Przetwarzanie języka naturalnego/01

2024-03-07

Krzysztof Misztal

Spis treści

Spis treści

Spis treści

Tematyka wykładu

Analizujemy język angielski

Tematyka jest w trakcie tworzenia, ale mniej więcej wygląda tak:

1. Statystyczne modelowanie języka naturalnego (wykłady 1-10)
2. Deep learning approach (wykłady 11-14)

Spis treści

Zasady zaliczenia przedmiotu

1. **50%: (K1, K2)** – kolokwia, na kartkach (stacjonarnie) lub online (zdalnie), terminy: ... oraz ...
2. **30%: (P)** – projekt, 1-2 osobowy (szczegóły wkrótce)
3. **30%: (A)** Aktywność na ćwiczeniach
4. **(E)** Egzamin: test + zadania otwarte:

   if *K1* $<$ *10% lub K2* $<$ *10% lub K1+K2* $<$ *30%* **then**
   | egzamin z maksymalną oceną 3.0
   else
   | przepisana ocena z ćwiczeń
   end
5. Ocena końcowa:

$$W = K1 + K2 + P + A$$

- 5.0 if $90\% < W$
- 4.5 if $80\% < W \leqslant 90\%$
- 4.0 if $70\% < W \leqslant 80\%$
- 3.5 if $60\% < W \leqslant 70\%$
- 3.0 if $50\% \leqslant W \leqslant 60\%$

Zasady zaliczenia przedmiotu

Ocena z wykładu

Ostateczna ocena będzie wystawiona na podstawie oceny z ćwiczeń i ewentualnego egzaminu ustnego.

Osoby z oceną z ćwiczeń 4.5 lub wyższą dostaną możliwość przepisania oceny z ćwiczeń jako ocenę końcową bez przychodzenia na egzamin.

Laboratoria

Zaliczenie lab jest konieczne do zaliczenia przedmiotu.

Dopuszczalne są dwie nieobecności. Projekt musi dotyczyć języka polskiego + Clarin-PL

Spis treści

Laboratoria

Oprogramowanie

- Python `http://python.org`
- NLTK `http://nltk.org`
- scikit-learn `http://scikit-learn.org`

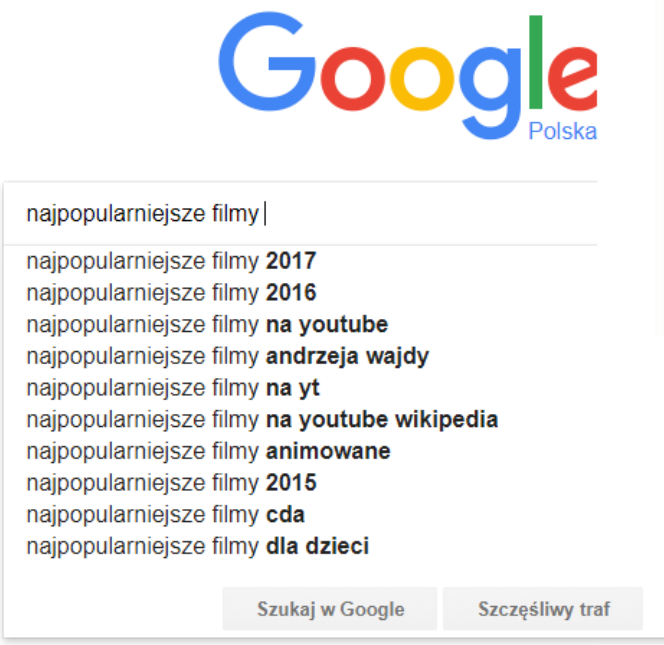
Na laboratoriach nie uczymy się programować w Pythonie!

Spis treści

Spis treści

Czy kiedykolwiek ukorzystali Państwo z NLP?

Czy kiedykolwiek ukorzystali Państwo z NLP?



Co to jest NLP?

Przetwarzanie języka naturalnego (ang. Natural Language Processing - NLP) jest dziedziną z pogranicza

- informatyki
- sztucznej inteligencji
- lingwistyki

Co to jest NLP?

Przetwarzanie języka naturalnego (ang. Natural Language Processing - NLP) jest dziedziną z pogranicza

- informatyki
- sztucznej inteligencji
- lingwistyki

Celem NLP jest nauczenie komputerów "rozumienia" języka naturalnego, tak aby były w stanie, np.

- wykonywać zadania, jak np. umawiać spotkania, kupować przedmioty;
- odpowiadać na pytania.

Co to jest NLP?

Przetwarzanie języka naturalnego (ang. Natural Language Processing - NLP) jest dziedziną z pogranicza

- informatyki
- sztucznej inteligencji
- lingwistyki

Celem NLP jest nauczenie komputerów "rozumienia" języka naturalnego, tak aby były w stanie, np.

