Text mining - odkrywanie wiedzy z tekstowych zbiorów danych

Wykład 3

Eksploracja danych testowych

Plan

- Adaptacja wybranych metod eksploracji danych do analizy danych tekstowych
- Przykłady
  - Wykrywanie synonimów
  - Wykrywanie homonimów
  - Wykrywanie związków nietaksonomicznych

Eksploracja danych tekstowych

Eksploracja danych tekstowych

=

Metody eksploracji danych

+

Klasyczne NLP

Zastosowanie reguł asocjacyjnych do analizy tekstu

Oznaczmy zbiór różnych wyrazów/ terminów (ang. terms) przez $T = \{t_1, t_2, ..., t_m\}$.
Zbiór *T*- jest pewnym słownikiem

Zbiór wyrazów X to dowolny podzbiór terminów z T

Zdanie jest zbiorem wyrazów, wyodrębnionym ze składniowo zdefiniowanej części tekstu (akapit, wyrażenie gramatyczne) za pomocą specjalnej procedury.

Zastosowanie reguł asocjacyjnych do analizy tekstu

Data mining	Text mining
Elementy (ang. items)	*Wyrazy (ang. terms)*
Zbiór elementów (ang. itemsets)	*Zbiór wyrazów (ang. termsets)*
Transakcja (ang. transaction)	*Zdanie (ang.sentence)*

Zbiory częste i reguły asocjacyjne w analizie tekstu (1)

Zbiory częste i wsparcie

Wsparcie zbioru wyrazów X, oznaczonego wsp(X), jest liczbą zdań w tekstach, które zawierają wszystkie terminy zawarte w X.

Zbiór wyrazów X jest częsty, jeśli wsp(X)> minWsp, gdzie minWsp jest zdefiniowaną przez użytkownika wartością progową.

Zbiory częste i reguły asocjacyjne w analizie tekstu (2)

Reguły asocjacyjne i zaufanie

- Reguła asocjacyjna jest wyrażeniem w postaci : $X \rightarrow Y$

  gdzie X,Y to zbiory wyrazów.

Reguła asocjacyjna wskazuje, że jeśli w zdaniu występują wszystkie wyrazy ze zbioru X, to prawdopodobnie występują również wszystkie wyrazy ze zbioru Y.

- Zaufanie reguły:

$$conf\,(X \rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)}$$

Wzorce sekwencyjne *versus* wzorce tekstowe

Wzorce sekwencyjne	Wzorce tekstowe
Element (ang. item)	**Słowo/ termin (ang. word/term)**
Transakcja – zbiór elementów	**Zdanie - lista słów**
Sekwencja transakcji – uporządkowany zbiór transakcji	**Dokument (paragraf) – lista zdań**
Baza danych – zbiór sekwencji transakcji	**Repozytorium tekstów – zbiór dokumentów/paragrafów**

Odkrywanie wzorców w danych tekstowych

Definicja wsparcia wzorców w przypadku analizy dokumentów tekstowych jest analogiczna do definicji wsparcia wzorców sekwencyjnych. Wsparcie wzorca tekstowego jest procentem jednostek tekstu zawierających ten wzorzec.

Sekwencja słów (ang. word-sequence)

- ***W* = {w_1, w_2, …, w_m}** – zbiór słów, w bardziej ogólnym ujęciu może to być zbiór terminów.
- ***Tw* –** taksonomia zdefiniowana dla zbioru W

- ***Sekwencja słów*** jest uporządkowaną listą słów. Dane słowo może wystąpić więcej niż raz w danej sekwencji słów. Sekwencja słów *ws* oznaczana jest jako < w_1, w_2, …, w_m>.

Odległość między słowami

Odległość między dwoma słowami *w1* i *w2* zawartymi w sekwencji *ws*, oznaczona jako $distance_{ws}$ (*w1*, *w2*), jest zdefiniowana jako liczba słów w rozważanej sekwencji znajdujących się pomiędzy tymi słowami.

Przykład

Dla sekwencji słów <*Jest to bardzo interesujący problem*>, odległość między słowem „*Jest*" a słowem „*interesujący*" jest równa 2, a odległość między „to" i „bardzo" jest równa 0.

