

Eksploracja danych i wyszukiwanie informacji w mediach społecznościowych

Wykład 4 - przetwarzanie języka naturalnego

dr inż. Julian Sienkiewicz

29 października 2018

Przetwarzanie języka naturalnego

- Natural Language Processing (NLP),

Przetwarzanie języka naturalnego

- Natural Language Processing (NLP),
- dziedzina, łącząca zagadnienia sztucznej inteligencji i językoznawstwa, zajmująca się automatyzacją analizy, rozumienia, tłumaczenia i generowania języka naturalnego przez komputer,

Przetwarzanie języka naturalnego

- Natural Language Processing (NLP),
- dziedzina, łącząca zagadnienia sztucznej inteligencji i językoznawstwa, zajmująca się automatyzacją analizy, rozumienia, tłumaczenia i generowania języka naturalnego przez komputer,
- termin **język naturalny** używany jest, by odróżnić języki ludzkie od języka formalnego czy komputerowego,

Przetwarzanie języka naturalnego

- Natural Language Processing (NLP),
- dziedzina, łącząca zagadnienia sztucznej inteligencji i językoznawstwa, zajmująca się automatyzacją analizy, rozumienia, tłumaczenia i generowania języka naturalnego przez komputer,
- termin **język naturalny** używany jest, by odróżnić języki ludzkie od języka formalnego czy komputerowego,
- najbardziej wymagające obecnie zadania: rozpoznawanie mowy, rozumienie języka naturalnego, generowanie języka,

Przetwarzanie języka naturalnego

- Natural Language Processing (NLP),
- dziedzina, łącząca zagadnienia sztucznej inteligencji i językoznawstwa, zajmująca się automatyzacją analizy, rozumienia, tłumaczenia i generowania języka naturalnego przez komputer,
- termin **język naturalny** używany jest, by odróżnić języki ludzkie od języka formalnego czy komputerowego,
- najbardziej wymagające obecnie zadania: rozpoznawanie mowy, rozumienie języka naturalnego, generowanie języka,
- część problemów jest **AI-complete**



Zarys procesu analizy tekstu

Na potrzeby tego wykładu interesuje nas głównie analiza i rozumienie (albo jakaś jego forma) tekstu. Zwykle, posiadając próbkę tekstu musimy wykonać następujące operacje (niekoniecznie wszystkie sekwencyjnie):

- tokenizacja,

Zarys procesu analizy tekstu

Na potrzeby tego wykładu interesuje nas głównie analiza i rozumienie (albo jakaś jego forma) tekstu. Zwykle, posiadając próbkę tekstu musimy wykonać następujące operacje (niekoniecznie wszystkie sekwencyjnie):

- tokenizacja,
- normalizacja,

Zarys procesu analizy tekstu

Na potrzeby tego wykładu interesuje nas głównie analiza i rozumienie (albo jakaś jego forma) tekstu. Zwykle, posiadając próbkę tekstu musimy wykonać następujące operacje (niekoniecznie wszystkie sekwencyjnie):

- tokenizacja,
- normalizacja,
- rozpoznawanie bytów,

Zarys procesu analizy tekstu

Na potrzeby tego wykładu interesuje nas głównie analiza i rozumienie (albo jakaś jego forma) tekstu. Zwykle, posiadając próbkę tekstu musimy wykonać następujące operacje (niekoniecznie wszystkie sekwencyjnie):

- tokenizacja,
- normalizacja,
- rozpoznawanie bytów,
- lematyzacja lub

Zarys procesu analizy tekstu

Na potrzeby tego wykładu interesuje nas głównie analiza i rozumienie (albo jakaś jego forma) tekstu. Zwykle, posiadając próbkę tekstu musimy wykonać następujące operacje (niekoniecznie wszystkie sekwencyjnie):

- tokenizacja,
- normalizacja,
- rozpoznawanie bytów,
- lematyzacja lub
- stemowanie,

Zarys procesu analizy tekstu

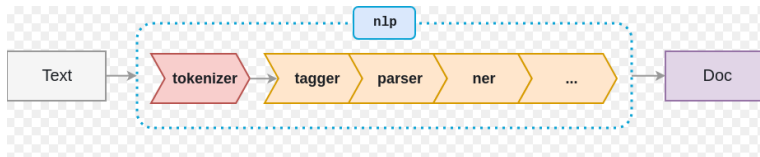
Na potrzeby tego wykładu interesuje nas głównie analiza i rozumienie (albo jakaś jego forma) tekstu. Zwykle, posiadając próbkę tekstu musimy wykonać następujące operacje (niekoniecznie wszystkie sekwencyjnie):