- wykonywać zadania, jak np. umawiać spotkania, kupować przedmioty;
- odpowiadać na pytania.

Całkowite zrozumienie i reprezentacja znaczenia języka naturalnego jest niezwykle trudnym zadaniem.

Dlaczego przetwarzanie języka naturalnego?

- Język naturalny?

Dlaczego przetwarzanie języka naturalnego?

- Język naturalny?
- Język naturalny to język powstały na drodze rozwoju historycznego, zróżnicowany geograficznie i społecznie, przeciwstawiający się z jednej strony językom sztucznym (jak np. esperanto), z drugiej zaś językom formalnym i językom programowania. wyrażeń oraz tym, że podlega ciągłym zmianom.
- Język naturalny to np. angielski, japoński, w przeciwieństwie do sztucznych języków takich jak C++, Java itd.

Definicja

- Przetwarzanie języka naturalnego (ang. natural language processing, NLP) – interdyscyplinarna dziedzina, łącząca zagadnienia sztucznej inteligencji i językoznawstwa, zajmująca się automatyzacją analizy, rozumienia, tłumaczenia i generowania języka naturalnego przez komputer.
- Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts/speech at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

Definicja

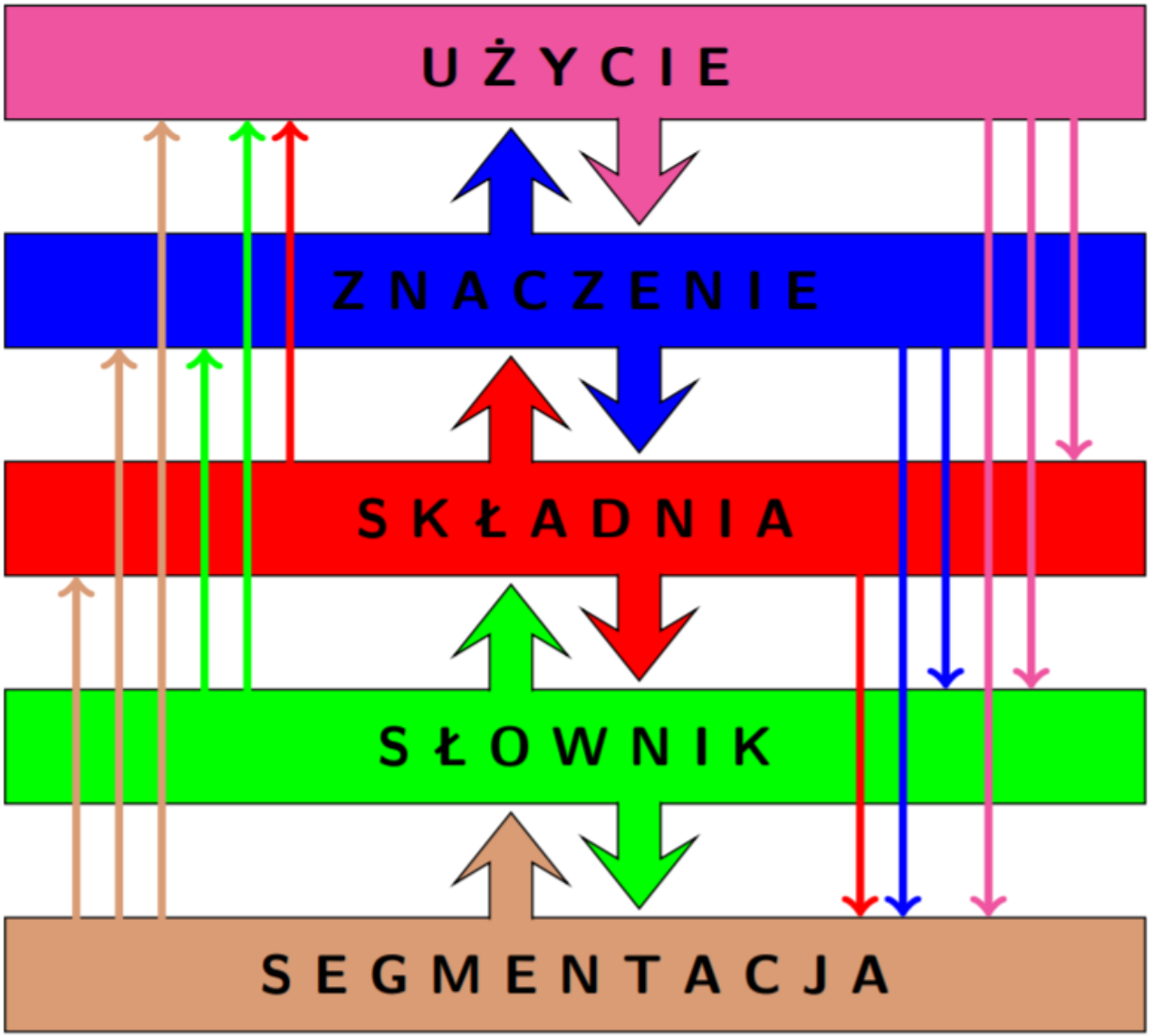
- Liczenie znaków w tekście nie jest przetwarzaniem języka naturalnego.
- Liczenie zdań – tak.

