Przetwarzanie języka naturalnego/02

2024.03.14

Krzysztof Misztal misztal.edu.pl

Spis treści

Spis treści

Statystyczne NLP

- Jak przewidzieć wystąpienie kolejnego słowa w sekwencji słów?

Statystyczne NLP

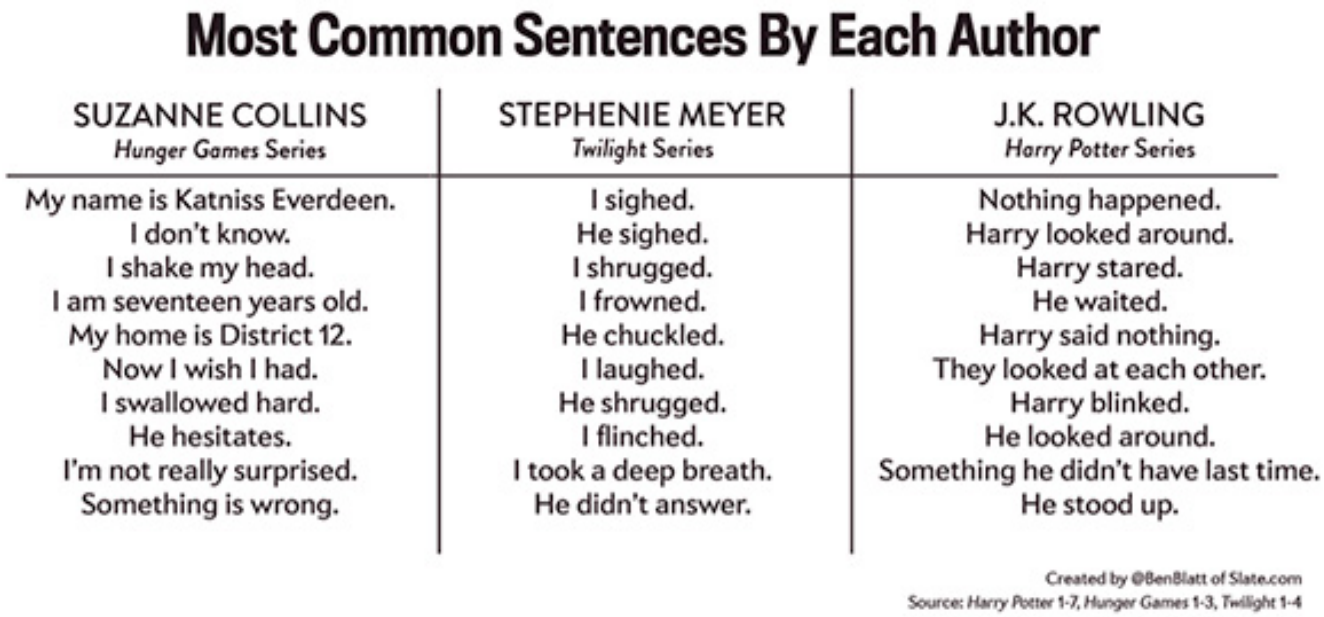
- Jak przewidzieć wystąpienie kolejnego słowa w sekwencji słów?
- Do tego potrzebny jest model generacji słów w języku, określający prawdopodobieństwa wystąpienia pewnych słów pod warunkiem wystąpienia słów poprzedzających.

Najczęściej występujące słowa w języku angielskim

https://www.youtube.com/watch?v=7XQRduB6oTM

Najczęściej występujące słowa w języku angielskim

Najczęściej występujące słowa dla autorów



Most Common Sentences By Each Author

SUZANNE COLLINS *Hunger Games* Series	STEPHENIE MEYER *Twilight* Series	J.K. ROWLING *Harry Potter* Series
My name is Katniss Everdeen.	I sighed.	Nothing happened.
I don't know.	He sighed.	Harry looked around.
I shake my head.	I shrugged.	Harry stared.
I am seventeen years old.	I frowned.	He waited.
My home is District 12.	He chuckled.	Harry said nothing.
Now I wish I had.	I laughed.	They looked at each other.
I swallowed hard.	He shrugged.	Harry blinked.
He hesitates.	I flinched.	He looked around.
I'm not really surprised.	I took a deep breath.	Something he didn't have last time.
Something is wrong.	He didn't answer.	He stood up.