- tokenizacja,
- normalizacja,
- rozpoznawanie bytów,
- lematyzacja lub
- stemowanie,
- rozpoznawanie części mowy (POS),

Zarys procesu analizy tekstu

Na potrzeby tego wykładu interesuje nas głównie analiza i rozumienie (albo jakaś jego forma) tekstu. Zwykle, posiadając próbkę tekstu musimy wykonać następujące operacje (niekoniecznie wszystkie sekwencyjnie):

- tokenizacja,
- normalizacja,
- rozpoznawanie bytów,
- lematyzacja lub
- stemowanie,
- rozpoznawanie części mowy (POS),
- parsowanie (czyli w tym kontekście analizy składniowej)





Tokenizacja

- token to oczywiście “żeton”, jednak na nasze potrzeby możemy w przybliżeniu określić to jako słowo,



Tokenizacja

- token to oczywiście “żeton”, jednak na nasze potrzeby możemy w przybliżeniu określić to jako słowo,
- tokenizacja może być trywialna i sprowadzić się do zwykłego podziału na słowa na podstawie spacji je rozdzielających,



Tokenizacja

- token to oczywiście “żeton”, jednak na nasze potrzeby możemy w przybliżeniu określić to jako słowo,
- tokenizacja może być trywialna i sprowadzić się do zwykłego podziału na słowa na podstawie spacji je rozdzielających,
- jednak np. słowo "humanistycznospołeczny" może zostać rozdzielone na dwa tokeny,



Tokenizacja

- token to oczywiście “żeton”, jednak na nasze potrzeby możemy w przybliżeniu określić to jako słowo,
- tokenizacja może być trywialna i sprowadzić się do zwykłego podziału na słowa na podstawie spacji je rozdzielających,
- jednak np. słowo "humanistycznospołeczny" może zostać rozdzielone na dwa tokeny,
- podobnie niemiecki "im"= "in dem",



Tokenizacja

- token to oczywiście “żeton”, jednak na nasze potrzeby możemy w przybliżeniu określić to jako słowo,
- tokenizacja może być trywialna i sprowadzić się do zwykłego podziału na słowa na podstawie spacji je rozdzielających,
- jednak np. słowo "humanistycznospołeczny" może zostać rozdzielone na dwa tokeny,
- podobnie niemiecki "im"= "in dem",
- czy hiszpańskie "damelo"(daj mi to)



Tokenizacja

- token to oczywiście “żeton”, jednak na nasze potrzeby możemy w przybliżeniu określić to jako słowo,
- tokenizacja może być trywialna i sprowadzić się do zwykłego podziału na słowa na podstawie spacji je rozdzielających,
- jednak np. słowo "humanistycznospołeczny" może zostać rozdzielone na dwa tokeny,
- podobnie niemiecki "im"= "in dem",
- czy hiszpańskie "damelo"(daj mi to)

Normalizacja

Czasami wraz z poprzednią operacją przeprowadza się jednocześnie tzw. **normalizację**, czyli np. zamienia słowa “trzy”, “trzeci” itd. na 3. Wszystko to po to, żeby uprościć sobie życie i zmniejszyć rozmiary modelu.

Rozpoznawanie bytów – named entity recognition(NER)

- wyszukiwanie **bytów nazwanych** (*named entities*) w tekście,

Rozpoznawanie bytów – named entity recognition(NER)

- wyszukiwanie **bytów nazwanych** (*named entities*) w tekście,
- mogą to być nazwy własne, ale również wyrażenia dotyczące czasu, jednostek, walut etc,

Rozpoznawanie bytów – named entity recognition(NER)

- wyszukiwanie **bytów nazwanych** (*named entities*) w tekście,
- mogą to być nazwy własne, ale również wyrażenia dotyczące czasu, jednostek, walut etc,
- kilka znanych metod statystycznych wykorzystywanych do tego zadania to łańcuchy Markowa, ukryte łańcuchy Markowa or Conditional Random Fields (CRF),

Rozpoznawanie bytów – named entity recognition(NER)

- wyszukiwanie **bytów nazwanych** (*named entities*) w tekście,
- mogą to być nazwy własne, ale również wyrażenia dotyczące czasu, jednostek, walut etc,
- kilka znanych metod statystycznych wykorzystywanych do tego zadania to łańcuchy Markowa, ukryte łańcuchy Markowa or Conditional Random Fields (CRF),

"There was nothing about this storm that was as expected," said **Jeff Masters**, a meteorologist and founder of **Weather Underground**. "**Irma** could have been so much worse. If it had traveled 20 miles north of the coast of **Cuba**, you'd have been looking at a (Category) 5 instead of a (Category) 3."