Spis treści

Podstawowe terminy w NLP

- syntaktyka – zajmuje się szykiem, związkami i stosunkami zachodzącymi pomiędzy wyrazami w zdaniu
- semantyka – zajmuje się zależnościami pomiędzy elementami języka, a ich odpowiednikami ze świata rzeczywistego, czyli znaczeniem tych elementów
- fleksja – zajmuje się budową form wyrazowych i ich odmianą
- składnia – zajmuje się regułami, według których wyrazy łączą się tworząc poprawne zdania
- gramatyka – zajmuje się opisem języka, w jej skład wchodzą fleksja oraz składnia
- wypowiedzenie – to komunikat językowy, podstawowa jednostka tekstu. Można powiedzieć, że to tekst rozpoczynający się od dużej litery, a kończący się kropką lub innym znakiem przestankowym.
- zdanie – to rodzaj wypowiedzenia, który zawiera podmiot i orzeczenie

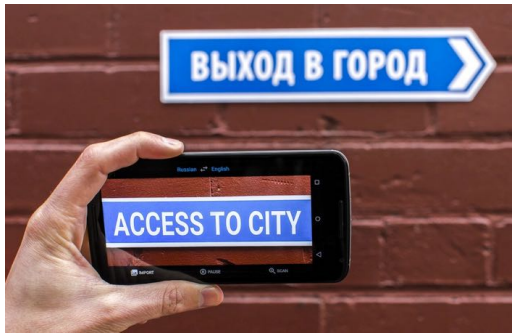
Poziomy w NLP



Spis treści

Zastosowania NLP

- tłumaczenia (https://www.youtube.com/watch?v=_GdSC1Z1Kzs)



Zastosowania NLP

- ekstrakcja informacji

Hi Jake, lets meet tomorrow at 10 for a lunch!

calendar entry
date: 4th October
time: 10:00
what: lunch

Zastosowania NLP

- analiza sentymentu

Zastosowania NLP

- analiza sentymentu

Zastosowania NLP – inne

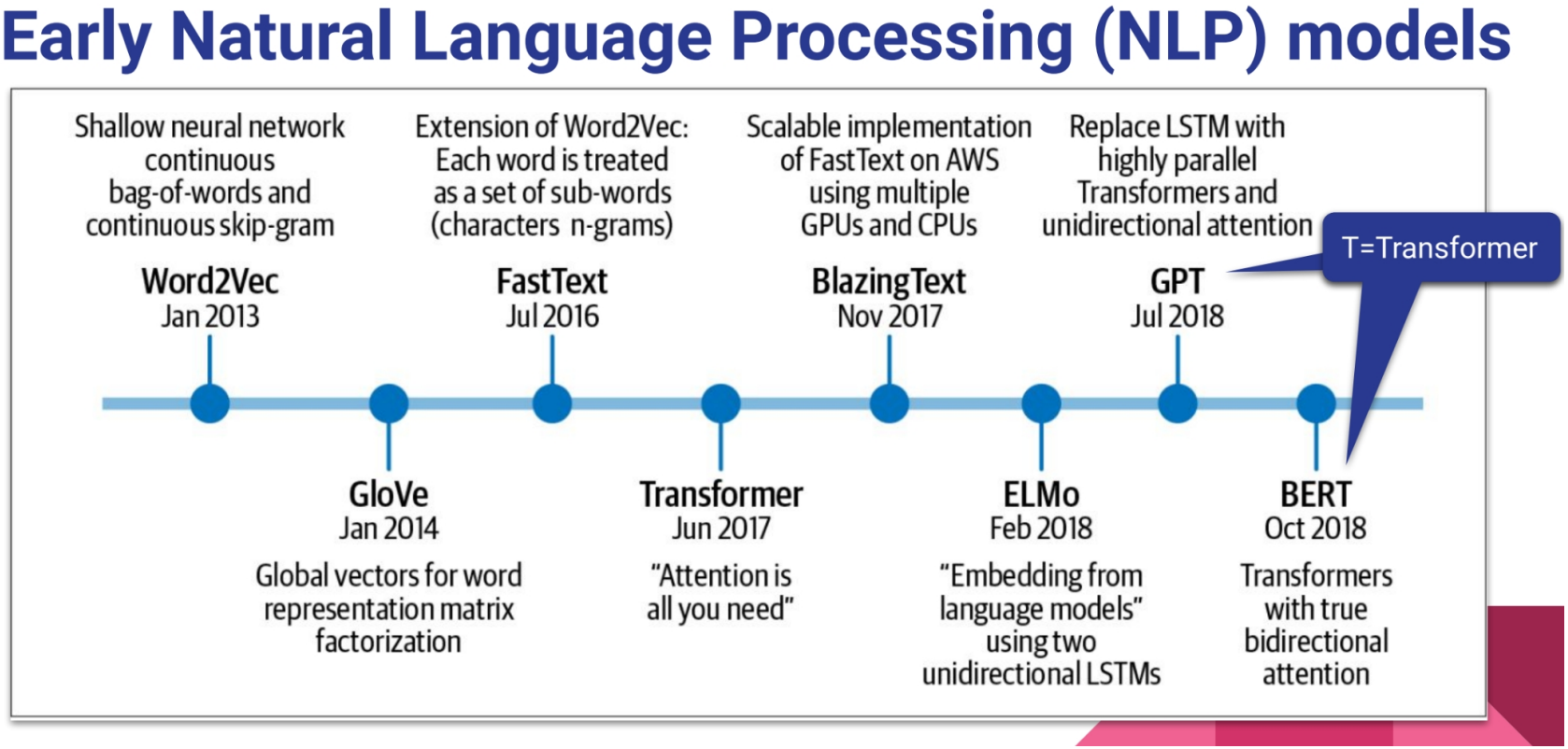
- sprawdzanie błędów (spell checking), wyszukiwania (keyword search), znajdowanie synonimów (finding synonyms);
- klasyfikacja testów: pozytywna/negatywna, tematyczna, itd.
- udzielanie odpowiedzi na pytania (Complex question answering)
- wykrywanie spamu
- rozpoznawanie części mowy
- parafrazowanie
- ..