Created by @BenBlatt of Slate.com
Source: *Harry Potter* 1-7, *Hunger Games* 1-3, *Twilight* 1-4

Źródło: *https://intothewonder.wordpress.com/2013/11/24/textual-analysis-of-hunger-games-twilight-and-harry-potter/*

Statystyczne NLP

- Podejście statystyczne zakłada wykorzystanie metod wnioskowania statystycznego do analizy języka naturalnego
- Wnioskowanie statystyczne – analiza pewnych danych eksperymentalnych (wyników doświadczenia, odpowiedzi na pytania ankietera itp.), generowanych zgodnie z pewnym nieznanym rozkładem prawdopodobieństwa, w celu określenia cech tego rozkładu

Spis treści

Spis treści

Zdarzenie losowe

- Zdarzenie losowe to zdarzenie, którego zaistnienie zależy od przypadku...
- ... a formalniej to mierzalny podzbiór zbioru zdarzeń elementarnych danego doświadczenia losowego.

Zdarzenie losowe

- Zdarzenie losowe to zdarzenie, którego zaistnienie zależy od przypadku...
- ... a formalniej to mierzalny podzbiór zbioru zdarzeń elementarnych danego doświadczenia losowego.
- Zdarzenie elementarne - to pojedynczy wynik eksperymentu losowego.
- Przestrzeń zdarzeń elementarnych Ω - to zbiór możliwych wyników eksperymentu losowego.
- Zdarzenie przeciwne do danego zdarzenia - to zdarzenia będące dopełnieniem danego zdarzenia do zbioru Ω.

Zdarzenie losowe – przykłady

- obserwacja rzutu monetą:
  przestrzeń zdarzeń: $\Omega = \{O, R\}$
  zdarzenie losowe: wyrzucenie orła
- badanie błędów ortograficzne w danym języku
  przestrzeń zdarzeń: $\Omega = Z^*$, gdzie Z – alfabet, Z^* – ciąg znaków nad tym alfabetem
  zdarzenie losowe: wystąpienie błędu ortograficznego – **czy to jest proste?**

Spis treści

Prawdopodobieństwo

Jakie jest prawdopodobieństwo zdarzenia?

- Powtarzamy eksperyment t razy, zliczając liczbę c wystąpień zdarzenia A
- Powyższe serie powtarzamy wielokrotnie
- Wartość c_i/t_i zbliża się do pewnej (nieznanej a priori) stałej wartości
- Ta stała wartość to prawdopodobieństwo zdarzenia A (definicja częstościowa, von Misesa),

$$P(A) = \lim_{t_i \to \infty} \frac{c_i}{t_i}$$

- Oczywiście w praktyce nie da się przeprowadzić nieskończonej liczby doświadczeń, zatem wartość prawdopodobieństwa możemy jedynie estymować z c_1/t_1

Prawdopodobieństwo – definicja aksjomatyczna Kołmogorowa

Niech $\mathcal{F}$ będzie σ-ciałem podzbiorów zbioru Ω, tzn. spełnia następujące warunki:

- $\mathcal{F}$ jest nie pusta,
- jeżeli $A \in \mathcal{F}$, to $A' = \Omega \setminus A \in \mathcal{F}$,
- jeżeli $A_i \in \mathcal{F}$, to $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Podzbiór A zbioru Ω nazywamy **zdarzeniem losowym**, gdy $A \in \mathcal{F}$. Zbiór pusty nazywamy **zdarzeniem niemożliwym**, zbiór Ω **zdarzeniem pewnym**, natomiast A' **zdarzeniem przeciwnym** do zdarzenia A.