Wzorzec tekstowy (ang. text-pattern)

- **Wzorzec tekstowy *tp*** jest uporządkowaną listą sekwencji słów i jest oznaczany jako $tp=<ws_1, ws_2, \ldots, ws_m>$.
- Baza danych wejściowych *tDB* składa się z dokumentów tekstowych.
- **Dokument *d*** jest listą zdań, $d=<snt_1,snt_2,\ldots,snt_n>$.
- **Zdanie *snt*** jest opisane przez identyfikator dokumentu, do którego należy, numer wskazujący pozycję na liście zdań oraz uporządkowaną listę słów obecnych w zdaniu;
  $snt = <w_1 ,w_2, \ldots, w_m>$.

Wzorzec tekstowy (2)

Odległość między dwoma słowami w zdaniu jest zdefiniowana identycznie jak w przypadku sekwencji słów.
W niektórych aplikacjach przydatne może być połączenie kilku zdań w jedno zdanie.

Dla dwóch zdań: *snt1*= $<w_1, w_2, \ldots, w_m>$ oraz *snt2*= $<v_1, v_2, \ldots, v_n>$ operacja łączenia, oznaczona jako *snt1*$\cup$ *snt2*, zdefiniowana jest jako:

$$snt3 = snt1 \cup snt2 = < w_1, w_2, \ldots, w_m, v_1, v_2, \ldots, v_n>.$$

Wzorzec tekstowy (3)

- **Dane słowo *w* odpowiada słowu *v***, co jest oznaczono jako $w \sim v$, jeśli te słowa są takie same lub *v* jest przodkiem *w* w danej taksonomii Tw.
- Zdanie $snt = < w_1, w_2, \ldots, w_m >$ zawiera sekwencję słów $ws =< v_1, v_2, \ldots, v_n >$, co jest oznaczone jako $ws \subseteq snt$, jeśli istnieją takie liczby naturalne $i_1 < i_2 < \ldots < i_n$ że: $v_1 \sim w_{i1}$, $v_2 \sim w_{i2}, \ldots, v_n \sim w_{in}$.
- Dokument $d = < snt_1, snt_2, \ldots, snt_m >$ zawiera wzorzec tekstowy $tp =< ws_1\ ws_2\ \ldots\ ws_n >$ (za pominięciem dodatkowych ograniczeń) jeśli istnieją takie liczby naturalne $i_1 < i_2 < \ldots < i_n$, że: $ws_1 \subseteq snt_{i1}$, $ws_2 \subseteq snt_{i2}, \ldots, ws_n \subseteq snt_{in}$.

Parametry algorytmu GSP w odniesieniu do danych tekstowych

- **Rozmiar okna** - wskazuje liczbę zdań, w których dana sekwencja słów powinna być obecna. Wartość 1 tego parametru oznacza, że sekwencja słów powinna być obecna w jednym zdaniu.
- **max-gap** – dla dwóch kolejnych sekwencji słów z danego wzorca tekstowego parametr ten wskazuje maksymalną liczbę zdań w dokumencie między dwoma zdaniami A i B, zawierającymi odpowiednio: pierwszą sekwencję słów (A) i drugą sekwencję słów (B), przy której dokument nadal zawiera wzorzec.

Parametry algorytmu GSP w odniesieniu do danych tekstowych (2)

- **min-gap** – dla dwóch zdań zaczerpniętych z danego dokumentu parametr ten wskazuje minimalną liczbę zdań, które powinny wystąpić w tym dokumencie pomiędzy tymi dwoma zdaniami, aby te dwa zdania można było uznać za dwa osobne zdania.

Parametr maxWord-gap (1)

- **Parametr *maxWord-gap*** definiuje maksymalną przerwę między dwoma kolejnymi słowami w sekwencji słów.

- **Zdanie zawiera sekwencję**, jeśli zawiera wszystkie słowa z sekwencji, a liczba słów między dowolnymi dwoma kolejnymi słowami w sekwencji w zdaniu jest niewiększa niż wartość parametru maxWord-gap.

- Formalnie, zdanie $snt = < w_1, w_2, \ldots, w_m >$ zawiera sekwencję słów $ws = <v_1, v_2, \ldots, v_n>$, jeśli istnieją takie liczby natrualne $i_1 < i_2 < \ldots < i_n$ że: $v_1 \sim w_{i1}$, $v_2 \sim w_{i2}, \ldots, v_n \sim w_{in}$ oraz $distance_{snt}(w_i, w_{i+1}) \leq maxWord\text{-}gap$, $1 \leq i \leq n\text{-}1$.