Czemu to jest trudne?

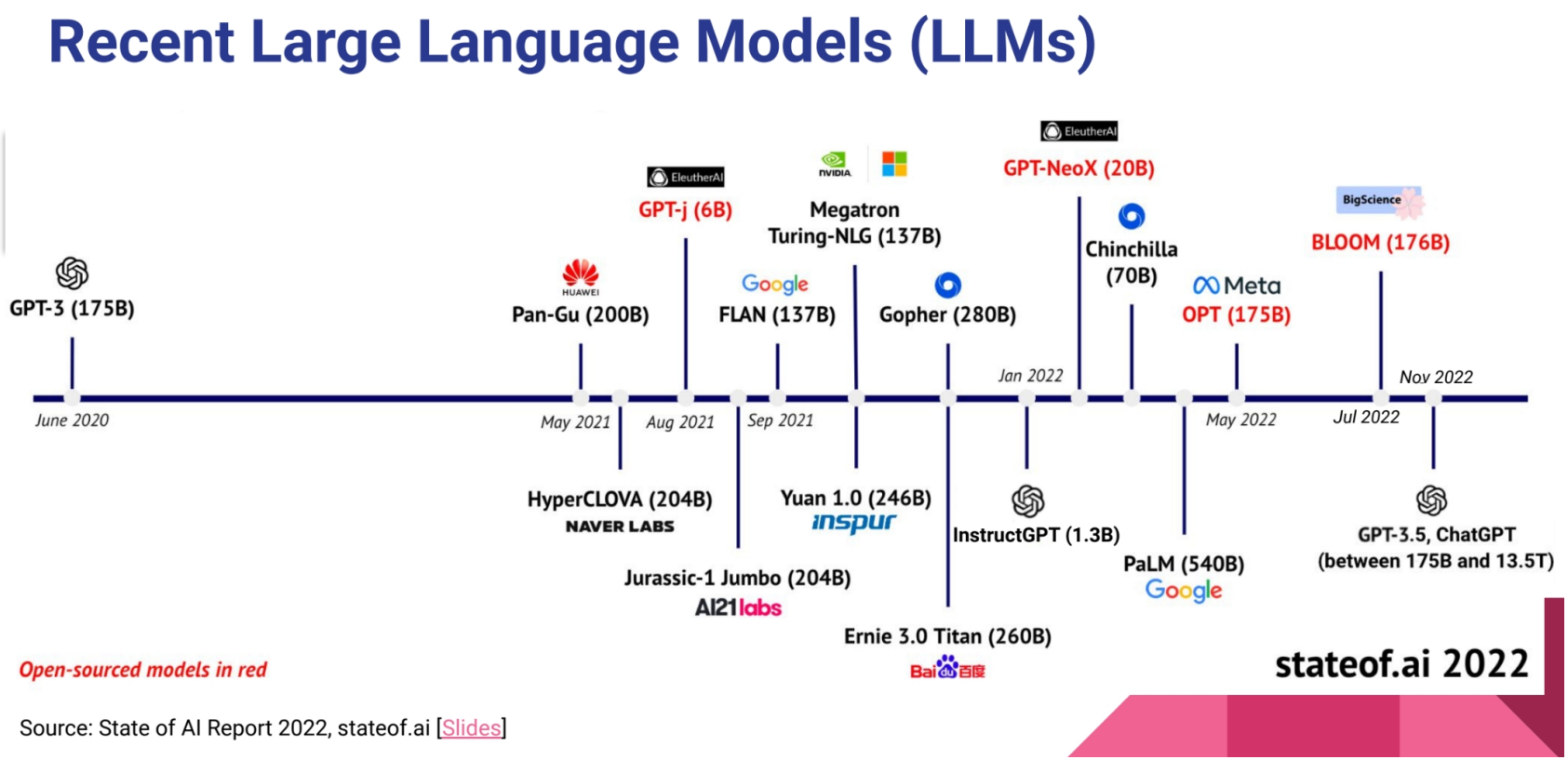
- Wymaga wiedzy o świecie i języku
- Pracujemy na olbrzymich zasobach danych
- Brak ścisłej struktury