Person

Organization

Location

Event Registry NER:

<https://eventregistry.org/documentation?tab=ner>

Albert Einstein (14 March 1879 – 18 April 1955) was a German-born theoretical physicist[5] who developed the theory of relativity, one of the two pillars of modern physics (alongside quantum mechanics)[3][6]:274 His work is also known for its influence on the philosophy of science.[7][8] He is best known to the general public for his *mass-energy* equivalence formula $E = mc^2$, which has been dubbed "the world's most famous equation".[9] He received the 1921 Nobel Prize in Physics "for his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect"[10] a pivotal step in the development of quantum theory.

Near the beginning of his career, Einstein thought that Newtonian mechanics was no longer enough to reconcile the laws of classical mechanics with the laws of the electromagnetic field. This led him to develop his special theory of relativity during his time at the Swiss Patent Office in Bern (1902–1909), Switzerland. However, he realized that the principle of relativity could also be extended to gravitational fields, and he published a paper on general relativity in 1916 with his theory of gravitation. He continued to deal with problems of statistical mechanics and quantum theory, which led to his explanations of particle theory and the motion of molecules. He also investigated the thermal properties of light which laid the foundation of the photon theory of light. In 1917, he applied the general theory of relativity to model the structure of the universe.[11][12]

Load sample document in:

LOAD

COMPUTE NAMED ENTITIES

Visual display

JSON data

Identified named entities

Text	Start position	End position	Type
Albert Einstein	0	15	PERSON
14 March 1879	16	29	DATE
18 April 1955	32	45	DATE
German-born	53	64	NATIONALITY
theoretical physicist	65	86	TITLE
5	87	88	NUMBER
one	130	133	NUMBER
two	141	144	NUMBER
3	202	203	NUMBER

Stemowanie (stemming)

- jest to, ogólnie mówiąc, obcięcie wszelkiego rodzaju przedrostków i przyrostków, mające na celu dotarcie do nieodmiennego “rdzenia” reprezentującego wyraz,

Stemowanie (stemming)

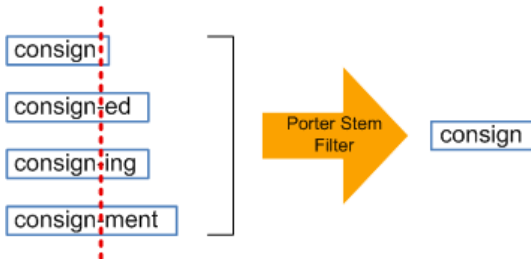
- jest to, ogólnie mówiąc, obcięcie wszelkiego rodzaju przedrostków i przyrostków, mające na celu dotarcie do nieodmiennego “rdzenia” reprezentującego wyraz,
- istotne jest to, że ów rdzeń niekoniecznie musi być poprawnym słowem,

Stemowanie (stemming)

- jest to, ogólnie mówiąc, obcięcie wszelkiego rodzaju przedrostków i przyrostków, mające na celu dotarcie do nieodmiennego “rdzenia” reprezentującego wyraz,
- istotne jest to, że ów rdzeń niekoniecznie musi być poprawnym słowem,
- stemmer działa najczęściej za pomocą pewnego zestawu reguł, np. w jęz. angielskim usuwania końcówek *ed*, *ing*, *ly*,

Stemowanie (stemming)

- jest to, ogólnie mówiąc obcięcie wszelkiego rodzaju przedrostków i przyrostków, mające na celu dotarcie do nieodmiennego “rdzenia” reprezentującego wyraz,
- istotne jest to, że ów rdzeń niekoniecznie musi być poprawnym słowem,
- stemmer działa najczęściej za pomocą pewnego zestawu reguł, np. w jęz. angielskim usuwania końcówek *ed*, *ing*, *ly*,
- bardzo znany stemmer Portera (lata 80-te)



Stemmer Portera:

<https://tartarus.org/martin/PorterStemmer/def.txt>

In the rules below, examples of their application, successful or otherwise, are given on the right in lower case. The algorithm now follows:

Step 1a

SSes -> SS	caresses -> caress
IES -> I	ponies -> poni
	ties -> ti
SS -> SS	caress -> caress
S ->	cats -> cat

Step 1b

(m>0) EED -> EE	feed -> feed
(*v*) ED ->	agreed -> agree
	plastered -> plaster
	bled -> bled
(*v*) ING ->	motoring -> motor
	sing -> sing

If the second or third of the rules in Step 1b is successful, the following is done:

AT -> ATE	conflat(ed) -> conflate
BL -> BLE	troubl(ed) -> trouble
IZ -> IZE	siz(ed) -> size
(*d and not (*L or *S or *Z))	
-> single letter	
	hopp(ing) -> hop
	tann(ed) -> tan
	fall(ing) -> fall
	hiss(ing) -> hiss
	fizz(ed) -> fizz
	fail(ing) -> fail
	fil(ing) -> file
(m=1 and *o) -> E	

Lematyzacja

- to sprowadzenie słowa do jego podstawowej postaci,

Lematyzacja

- to sprowadzenie słowa do jego podstawowej postaci,
- W przypadku czasownika będzie do bezokolicznik, w przypadku rzeczownika – mianownik liczby pojedynczej etc,

Lematyzacja

- to sprowadzenie słowa do jego podstawowej postaci,
- W przypadku czasownika będzie to bezokolicznik, w przypadku rzeczownika – mianownik liczby pojedynczej etc,
- do wykonania tego zadania potrzebny jest słownik lub rozbudowany zestaw reguł fleksyjnych dla danego języka.

Dan Jurafsky



Lemmatization

- Reduce inflections or variant forms to base form
 - *am, are, is* → *be*
 - *car, cars, car's, cars'* → *car*
- *the boy's cars are different colors* → *the boy car be different color*
- Lemmatization: have to find correct dictionary headword form

Part of speech tagging (POS)

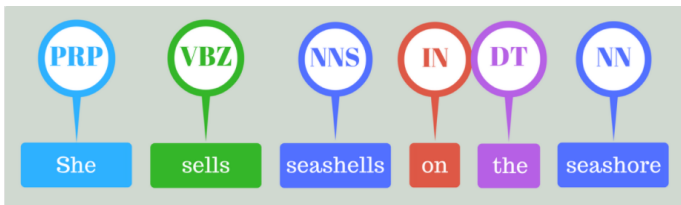
- musimy się dowiedzieć z jakimi częściami mowy mamy do czynienia, najlepiej również w jakich odmianach,

Part of speech tagging (POS)

- musimy się dowiedzieć z jakimi częściami mowy mamy do czynienia, najlepiej również w jakich odmianach,
- z reguły informację tę zwraca lematyzator,

Part of speech tagging (POS)

- musimy się dowiedzieć z jakimi częściami mowy mamy do czynienia, najlepiej również w jakich odmianach,
- z reguły informację tę zwraca lematyzator,
- analiza może być oparta o słownik, ale nie zawsze będzie to działać – tu też mogą się przydać metody statystyczne,



Morfeusz R:

<http://sgjp.pl/morfeusz/morfeusz.html>


Analizator morfologiczny Morfeusz

Podstawowe pojęcia

Słowem nazywamy ciąg znaków w tekście w języku naturalnym zwykle wydzielony odstępami lub znakami interpunkcyjnymi. I słowo zinterpretowane — przypisane do konkretnego leksemu i opisane co do jego funkcji gramatycznej.

Analiza morfologiczna polega na określeniu dla danego słowa wszystkich form wszystkich leksatów, których może ono być wyrazem. W językoznawstwie termin analiza morfologiczna odnosi się raczej do rozkładania wyrazów na na elementarne składniki morfol. wydaje się, że to ten pierwszy termin utarł się w środowisku językoznawstwa komputerowego.

Ujednoznacznianiem morfologicznym nazywamy określanie na podstawie kontekstu, jaką formę realizuje dane wystąpienie wyrazu.

Następujące po sobie analizy i ujednoznacznianie morfologiczne nazywa się żargonowo **tagowaniem**.

Celem **hasłowania (lematyzacji)** jest wskazanie dla każdego słowa tekstowego opisującej je jednostki słownika morfologicznego informacji o formach — do lematów.

Przybliżone hasłowanie polegające na odcięciu ze słów części zmieniającej się przy odmianie bywa nazywane **stemowaniem**. Me niezadowolające. W kontekście Morfeusza mówimy więc o prawdziwym hasłowaniu.