Parametr maxWord-gap (2)

Przykład.

zdanie *snt* = <*To jest bardzo interesujący problem*>,

maxWord-gap = 1,

word-sequence *ws* = <*jest problem*>.

Podane zdanie nie zawiera rozważanej sekwencji słów, gdyż:

$$distance_{snt}(jest, problem) = 2 > maxWord\text{-}gap,$$

Odkrywanie wzorców tekstowych

Dokument $d = <snt_1, snt_2, \ldots, snt_m>$ zawiera wzorzec tekstowy $p = < ws_1\ ws_2\ \ldots\ ws_n>$ jeśli istnieją takie liczby naturalne $i_1 < i_2 < \ldots < i_n$, że:

1. ws_i jest zawarty w $\cup_{k=l_i}^{u_i} snt_k$, $1 \leq i \leq n$.
2. $position(snt_{u_i}) - position(snt_{l_i}) \leq$ window-size, $1 \leq i \leq n$.
3. $position(snt_{l_i}) - position(snt_{u_{i-1}}) >$ min-gap, $2 \leq i \leq n$.
4. $position(snt_{u_i}) - position(snt_{l_{i-1}}) \leq$ max-gap, $2 \leq i \leq n$.

gdzie position(*snt*) jest pozycją zdania *snt* na liście zdań dokumentu, z którego pochodzi dane zdanie.

Wykrywanie synonimów

Podstawowe założenia

„The proportion of words common to the context of word A and to the context of word B is a function of the degree to which A and B are similar in meaning.”

H. Rubenstein, J. B. Goodenough, 1965

„If two words (descriptors) are synonymous then they very infrequently, or never co-occur as words in the same sentence, but in the separate occurrences they tend to have similar contexts”

P. A. W. Lewis , P. B. Baxendale , J. L. Bennett, 1967

Definicja kontekstu słowa

Kontekst słowa jest generowany w następujący sposób:

- wszystkie słowa występują razem w przetwarzanej jednostce tekstu (np. zdaniu)
- wszystkie rozważane słowa spełniają pewne ograniczenia (np. są określone części mowy, są częste itp.).

Kiedy dwa słowa są swoimi synonimami

- Na podstawie pewnej miary synonimii
- Miara synonimii między wyrazami A i B jest zwykle definiowana jako podobieństwo między kontekstami słów A i B.
- Zwykle brana jest jedna miara definiująca podobieństwo słów (redukcja semantyki)
- Nie ma uniwersalnej miary

Definicja kontekstu

Kontekst C(X) (częsty zbiór słów)

$$C(X) = \{Z \setminus X |\ X \subset Z \wedge sup(Z) > minSup\}$$

Wspólny kontekst dla X i Y - CC(X,Y) - część wspólna kontekstów

$$CC(X,Y) = context(X) \cap context(Y).$$

Przykłady

- *C(student) = {{Warsaw}, {university}, {Warsaw, University}, {exams}, {passed}, {passed, exams} }*
- *C(capital) = {{Warsaw}, {university}}*
- *C(rector) = {{Warsaw}, {university}, {Warsaw, University}}*
- *C(lecturer) = {{Warsaw}, {university}, {Warsaw, University}, {exams}, {performed}, {performed, exams}}*
- *C(professor) = {{Warsaw}, {university}, {Warsaw, University}, {exams}, {performed}, {performed, exams}}*

Miara podobieństwa kontekstów CSIM (ang. contexts similarity measure)

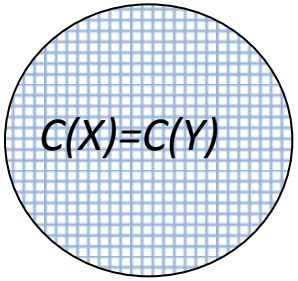
$$CSim\ (X,Y) = \frac{|CC(X,Y)|}{|context(X) \cup context(Y)|}$$

Kontekst może zawierać nie tylko pojedyncze słowa, ale również terminy wielowyrazowe.