Prawdopodobieństwo – definicja aksjomatyczna Kołmogorowa

Definicja

Prawdopodobieństwem nazywamy funkcję P spełniającą następujące aksjomaty

- $P : \mathcal{F} \to \mathbb{R}_+$;
- $P(\Omega) = 1$;
- *jeżeli* $A_i \in \mathcal{F}$, $i \in \mathbb{N}$ *są parami rozłączne, to* $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Prawdopodobieństwo własności

Twierdzenie

Niech $A, B \in \mathcal{F}$. Miara probabilistyczna P ma następujące własności:

- *jeśli $A \subset B$, to $P(A) \leqslant P(B)$;*
- $P(A) \leqslant 1$;
- $P(A') = 1 - P(A)$;
- $P(\emptyset) = 0$;
- $P(A \setminus B) = P(A) - P(A \cap B)$;
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Spis treści

Prawdopodobieństwo warunkowe

W wielu przypadkach, informacja o zajściu zdarzenia B ma pewien wpływ na wartość obliczonego prawdopodobieństwa zdarzenia A. Zdarzenie polegające na zajściu zdarzenia A przy założeniu, że zaszło zdarzenie B, oznaczamy symbolem $A|B$, prawdopodobieństwo tego zdarzenia $P(A|B)$ nazywamy prawdopodobieństwem warunkowym.

Definicja

Prawdopodobieństwem zajścia zdarzenia A pod warunkiem, że zdarzenie B zajdzie, nazywamy liczbę

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

gdzie $A, B \subset \Omega$, $P(B) > 0$.

Prawdopodobieństwo warunkowe

Zatem

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A).$$

co daje w ogólności

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|\cap_{i=1}^{n-1} A_i).$$

Prawdopodobieństwo warunkowe

Zatem

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A).$$

co daje w ogólności

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|\cap_{i=1}^{n-1} A_i).$$

Oczywiście zdarzenia A i B są niezależne, wtw

$$P(A \cap B) = P(A)P(B).$$

Prawdopodobieństwo całkowite

Jeżeli zdarzenia $B_1, B_2, \ldots, B_n$ są parami rozłączne oraz mają prawdopodobieństwa dodatnie, które sumują się do jedynki, to dla dowolnego zdarzenia A zachodzi wzór:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n)$$

Spis treści

Wzór Bayesa

Wzór Bayesa pozwala nam odwrócić stosunek zależności pomiędzy zdarzeniami – czyli obliczyć $P(B|A)$, gdy znane jest $P(A|B)$

Twierdzenie (Twierdzenie Bayesa)

Jeżeli zdarzenia $B_1, B_2, \ldots, B_n$ wykluczają się parami i mają prawdopodobieństwa dodatnie, to dla każdego zdarzenia A zawartego w sumie zdarzeń $B_1 \cup B_2 \cup \ldots \cup B_n$:

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n)}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

Wzór Bayesa - przykład

- Algorytm wykrywa "parastic gap" – czyli "pasożytnicza luka", to przerwa, która nie może istnieć (czyli musi być wypełniona) *Which book did she review __ without reading __?*
- G: w zdaniu jest parasitic gap, T: algorytm wykrył pg
- Algorytm myli się "w obie strony" tzn. $P(T|G) = 0.95$ oraz $P(T|\bar{G}) = 0.05$
- $P(G) = 0.00001$
- Algorytm wykrył pg, czy pg jest rzeczywiście w zdaniu?

$$P(G|T) = \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|\bar{G})P(\bar{G})}$$

$$= \frac{0.95 \cdot 0.00001}{0.95 \cdot 0.00001 + 0.05 \cdot 0.99999} \approx 0.002$$

Spis treści

Wstępne przetwarzanie tekstu - cele:

- **Wyodrębnianie zwykłego tekstu** – dane tekstowe mogą pochodzić z wielu różnych źródeł (internet, pliki PDF, dokumenty tekstowe, systemy rozpoznawania mowy, skany książek itp.), celem jest wyodrębnienie zwykłego tekstu, który jest wolny od jakichkolwiek znaczników lub konstrukcji źródłowych, które nie są istotne dla zadania.
- **Zmniejszenie złożoności** – niektóre cechy naszego języka, takie jak wielkie litery, interpunkcja i popularne słowa, takie jak a, of i the, często pomagają w nadaniu struktury, ale nie dodają zbyt wiele znaczenia. Czasami najlepiej je usunąć, jeśli pomoże to zmniejszyć złożoność procedur, które chcemy zastosować później.