Spis treści

Współczesność

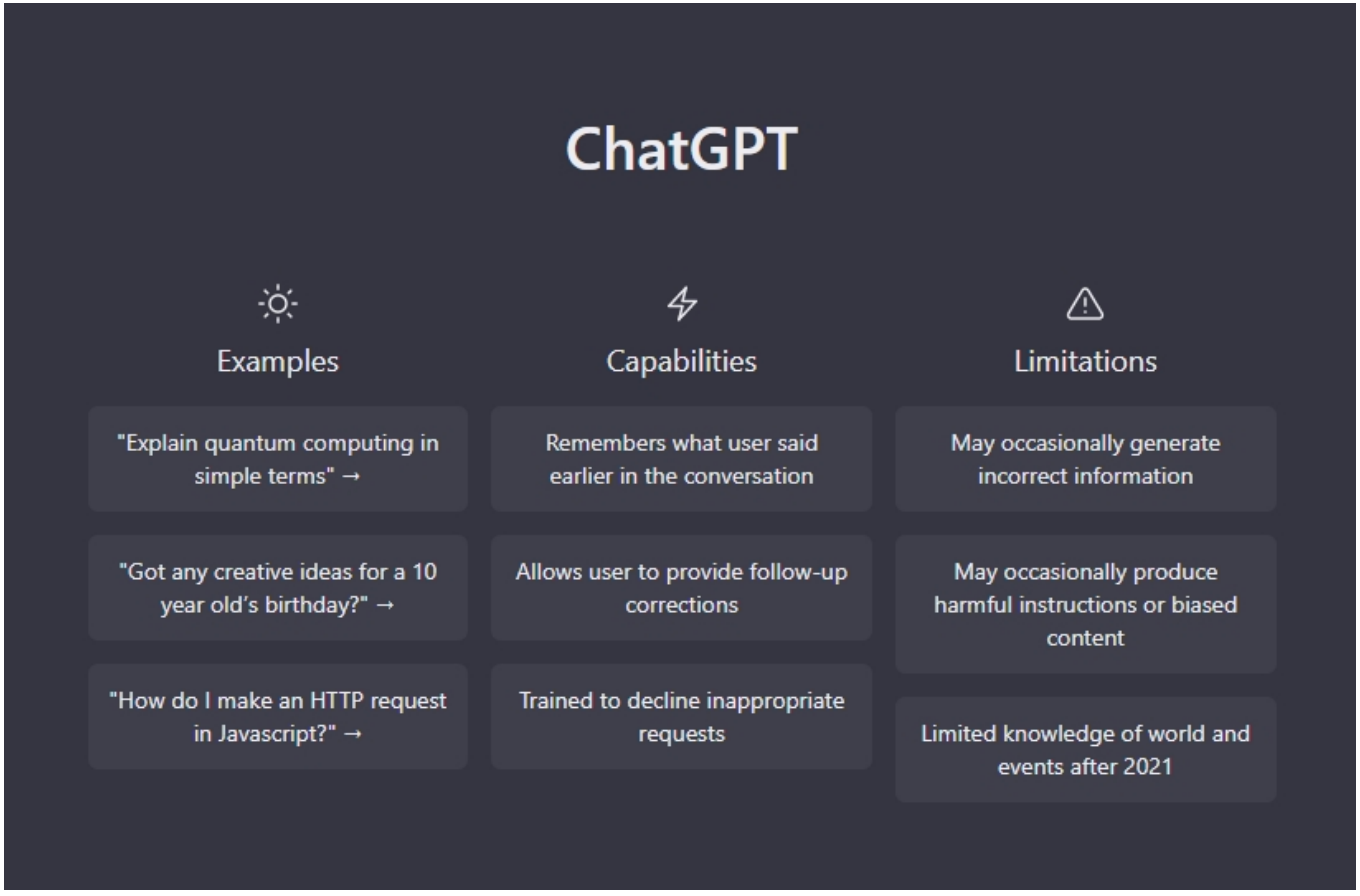


Współczesność



Source: State of AI Report 2022, stateof.ai [Slides]