Wspólny kontekst - przypadki

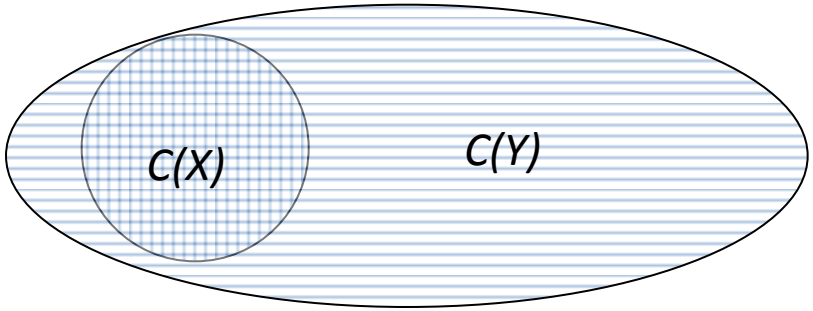
CSIM(X,Y)=1

(a)



C(X)⊂C(Y), CC(X,Y)=C(X)

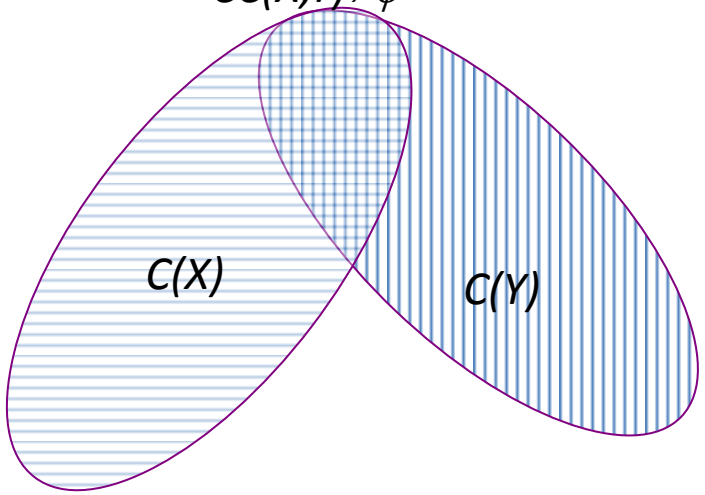
(b)



CSIM(X,Y) < 1

CC(X,Y)≠ϕ

(c)



Wspólny kontekst - przypadki

(a) Konteksty są takie same lub bardzo podobne, Przykłady: X i Y mogą być instancjami kategorii,
np. X = poniedziałek, Y = wtorek

(b) Relacja między X i Y jest typu szerszy-węższy, lub kategoria-instancja np. X = relacja, Y = hierarchia

(c) Słowa X i / lub Y mogą mieć więcej niż jedno znaczenie; mają bliskie znaczenie tylko w danym wspólnym kontekście

Miara podobieństwa kontekstów ASim (ang. association similarity measure)

Jeśli $CC(X,Y) = \emptyset$, lub dla każdego $Z \in CC(X,Y),\ |Z|=1$ to $ASim(X,Y) = 0$, w przeciwnym przypadku:

$$ASim(X,Y) = \frac{\sum_{Z \in CC(X,Y)} \left(SConf(X \to Z, Y \to Z) \right)}{\sum_{Z \in CC(X,Y)} |Z|}$$

gdzie

$$SConf(X \to Z, Y \to Z) = \min\left(\frac{conf(X \to Z)}{conf(Y \to Z)}, \frac{conf(Y \to Z)}{conf(X \to Z)} \right)$$

Odkrywanie synonimów – etapy procesu

1. Wstępne przetwarzanie korpusu
2. Konwertowanie korpusu tekstowego na bazę danych zawierającą transakcje
3. Znajdowanie częstych zbiorów słów
4. Znajdowanie nieczęstych par słów
5. Generowanie kontekstu słów
6. Wyliczenie miar synonimii dla nieczęstych par słów.

Wstępne przetwarzanie korpusu

- Podział dokumentów na zdania
- Usuwanie tzw. stopwords
- Oznaczenie słów częściami mowy
- Filtrowanie słów w oparciu o części mowy

Wykrywanie homonimów

Definicje

Homonim - wyraz o takim samym brzmieniu i zapisie, lecz odmiennym znaczeniu

Homonimia – relacja wyrażania różnych znaczeń za pomocą identycznych form językowych.

Odkrywanie homonimów

“It has been estimated that more than 73% of words used in common English texts are polysemous, i.e., more than 73% of words have more than one sense.”