Wstępne przetwarzanie tekstu - etapy:

- Czyszczenie w celu usunięcia nieistotnych elementów, takich jak znaczniki HTML (Removal of HTML tags), emotikony (Removal of emoticons), urle (Removal of URLs), itp.
- Normalizowanie poprzez konwersję na wszystkie małe litery (Lower casing) i usunięcie znaków interpunkcyjnych (Removal of Punctuations)
- Dzielenie tekstu na słowa lub tokeny
- Usuwanie zbyt popularnych słów (Removal of Frequent words), rzadkich słów (Removal of Rare words) oraz tzn. stopwordsów (Removal of Stopwords)
- Rozpoznawanie różnych części mowy i nazwanych jednostek
- Konwersja słów na ich formy słownikowe przy użyciu Stemming i Lemmatization
- Korekcja błędów językowych
- ...

Wstępne przetwarzanie tekstu - etapy:

https://ichi.pro/pl/post/262250151052646
https://github.com/A2Amir/NLP-and-Pipelines

Spis treści

Tokenizacja

Intuicja

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

Tokeny są w jakimś sensie elementarnymi "wyrazami".

Typ wyrazu

Intuicja

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

Typy i tokeny

- Typ - element słownika (unikalne wyrazy)
- Token - konkretne wystąpienie (każdy wyraz pojawiający się w tekście/korpusie)

Typy i tokeny

- Typ - element słownika (unikalne wyrazy)
- Token - konkretne wystąpienie (każdy wyraz pojawiający się w tekście/korpusie)

A good wine is a wine that you like.

- Tokenów: 9
- Typów: 7

Tokenizacja – problemy

Problem: podzielić tekst na tokeny (“słowa” elementarne w jakimś sensie)

Tokenizacja – wiele konwencji

- *I like trains* $\longrightarrow$ I; like; trains
- *Poland's capital* $\longrightarrow$ Poland; 's; capital

Tokenizacja – wiele konwencji

- *I like trains* $\longrightarrow$ I; like; trains
- *Poland's capital* $\longrightarrow$ Poland; 's; capital/Poland; '; s; capital/Poland's; capital
- *state-of-the-art* $\longrightarrow$ state; of; the; art

Tokenizacja – wiele konwencji

- *I like trains* $\longrightarrow$ I; like; trains
- *Poland's capital* $\longrightarrow$ Poland; 's; capital/Poland; '; s; capital/Poland's; capital
- *state-of-the-art* $\longrightarrow$ state; of; the; art/state-of-the-art
- *U.S.A.* $\longrightarrow$ U.S.A.

Tokenizacja – wiele konwencji

- *I like trains* $\longrightarrow$ I; like; trains
- *Poland's capital* $\longrightarrow$ Poland; 's; capital/Poland; '; s; capital/Poland's; capital
- *state-of-the-art* $\longrightarrow$ state; of; the; art/state-of-the-art
- *U.S.A.* $\longrightarrow$ U.S.A./U.; S.; A.

Tokenizacja – wiele konwencji

- *Kot-Dziwak* $\rightarrow$ Kot; Dziwak / Kot-Dziwak?
- *Zielona Góra* $\rightarrow$ Zielona; Góra/Zielona Góra?

Tokenizacja - typowe podejscie

Zakładając prace z językiem **angielskim**

- Dla dużych danych - **prosta tokenizacja** + ew. postprocessing statystyczny
- Dla małych danych - analiza alternatyw i wybór najbardziej prawdopodobnej wersji

Tokenizacja - SNLP group

- Jedna z najlepszych grup NLP na swiecie
- Bardzo wydajny tokenizator oparty o gigantyczny (11 500 linii kodu w Javie) automat skonczony
- Konwencja za Penn TreeBank

`http://nlp.stanford.edu/software/tokenizer.shtml`

Spis treści

Lematyzacja

Zadanie, polegające na ustaleniu, że dane wyrazy mają ten sam lemat, mimo różnicy w formie i pogrupowaniu ich form tak aby były identyfikowane przez lemat lub element słownika.

Lematyzacja

Zadanie, polegające na ustaleniu, że dane wyrazy mają ten sam lemat, mimo różnicy w formie i pogrupowaniu ich form tak aby były identyfikowane przez lemat lub element słownika.

jestem, jesteś, są $\longrightarrow$ być
kotów, koty $\longrightarrow$ kot

Morfologia

Morfem

Najmniejsza jednostka gramatyczna, część wyrazu (może stanowić samodzielny wyraz).