Współczesność



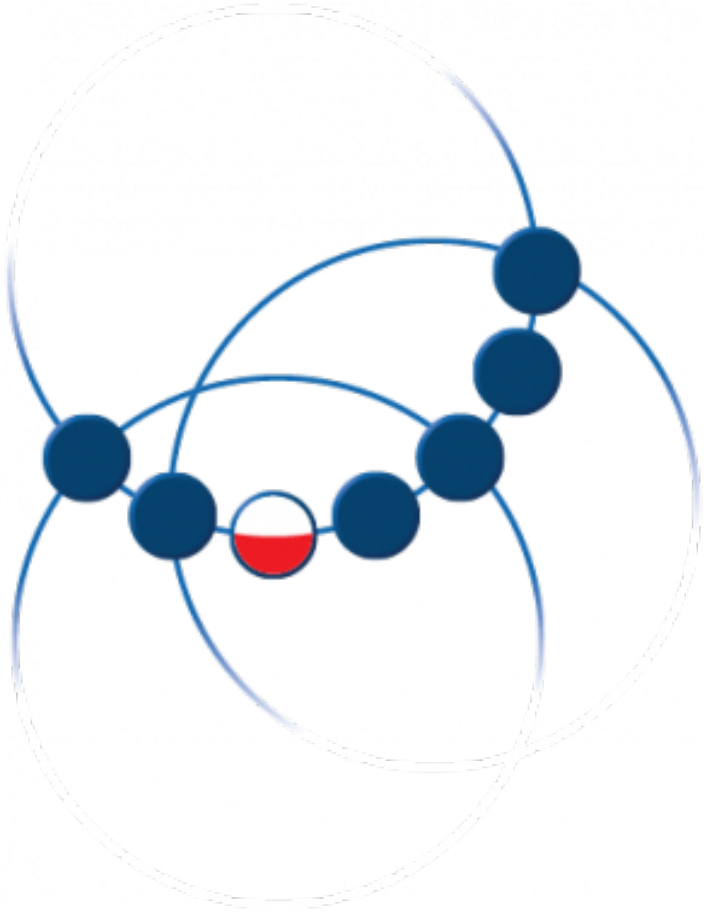
`https://chat.openai.com/chat`

Spis treści

Współczesność

CLARIN

Common Language Resources and Technology Infrastructure



https://clarin-pl.eu

chatGPT

- Czy możesz mi opowiedzieć tym jak rozwijało się przetwarzanie języka naturalnego?
- Czy przetwarzanie języka polskiego jest trudne?
- Czy możesz podać przykłady powyższych problemów?

Spis treści

Korpusy

„Korpus to dowolny zbiór tekstów, w którym czegoś szukamy. O korpusach w tym znaczeniu mówią najczęściej językoznawcy, ale także archiwiści, historycy i informatycy” – wydawnictwo PWN

- `http://korpus.pwn.pl` (7.5 mln słów dostępnych za darmo – ale bez możliwości pobrania)
- `http://korpus.pl` (korpus IPI PAN – 300 mln)

Korpusy

Korpus – zbiór tekstów reprezentatywnych dla języka, zapisany w formie elektronicznej, o ile to możliwe zawierający metadane

- niezbilansowany – niereprezentatywny dla języka, np. zawierający jedynie teksty o pewnej tematyce, albo też
- zbilansowany - reprezentatywny dla całego języka
- jednojęzykowy vs. wielojęzyczny (bitext)
- anotowany – zawierający metadane, w szczególności POS tags i/lub informacje o rozbiorze zdania

Korpus jest zwykle statyczny i jako taki jest „fotografią” języka w pewnej chwili – np. Brown corpus – język angielski z lat 60-tych

Korpusy

Pierwszy duży korpus – Brown Corpus: Kucera, Francis, lata 60-te XX w.

Wiele korpusów nie jest niestety dostępnych bezpłatnie (np. Penn TreeBank), większość repozytoriów bezpłatnych to zwykle czysty tekst, co najwyżej z podziałem na kategorie tematyczne (OTA, Project Gutenberg) i ew. pewnymi dodatkowymi metadanymi dotyczącymi całego pojedynczego dokumentu (Reuters 21578, Reuters RCV1)

Tworzenie korpusów – problemy

- **ilość danych** – im korpus większy, tym lepszy (miliony słów, setki MB czystego tekstu), dzisiaj to stosunkowo niewielki problem, lecz np. każda operacja sortowania listy słów podczas tworzenia Brown Corpus zajmowała 17 godzin na IBM 7070
- **formatowanie** – znaki niealfabetyczne, wielkość znaków
- **tokenizacja** – podział tekstu na: słowa i zdania
- **metadane** – wybór pomiędzy opisem ręcznym (często niemożliwym do wykonania) a automatycznym (często dającym nie najlepsze wyniki)