Miller, G., Chadorow, M., Landes, S., Leacock, C., and Thomas, R. G., 1994,

Podstawowe założenia

Znaczenie(a) słowa A jest (są) dobrze zdefiniowane przez jego kontekst(y)

Jeśli słowo A ma dwa lub więcej charakterystycznych kontekstów, prawdopodobnie każdy kontekst określa inne, specyficzne znaczenie

Miara homonimii

- Miarę homonimii słowa A można zdefiniować jako rozróżnienie między różnymi kontekstami w jakich występuje słowo A.
- Jeśli są znalezione więcej niż 2 konteksty, które są wystarczająco odróżnialne, słowa A może być postrzegane jako homonim, a jego znaczenie definiowane jest przez te konteksty.

Odkrywanie homonimów – definicje (1)

MF(*x*) – wszystkie maksymalne częste zbiory słów zawierające słowo *x*.

Zbiór słów *X*, $x \notin X$, jest **kontekstem atomowym** słowa *x* jeśli $\{x\} \cup X \in MF(x)$.

Zbiór wszystkich atomowych kontekstów słowa *x* jest oznaczony przez jako *AC(x) :*

$$AC(x) = \{X \setminus \{x\} \mid X \cup \{x\} \in MF(x)\}.$$

Odkrywanie homonimów – definicje (2)

Dane są dwa zbiory słów $Y, Z \in AC(x)$:

Y różni się od Z przynajmniej jednym terminem i odwrotnie.

y – słowo w $Y \setminus Z$, z – słowo w $Z \setminus Y$

Załóżmy, że {xyz} jest zbiorem słów, którego wsparcie jest znacznie mniejsza niż wsparcie zbiorów Y i Z. Może to sugerować, że zbiory Y i Z prawdopodobnie reprezentują różne znaczenia słowa *x*.

{xyz} odgrywa rolę potencjalnego wyróżnika dla słowa *x* i jego atomowych kontekstów Y i Z.

Odkrywanie homonimów – definicje (3)

Zbiór wszystkich potencjalnych wyróżników:

$D(x, Y, Z) = \{\{xyz\} \mid y \in Y \setminus Z \wedge z \in Z \setminus Y\}, Y, Z \in AC(x)$

Właściwy wyróżnik dla słowa *x* – $pd(x, Y, Z)$:

$X \in D(x, Y, Z)$ and $relSup(x, X, Y, Z) \leq \delta$,

gdzieie

- $relSup(x, X, Y, Z) = sup(X) / min(sup(xY), sup(xZ))$,
- δ jest wartością podaną przez użytkownika.

Odkrywanie homonimów – definicje (4)

Zbiory *Y* oraz *Z* zwarte *AC*(*x*) **są rozróżnialne** (ang. *discriminable),* jeśli istnieje przynajmniej jest właściwy wyróżnik PD(*x*, *Y*, *Z*).

W przeciwnym przypadku Y i Z **są nierozróżnialne** wyróżniające.

Odkrywanie homonimów – definicje (5)

Kontekst wyróżniający znaczenie SDC(x, X) słowa x dla zbiorów słów X zawartych w $AC(x)$ jest sumą tych zbiorów słów w $AC(x)$, które są nierozróżniale względem słowa x:

$$\mathrm{SDC}(x, X) = \{Y \in AC(x) \mid \mathrm{PD}(x, X, Y) = \emptyset\}.$$

Odkrywanie homonimów – etapy procesu

1. Wstępne przetwarzanie korpusu
2. Zbudowanie słownika D składającego się z rzeczowników oraz nazw własnych
3. Konwertowanie korpusu tekstowego na bazę danych zawierającą transakcje
4. Znalezienie maksymalnych częstych zbiorów słów
5. Dla każdego słowa $x \in D$ znalezienie $AC(x)$
6. Dla każdego słowa obliczenie potencjalnych wyróżników (wraz z ich wsparciem), a następnie zidentyfikowanie właściwych wyróżników
7. Wyszukanie kontekstów wyróżniający znaczenie. Jeśli dla danego słowa *x* znaleziono więcej niż 1 taki kontekst to słowo *x* jest kandydatem na homonim

Wstępne przetwarzanie korpusu

- Podział dokumentów na paragrafy
- Oznaczenie słów częściami mowy
- Filtrowanie słów w oparciu o części mowy

Wykrywanie wyrażeń wielowyrazowych

Typy wyrażeń wielosłowowych

- **Rzeczowniki złożone** - reprezentują pojęcia, mają sztywną strukturę składniową; np.: „wyszukiwanie informacji", "Politechnika Warszawska", fizyka ciała stałego.
- **Idiomy** - wyrażenia, których znaczenia prawie nigdy nie można wyprowadzić ze znaczenia słów tworzących te wyrażenia (mała możliwość modyfikowania składni).
  Kolokacje - ta klasa składa się z powiązanych słów, tj. słów, które często współwystępują w tekście (struktura syntaktyczna nie jest sztywna). Do tej grupy należą również asocjacje.

