Statystyczna reprezentacja tekstu w NLP

Marzena Nowakowska
Wydział Zarządzania i Modelowania Komputerowego
Politechnika Świętokrzyska

Porządek reprezentacji statystycznej tekstu – jednostki/obiekty

Term (Termin) - najprostsza jednostka, zwykle pojedyncze słowo lub token. W statystycznej reprezentacji tekstu słowo najczęściej jest określane jako „atom znaczeniowy”. Powstaje w wyniku tokenizacji.

Stop list (Stoplista) - wykaz (lista) słów, które w kontekście analizy statystycznej są mało informacyjne i mogą wprowadzać szum. Najczęściej są to: spójniki, przyimki, zaimki, partykuły itp.

Zazwyczaj korzysta się nie tylko ze standardowej listy dostępnej w środowisku programistycznym (Python) - dostosowuje ją do własnego projektu, dodając słowa nieistotne dla danej domeny, co wpływa na ostateczną postać reprezentacji tekstu.

Dla języka polskiego:

['i', 'a', 'się', 'w', 'na', 'do', 'że', 'z', 'dla', 'o', 'przy', 'od', 'za', 'ale', 'też', 'czasem']

Dla języka angielskiego:

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're"]

Porządek reprezentacji statystycznej tekstu – jednostki/obiekty

Vocabulary (Słownik) - zbiór wszystkich unikalnych terminów w zbiorze tekstów (korpusie), które będą używane w reprezentacji wektorowej. Powstaje w wyniku wyboru wartości unikatowych.

Zaleca się tworzenie słownika po lematyzacji, dzięki czemu analiza statystyczna nie „rozprasza” informacji między różne formy tego samego słowa.

- Zmniejszenie wymiarowość reprezentacji
- Łączenie statystyk dla różnych form słowa
- Poprawa semantyki tekstu.

Porządek reprezentacji statystycznej tekstu – jednostki/obiekty

Thesaurus (Tezaurus) - ustrukturyzowany słownik, który grupuje wyrazy na podstawie ich relacji semantycznych (znaczeniowych).

Najważniejsze relacje w tezaurusie:

- synonimy: wyrazy o tym samym znaczeniu (np. samochód – auto)
- antonimy: wyrazy o przeciwnym znaczeniu (np. ciepły – zimny)
- homonimy: wyrazy o identycznym brzmieniu lub pisowni, ale zupełnie odmiennym znaczeniu i etymologii (np. zamek: budowla – mechanizm)
- hiperonimy: pojęcia ogólniejsze (np. mebel dla słowa krzesło)
- hiponimy: pojęcia bardziej szczegółowe (np. wróbel dla słowa ptak)

Wykorzystanie:

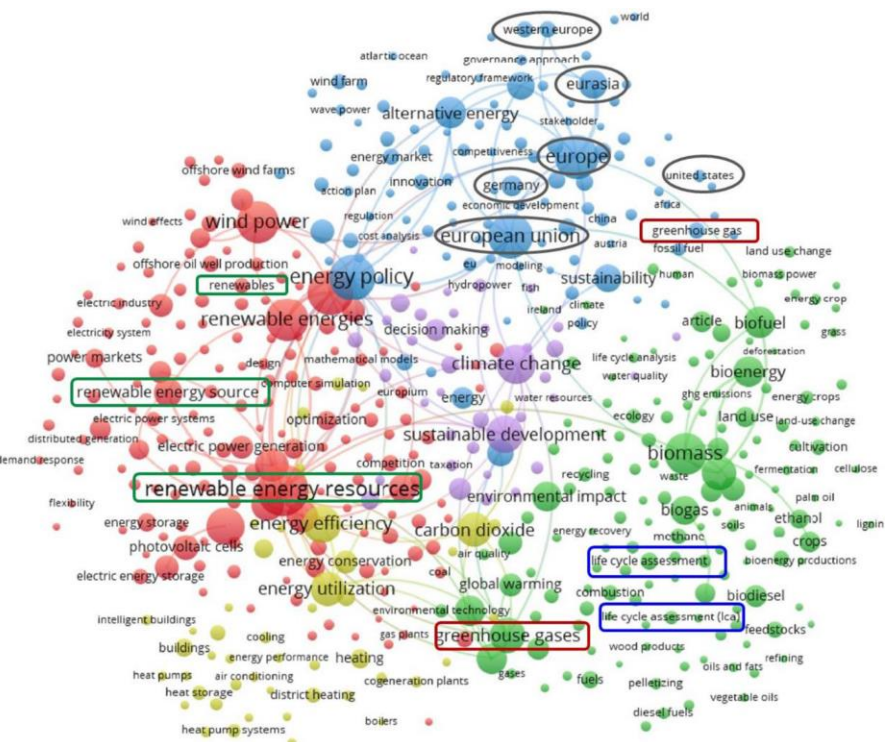
- ekspansja zapytań: dodawanie synonimów do wyszukiwanej frazy, aby znaleźć więcej trafnych dokumentów.
- redukcja wymiarowości: zastępowanie bliskoznacznych słów jednym wspólnym terminem.