Operacją odwrotną do analizy morfologicznej jest **synteza morfologiczna** — utworzenie wykładnika formy odmiany danej przez słowo.

Program Morfeusz

Program Morfeusz wykonuje analizę morfologiczną dla języka polskiego. W obecnej wersji nie zawiera modułu zgadującego nieznaną formę wyrazu.

Oto przykład wyników działania programu dla tekstu „Mam próbkę analizy morfologicznej.”:

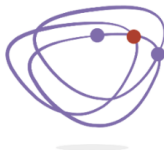
0	1	Mam	mama	subst-pl-gen:f
			mamié	impt-sg-sec:imperf
			mieć	fin-sg-pri:imperf
1	2	próbkę	próbka	substat-ag-acc:f
2	3	analizy	analiza	subst-sg-gen:f
				subst-pl-nom-acc-voef
3	4	morfologicznej	morfologiczny	adj-sg-gen-dat-loc:f:pos
4	5	.	.	interp

Lematyzacja w R:

<https://wilkowski.org/notka/1587>

Prosta i szybka lematyzacja w R z wykorzystaniem usług Clarin

Opublikowano: 17.01.2018



CENTRUM TECHNOLOGII
JEZYKOWYCH **CLARIN-PL**

Konsorcjum Clarin udostępnia interfejs programistyczny, pozwalający na zdalne przetwarzanie dokumentów tekstowych. Można swobodnie wykorzystywać te usługi do pracy z tekstem w R. Oto prosty sposób na sprowadzanie do wspólnej, podstawowej postaci wyrazów z wektora (czyli – w dużym skrócie – wskazanie lematów, podstawowych form hasłowych). Dzięki temu jesteśmy w stanie np. przygotować dobrą statystykę wyrazów czy wygenerować poprawną chmurę słów kluczowych dla tekstu.

Poniższy przykład wykorzystuje tager WCRFT. Tager to narzędzie, które dzieli tekst na wyrazy i dla każdego rozpoznaje określone właściwości gramatyczne.

```
function(t,u) {
```

Pasrowanie

- to analiza składniowa,

Parsowanie

- to analiza składniowa,
- jednak w praktyce już na tym etapie często wydobywa się także informacje semantyczną, czyli znaczenie wypowiedzenia,

Parsowanie

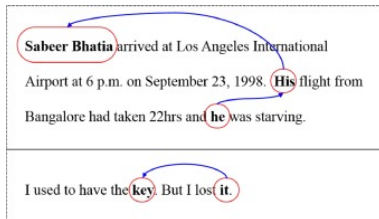
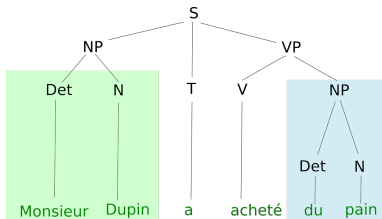
- to analiza składniowa,
- jednak w praktyce już na tym etapie często wydobywa się także informacje semantyczną, czyli znaczenie wypowiedzenia,
- tu pojawiają się drzewka składniowe,

Parsowanie

- to analiza składniowa,
- jednak w praktyce już na tym etapie często wydobywa się także informacje semantyczną, czyli znaczenie wypowiedzenia,
- tu pojawiają się drzewka składniowe,
- często używamy sieci semantycznych,

Parsowanie

- to analiza składniowa,
- jednak w praktyce już na tym etapie często wydobywa się także informację semantyczną, czyli znaczenie wypowiedzenia,
- tu pojawiają się drzewka składniowe,
- często używamy sieci semantycznych,
- problem z odwołaniami (*anaphora resolution*)



n-gram

- **n-gram** jest formalnie modelem językowym,

n-gram

- **n-gram** jest formalnie modelem językowym,
- opiera się na statystykach występowania słów (ale także fonemów czy kodów genetycznych),

n-gram

- **n-gram** jest formalnie modelem językowym,
- opiera się na statystykach występowania słów (ale także fonemów czy kodów genetycznych),
- służy do wyznaczania przewidywania kolejnego elementu sekwencji,

n-gram

- **n-gram** jest formalnie modelem językowym,
- opiera się na statystykach występowania słów (ale także fonemów czy kodów genetycznych),
- służy do wyznaczania przewidywania kolejnego elementu sekwencji,