Zastosowanie wzorców tekstowych

Pojedyncza, częsta sekwencja słów (wzorzec tekstowy) może być uważana za kandydata do różnych typów wyrażeń wielowyrazowych.

Rzeczowniki złożone (1)

Rzeczowniki złożone lub pojęcie wielowyrazowe - ciąg co najmniej dwóch słów użyty jako nazwa (wskaźnik) jednego pojęcia lub jednej rzeczy.

Przykłady:

Politechnika Warszawska
programowanie obiektowe
reguła asocjacyjna

Rzeczowniki złożone (2)

- Słowa budujące pojęcia wielowyrazowe lub wielosłowowe nazwy własne występują zwykle jeden po drugim w tekstach.
- W większości przypadków terminy wielowyrazowe składają się z: *rzeczowników*, *przyimków* i *przymiotników*.
- Wiele nazw własnych składa się tylko z rzeczowników i przyimka.

Rzeczowniki złożone (3)

Termin złożony (wielowyrazowy) - ciąg słów należących do następującej części mowy: przymiotnik, rzeczownik, przyimek, który występuje w danym korpusie co najmniej w *minSup* jednostkach tekstu.

Próg *minSup* jest podany przez użytkownika.

Jednostką tekstu może być dokument, paragraf lub zdanie.

Asocjacje – założenia (1)

Relacje asocjacyjne i obiekty w nich uczestniczące są reprezentowane w zdaniach ciągami słów spełniającymi następujący wzorzec: <rzeczownik, czasownik, rzeczownik>

Przykłady:

Agent sprzedaje wiele ***samochodów.***

Eksperymenty w obszarze TM ***zajmują*** dużo ***czasu***.

Asocjacje – założenia (2)

Interesujące są tylko te sekwencje, które wystąpiły w więcej niż podanej liczbie jednostek tekstu.

Jednostką tekstu może być dokument lub paragraf.

Częste wzorce

Pojedyncza częsta sekwencja słów może być uważana jako kandydat na wyrażenie wielowyrazowe

Wzorce gramatyczne (1)

Wzorzec gramatyczny jest parą [E,G],

Gdzie: **E** to zbiór par **: *e* = <V,P>,**

gdzie

V jest zbiorem lub ciągiem słów lub znaczników POS,

P jest trójką **(*mdt*, *minO*, *maxO*),**

gdzie

mdt **-** wskazuje, czy element *e* wzoru jest opcjonalny czy obowiązkowy**,**

minO **-** wskazuje minimalną liczbę pasujących słów do elementu *e* w rozważanej jednostce tekstu**,**

maxO **-** wskazuje maksymalną liczbę pasujących słów do elementu *e* w rozważanej jednostce tekstu (jeśli maxO = 0 oznacza "dowolną liczbę razy")

$G = \{max\_gap_{12}, max\_gap_{23}, \ldots, max\_gap_{n-1,n}\}$,

gdzie $max\_gap_{i,i+1}$ to maksymalna odległość między kolejnymi słowami w zdaniu, które są używane do wypełnienia wzorca gramatycznego na pozycjach *i* oraz *i*+1.

Wzorce gramatyczne (2)

Przykłady

- **gp1 = [{<{noun},(true,1,1)>, <{verb},(true,1,1)>, <{verb},(false,1,4)>, <{noun},(true,1,1)>},{3,0,3}]**
- **gp2 = [{<{noun},(true,1,1)>, <{preposition},(true,1,1)>, <{noun},(true,1,1)>},{0,1}]**
- **gp3 = [{<{noun},(true,1,1)>, <{verb},(false,1,1)>, <{adjective},(false,1,2)>, <{noun},(true,1,1)>},{3,1,3}]**

Wyszukiwanie fraz „rzeczownik czasownik czasownik rzeczownik” z opcjonalnym drugim czasownikiem z różną specyfikacją przerw między wyrazami.

