(i) = Y(j) \end{cases}$ 
4:   end for
5: end for
```

- Zakończenie

```
1:  $LD(N, M)$  jest odległością
```

Odległość Levenshteina: działanie algorytmu

Ustalenie długości łańcuchów i utworzenie macierzy $N \times M$

k o t

t

o

m

Odległość Levenshteina: działanie algorytmu

Inicjalizacja pierwszego wiersza wartościami od 0 do N i pierwszej kolumny od 0 do M

	k	o	t
0	1	2	3
t	1		
o	2		
m	3		

Odległość Levenshteina: działanie algorytmu

Liczymy wartości dla drugiej kolumny (piewsza litera wyrazu kot):
 t różne od t , więc koszt $1_{X_i \neq Y_j} = 1$. Obliczamy minimum z: $1+1$ (komórka powyżej plus koszt), $1+1$ (komórka z lewej plus koszt) oraz $1+0$ (komórka po skosie lewa-górną plus koszt).

	k	o	t
0	1	2	3
t	1	1	
o	2		
m	3		

Odległość Levenshteina: działanie algorytmu

Liczymy wartości dla kolejnej pary (**k** i **o**): koszt $1_{X_i \neq Y_j} = 1$.
Obliczamy minimum z: 1+1 (komórka powyżej plus koszt), 2+1
(komórka z lewej plus koszt) oraz 1+1 (komórka po skosie
lewa-górną plus koszt).

	k	o	t
0	1	2	3
t	1	1	
o	2	2	
m	3		

Odległość Levenshteina: działanie algorytmu

Liczymy wartości dla kolejnej pary (**k** i **m**): koszt $1_{X_i \neq Y_j} = 1$.
Obliczamy minimum z: 2+1 (komórka powyżej plus koszt), 3+1 (komórka z lewej plus koszt) oraz 2+1 (komórka po skosie lewa-górną plus koszt).

	k	o	t
0	1	2	3
t	1	1	
o	2	2	
m	3	3	

Odległość Levenshteina: działanie algorytmu

$\min(2+1, 1+1, 1+1)$.

	k	o	t
0	1	2	3
t	1	1	2
o	2	2	
m	3	3	

Odległość Levenshteina: działanie algorytmu

$\min(2+0, 2+0, 1+0).$

	k	o	t
0	1	2	3
t	1	1	2
o	2	2	1
m	3	3	

Odległość Levenshteina: działanie algorytmu

$\min(1+1, 3+1, 2+1).$

	k	o	t
0	1	2	3
t	1	1	2
o	2	2	1
m	3	3	2

Odległość Levenshteina: działanie algorytmu

W dolnym prawym narożniku znajduje się wynikowa odległość edycyjna: 2

	k	o	t	
0	1	2	3	
t	1	1	2	2
o	2	2	1	2
m	3	3	2	2

Odległość edycyjna: inne zastosowania w przetwarzaniu języka naturalnego

- Ocena tłumaczenia maszynowego i rozpoznawania mowy:
- Rozpoznawanie jednostek referencyjnych (NER)
 - Pekao SA ogłosił dzisiaj
 - Pekao wprowadzi
- Wykrywanie plagiatów

Dziękuję za uwagę!