Problemy – 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

Problemy – podział na słowa

- **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*

Problemy – podział na słowa

- **apostrof** – szczególne problemy w języku angielskim, apostrof może mieć znaczenie gramatyczne
  - *I'll* $\rightarrow$ *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 $\rightarrow$ I + ' + ll
  - *niestety apostrofy mogą się także pojawić jako znaczniki cytowania trzeba odróżnić boys' od she said 'hello boys'*
- więcej przykładów
  ang. it's a 'dog', dog's bone, dog's crazy, dogs' house
  fr. qu'est-ce que c'est, aujourd'hui, l'amour, je l'aime

Problemy – podział na słowa

- **myślniki** – zwykle trzy główne funkcje
  • *dzielenie słów przy formatowaniu – wystąpią, 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.*

Problemy – podział na słowa

- **myślniki** w języku polskim:
  • W 1900 r. trafił do Niemieckiej Południowo-Zachodniej Afryki.
  • Zakład Przemysłowo-Drzewny „Henryków”
  • Żydowskie Stowarzyszenie Kulturalno-Oświatowe Tarbut
  • SS-man Fuss aresztował Jankiela za sabotaż
  • Kazimierz Opel ukrył 6-osobową rodzinę Górskich
  • musieli oni nie tylko wykazać się znajomością programu 2-letniej
  • państwowej szkoły elementarnej. . .
  • Dochodząc w opowieści o PRL-u do takiego punktu,. . .
- czyli jedno czy dwa słowa???

Problemy – podział na słowa

- **homonimy** – słowa które mając tą samą formę typograficzną oznaczają różne leksemy np.
  - *zamek (do drzwi) – zamek (króla)*

Problemy – podział na słowa

- **białe znaki**

  • *nie zawsze są używane do podziału zdań na słowa np.:*
  język chiński – zupełny brak podziału na słowa;
  język niemiecki – niektóre rzeczowniki zapisywane bez spacji
  Lebensversicherungsgesellschaftsangesteller – 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

Podział na zdania (sentence boundary detection)

- **Gdzie kończy się zdanie? na kropce?**
  - . . . nie ma prawdy innej, jak cała prawda; to też wszelkie zatajanie jest popełnianiem kłamstwa.
  - Czy to nasza wina, że mamy takich władców? Myśmy ich sobie nie wybierali! W tysiącletniej afgańskiej historii żaden z władców nie został wyniesiony na tron z woli poddanych.
  - W 1885 r. znalazł się Stanach Zjednoczonych, następnie w Wielkiej Brytanii; w 1900 r. w Johannesburgu i Kapsztadzie. W 1900 r. trafił do Niemieckiej Południowo-Zachodniej Afryki. Zmarł prawdopodobnie w Brukseli w 1912 r.
- **a co z...?**
  - Skróty (także inicjały), liczby porządkowe (zapisane cyframi)?
  - Czy kropka należy do skrótu, czy stanowi odrębny znak?
  - Co ze skrótami na końcu zdania?

Podział na zdania (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”.

  ale też

  „You remind me”, she remarked, „of your mother.”

Podział na zdania (sentence boundary detection)

- **podejście lepsze – heurystyka**:
  1. wstępne podziały zdań po . ? !.
  2. uwzględnienie cudzysłowów występujących po powyższych zdaniach
  3. skasowanie podziału zdania jeśli:
     - jeśli jest poprzedzony znanym skrótem, po którym występuje zwykle nazwa własna – np. Prof. lub vs.
     - jeśli jest poprzedzony znanym skrótem po którym nie występuje słowo rozpoczęte wielką literą
     - jeśli podział zdania wynikał z wystąpienia '!' lub '?' oraz następuje po nim mała litera

Podział na zdania (sentence boundary detection)

- **jeszcze lepsze podejścia**:
  - 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 POS słów przed i po końcach zdań
  - ...

Spis treści

Zawartość korpusów – power laws

https://www.youtube.com/watch?v=fCn8zs912OE

Zipf's law

Zipf's law (1949)

Częstotliwość f występowania słowa w w korpusie jest proporcjonalna do jego pozycji r w liście wszystkich słów z korpusu posortowanych według częstości występowania w nim, czyli

$$f \propto \frac{1}{r}$$

Zipf's law

Zipf's law (1949)

Częstotliwość f występowania słowa w w korpusie jest proporcjonalna do jego pozycji r w liście wszystkich słów z korpusu posortowanych według częstości występowania w nim, czyli

$$f \propto \frac{1}{r}$$

czyli

$$f \cdot r = \text{constant}$$

Na przykład, gdy w danym tekście 100. wyraz został użyty 314 razy tzn. ($r \cdot f = 31400$), z kolei 200. wyraz został użyty 158 razy ($r \cdot f = 31600$), to odchylenie od normy między setnym a dwusetnym wyrazem – zgodnie z prawem Zipfa – wynosi około 0,008%.