Wzorce gramatyczne (3)

Kiedy zdanie spełnia wzór gramatyczny.

Zddanie $snt = <w_1, w_2, \ldots, w_m>$ wspiera wzorzec gramatycznyc $gp = [E,G]$,

$E = \{<POS^1,(true,1,1)>, <POS^2, (true,1,1)>,\ldots, <POS^{n,} (true,1,1)>\}$,

$G = \{max\_gap_{12}, max\_gap_{23},\ldots, max\_gap_{n-1,n}\}$,

jeśli istnieje taka sekwencja liczb naturalnych $(i_1, i_2, \ldots, i_n)$, $i_1<i_2<\ldots<i_n$, że $pos(w_{i,1}) \in POS^1$, $pos(w_{i,2}) \in POS^2,\ldots, pos(w_{i,n}) \in POS^n$,

gdzie $pos(w)$ oznacza część mowy w kontekście zdania snt, do którego należy słowo

oraz $distance_{snt}(w_{i,j}, w_{i,j+1}) \leq max\_gap_{j,j+1}$, $1 \leq j \leq n-1$.

Wzorce gramatyczne - przykład

A sentence:

"In Warsaw, in the buildings belonging to University there are auditoria with audio-visual equipment of advanced technology."

Wzorzec: <({noun}, {noun}, {preposition}, {noun}), (0,0,0)>

Szukanie nazwy własnej

Warsaw University of Technology

Zaznaczone słowa: "Warsaw", "University", "of" "Technology".

Słowa te nie spełniają wzorca, gdyż ograniczenia odległości nie są spełnione:

$distance_{ws}$(Warsaw, University) = 5 > max_gap_{12} = 0.

Odkrywanie terminów wielowyrazowych – etapy procesu

1. Wstępne przetwarzanie korpusu
2. Definicja wzorców gramatycznych.
3. Wykonanie algorytmu T-GSP

Wstępne przetwarzanie korpusu

- Podział dokumentów na paragrafy i zdania
- Oznaczenie słów częściami mowy
- Brak usuwania tzw. stopwords

Określanie autora tekstów

Stylometria

- **Stylometria** jest statystyczną analizą różnic stylu literackiego pomiędzy jednym autorem a drugim. Jest to badanie mierzalnych cech na podstawie wskaźników tekstowych, które charakteryzują styl autora.

Atrybuty

- częstotliwość występowania poszczególnych liter
- częstotliwość użycia wielkich liter
- całkowita liczba znaków na wyraz
- liczba znaków na zdanie
- częstotliwość użycia specyficznych wyrazów, n-gramów znakowych.
- Częstotliwość użycia słów funkcyjnych
- Częstotliwość występowania poszczególnych części mowy

Cechy strukturalne

- średnia długość akapitu,
- liczba punktów na dokument,
- obecność pewnych uporządkowanych układów, takich jak miejsce powitania czy adres e-mail

Inne atrybuty

Charakterystyczne cechy odnoszą się do występowania błędów, takich jak błędy pisowni i błędy składniowe w dokumencie.

Cechy stylometryczne - podsumowanie

Type of stylometric attribute	Example of attribute
Lexical	n-grammes word, distribution of word length, the average number of words per sentence, the richness of the vocabulary
Feature	Letter frequency, the frequency of uppercase letters, the total number of characters per word, the number of characters per phrase.
Syntax	Punctuation, functional words (of, which, above) and the part of speech labeling
Semantics	Synonym, hypernyms
Structural	Average paragraph length, the number of paragraphs per document, the greetings presence and their position into the document
Idiosyncrasy	There are misspelling and grammar collections

Określanie autora tekstu

Porównanie podejść: opartego na profilu (ang. profile-based) oraz opartego o przykłady (ang. instance based)

	Profile-based approach	**Instance-based approach**
Representation of the text	A cumulative representation of all the learning texts per author.	Each text is individually represented. The segmentation of the text can be required.
Stylometric specifications	Difficult to combine different characteristics.	Different characteristics are to be combined.
Classification	The generative model (i.e.: Bayesian) method based on similarity.	The discriminative model, SMV, method based on similarity.
Learning time	Low	Relatively high
Execution time	Low	Low