Temat wyrazu: część wyrazu, która jest nośnikiem znaczenia wyrazu (np.**kot**-)
Prefiks, interfiks, sufiks: niesamodzielne morfemy (np. **-ek** w kotek)

Stemming

Prostsza wersja lematyzacji, gdzie z wyrazu usuwana jest końcówka fleksyjna, pozostawiając tylko temat wyrazu

Stemming

Prostsza wersja lematyzacji, gdzie z wyrazu usuwana jest końcówka fleksyjna, pozostawiając tylko temat wyrazu

koty, kotów, kotek $\rightarrow$ kot

Stemming/lematyzacja

- **stemming** - jest to proces polegający na wydobyciu z wybranego wyrazu tzw. rdzenia, a więc tej jego części, która jest odporna na odmiany przez przyimki, rodzaje itp.
  politician, politicians, policy $\longrightarrow$ politics
  policeman, policemen $\longrightarrow$ police
- **lematyzacja** - pojęcie to jest bardzo podobne do powyższego, a oznacza sprowadzenie grupy wyrazów stanowiących odmianę danego zwrotu do wspólnej postaci, umożliwiającej traktowanie ich wszystkich jako te samo słowo
  am, are, is $\longrightarrow$ be

Stemming

Dwa z najbardziej znanych to algorytm stemmingu Portera (1979 r.) oraz algorytm Lancaster (1990 r.).
`http://text-processing.com/demo/stem/`

Algorytm Portera

- Ogólna idea: **iteracyjne usuwanie nadmiarowych sufiksów**. Algorytm składa się z 5 głównych kroków:
  - Depluralyzacja oraz proste koncówki (usuwanie -es, -ed, -ing etc.)
  - Redukcja podwójnych sufiksów (“ational” $\rightarrow$ “ate”, “tional” $\rightarrow$ “tion”, etc.)
  - Usuwanie form przysłówkowych, bezokolicznikowych i im podobnych (“ness” $\rightarrow$ “ ”, “alize” $\rightarrow$ “al”, “icate” $\rightarrow$ “ic”)
  - Usuwanie “ant”, “ence” etc.
  - Usuwanie koncówki “e” oraz redukcja podwójnych spółgłosek (“ll” $\rightarrow$ “l”)

`http://tartarus.org/~martin/PorterStemmer/python.txt`

Algorytm Portera

Do you really think it is weakness that yields to temptation? I tell you that there are terrible temptations which it requires strength, strength and courage to yield to Oscar Wilde

Do you really think it is weakn**ess** that yield*s* to temptation? I tell you that there ar**e** terrib**le** temptat**ions** which it requir**es** strength, strength and courag**e** to yield to Oscar Wild**e**

Do you realli think it is weak that yield to temptat I tell you that there ar terribl temptat which it requir strength strength and courag to yield to Oscar Wild

Algorytm Lancaster

Bardzo szybka, agresywna metoda stemmingu, oparta o iteratywne aplikowanie pojedynczych zasad redukcji sufiksu.

Algorytm Lancaster

Do you really think it is weakness that yields to temptation? I tell you that there are terrible temptations which it requires strength, strength and courage to yield to *Oscar Wilde*

do you real think it is weak that yield to tempt i tel you that ther ar terr tempt which it requir strength strength and cour to yield to osc wild

Wordnet lemmatizer

Bardzo prosty algorytm używający bardzo złożonej bazy danych.

- Jeśli słowo nie należy do wordnetu, zwróć to słowo
- Zwróć formę podstawowa słowa

Wordnet lemmatizer

Do you really think it is weakness that yields to temptation? I tell you that there are terrible temptations which it requires strength, strength and courage to yield to *Oscar Wilde*

Do you really think it is weakness that yield to temptation ? I tell you that there are terrible temptation which it requires strength , strength and courage to yield to Oscar Wilde

Do własnego przejrzenia

- Porównanie `https://www.machinelearningplus.com/nlp/lemmatization-examples-python/`
- Przykłady `https://www.datacamp.com/community/tutorials/stemming-lemmatization-python`
- Morfeusz `http://morfeusz.sgjp.pl/doc/about/`
- CLARIN-PL `http://clarin-pl.eu/en/uslugi/`

Dziękuję za uwagę.