Spis treści

Heaps's law

"Rozmiar słownika rośnie wraz z rozmiarem korpusu"

Heaps's law

"Rozmiar słownika rośnie wraz z rozmiarem korpusu"

Heaps's law

Ilość różnych słów V (słów w słowniku) rośnie wraz ze wzrostem rozmiaru korpusu n według formuły

$$V(n) = Kn^{\beta}$$

dla języka angielskiego, zwykle: $K \in [10, 100]$ oraz $\beta \in [0.4, 0.6]$

Spis treści

Wyrażenia regularne

- Formalny sposób opisu tekstu
- Jak możemy w prosty sposób sformułować zapytanie pasujące do następujących słów:
  - dog
  - Dog
  - DOG
  - dogs
  - dog's

Wyrażenia regularne

Wyliczenia w nawiasach kwadratowych []

$[dD]og$	dog; Dog
$d[io]g$	dog; dig
[0123456789]	pojedyncza cyfra

Zakresy w nawiasach kwaratowych [AZ]

$[A-Z]$	wielka litera
$[a-z]$	mała litera
$[0-9]$	cyfra
$[A-Za-z0-9]$	znak alfanumeryczy
$[a-c]$	‘a’, ‘b’, ‘c’
$[-a-c]$	‘a’, ‘b’, ‘c’ lub ‘-’

Wyrażenia regularne

Negacja

[^A-Z]	nie wielka litera	O**h**
[^a-z]	nie mała litera	**O**h
[^ab]	nie ‘a’ ani nie ‘b’	**c**ab
[^a^]	nie ‘a’ ani nie ^	a^**b**
[a ^b]	‘a’,^a potem ‘b’	x=**a^b**

Wyrażenia regularne

Alternatywa

dog|puppy	dog; puppy
a|b|c|d	a; b; c; d; =[abcd]
[dD]og|[pP]uppy	dog; Dog; puppy; Puppy

Wyrażenia regularne

Podstawowe operatory

dogs?	poprzednie dop. nie wymagane	dog; dogs
oo*h	poprzednie dop. dowolnie wiele razy	oh; ooh; oo...oh
o+h	poprzednie dop. co najmniej raz	oh; ooh; oo...oh
d.g	dowolny znak	dig; dog; dXg
o[ao]*h		oh; ooh; oah; ooaoaoah; ...

Wyrażenia regularne

Kotwice

^[A-Z]	początek tekstu	**D**og
[a-z]$	koniec tekstu	Do**g**

Wyrażenia regularne

Kotwice

.$	ostatni znak	Do**g**
\.$	kropka na końcu	Dog**.**
\[[0-9]\]	cyfra w nawiasach	**[8]**

Wyrażenia regularne

Grupowanie i ograniczenia liczności

d{3}	ddd
d{1,5}	d; dd; ddd; dddd; ddddd
d{3,}	ddd; dddd; ddddd, ...
a (dog){1,2}	a dog ; a dog dog
a (dog)+	a dog ; a dog dog ; a dog dog dog ... dog ;
(dog|cat) and \1	dog and dog; cat and cat

Wyrażenia regularne

Dopasowanie zachłanne

o.*h	**oohooh**
.+h	**oohooh**

Dopasowanie leniwe

o.*?h	**ooh**ooh
.+?h	**ooh**ooh

Przykład

Chcemy dopasować przedimki ‘a, an, the’

- a|an|he

  a,an,the,**A**,...
- [aA]n?|[tT]he

  a,an,the,A,An,The,**A**rbuz,...
- (^|[^a-zA-Z0-9])([aA]n?|[tT]he)([^a-zA-Z0-9]|$)

Przykład – analiza

Komplikowanie wyrażenia było spowodowane dwoma rodzajami błędów:

- Nie dopasowywaliśmy słów, które chcieliśmy (a,an,the,A)
- Dopasowywaliśmy słowa, których nie chcieliśmy (Arbuz)

Przykład – analiza

Komplikowanie wyrażenia było spowodowane dwoma rodzajami błędów:

- Nie dopasowywaliśmy słów, które chcieliśmy (a,an,the,A)
  False negatives (FN)
- Dopasowywaliśmy słowa, których nie chcieliśmy (Arbuz)
  False positives (FP)

Budowanie modeli

W NLP zajmujemy się głównie **budową** pewnych modeli.
Naszym celem jest zazwyczaj:

- Maksymalizacja **accuracy** bądź **precision** (minimalizacja FP)
- Maksymalizacja **coverage** bądź **recall** (minimalizacja FN)

Te dwie wartosci sa **antagonizmami**.

Po co nam wyrażenia regularne?

Mimo swej prostoty i zdawałoby się toporności, wyrażenia regularne są **podstawa NLP**, niezależnie od złożoności systemu, na którymś etapie niemal każdy ich używa.

Dziękuję za uwagę.