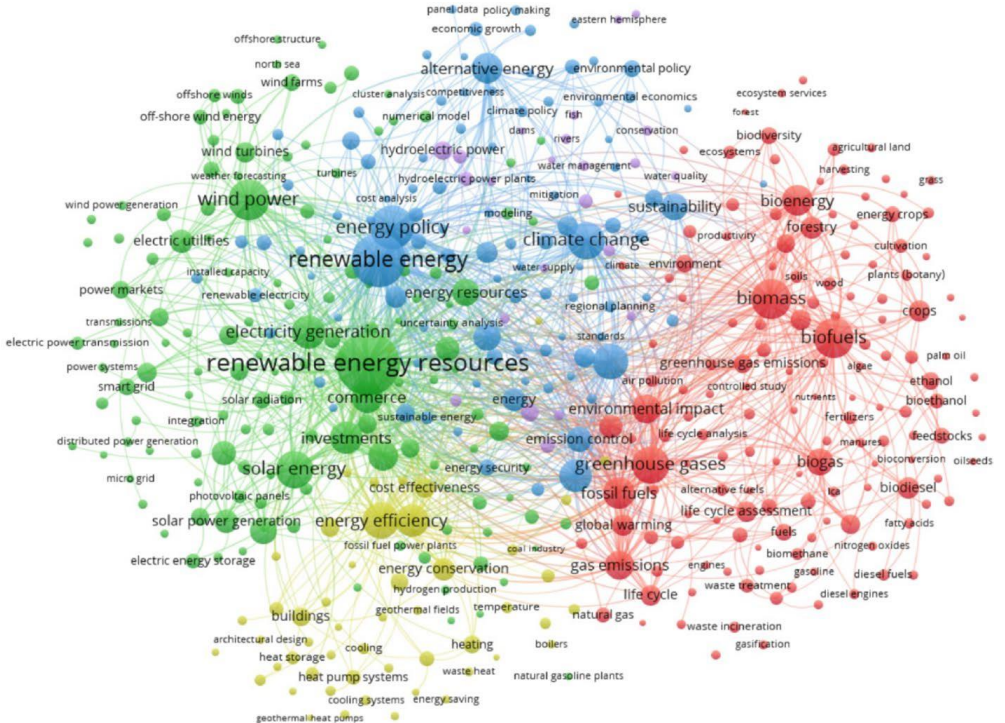
Cyfrowe tezaurusy: Wordnet, dla polskiego - Słowosieć

Tezaurus i stoplista - przykład znaczenia

Mapa 1 – bez zastosowanie tezaurusa i stoplisty	
163	renewable energy resources, renewable energy, renewable energies, wind power, solar energy, commerce, investments, costs
150	biomass, greenhouse gases, biofuels, biofuel, fossil fuels, gas emissions, bioenergy, environmental impact, biogas, carbon
103	energy policy, european union, europe, alternative energy, Eurasia
47	energy efficiency, carbon dioxide, energy utilization
37	climate change, sustainable development

Mapa 2 – po zastosowaniu tezaurusa i stoplisty	
149	biofuels, biomass, greenhouse gases, fossil fuels, bioenergy, gas emissions, environmental impact, biogas, carbon
136	renewable energy resources, wind power, solar energy, electricity generation, investments, commerce, costs
97	renewable energy, energy policy, climate change, sustainable development, carbon dioxide
51	energy efficiency, energy utilization, energy conservation
24	hydroelectric power, laws and legislation, hydropower, hydroelectric power plants





Reprezentacja statystyczna tekstu - BoF

Bag of Words (BOF, worek słów) - Najprostszy sposób reprezentacji tekstu w formie liczbowej. W tym modelu traktuje się dokument lub korpus jako dostawcę słów, całkowicie ignorując ich kolejność, gramatykę oraz kontekst. Liczy się tylko to, **czy** dane słowo wystąpiło i **ile razy**.

Ala ma kota i kot ma Alę po normalizacji i lematyzacji mają tę samą reprezentację w słowniku ['ala', 'mieć', 'kot']

Ilustracja graficzna BoF → **Cloud of Words (CoF)**

BoW

CoW

Reprezentacja statystyczna tekstu – T-DM, D-TM

Macierz termin-dokument (Term-Document Matrix, T-DM) - matematyczna struktura danych w formie tabeli służąca do reprezentacji zbioru tekstów.

- Wiersze odpowiadają unikalnym słowom (terminom) występującym w tekstach (cały korpus) – dostarczone przez słownik.
- Kolumny reprezentują poszczególne dokumenty (np. zdania, artykuły, pliki).
- Wartości w komórkach zawierają wagę danego słowa w konkretnym dokumencie.

	Doc 1	Doc 2	…	Doc N
Term 1	3	O	…	1
Term 2	1	2	…	3
…	…	…	…	…
Term M	2	3	…	0

Po transponowaniu otrzymuje się D-TM – Macierz Dokument-Termin

Wagi macierzy T-DM

Waga binarna (BW, Binary Weight) informuje, czy termin *t* wystąpił w danym dokumencie *d* korpusu:

$$BW(t,d) = \begin{cases} 1\ gdy\ t \in d \\ 0\ gdy\ t \notin d \end{cases}$$

Częstotliwość terminu (TF, Term Frequency) informuje ile razy termin *t* wystąpił w danym dokumencie *d* korpusu:

$$TF(t,d) = f(t,d)$$

Częstość dokumentowa (DF, Document Frequency) pokazuje liczbę dokumentów w korpusie, w których wystąpił termin *t* (*N* – liczność korpusu).

$$DF(t) = \sum_{d=1}^{N} BW(t,d)$$

Odwrotna częstość dokumentowa (IDF, Inverse Document Frequency) pokazuje, jak ważne lub rzadkie jest słowo *t* w korpusie.

$$IDF(t) = \log\left(\frac{N}{DF(t)}\right)$$

TF-IDF (Term Frequency Inverse Document Frequency) pokazuje, jak ważne jest dane słowo *t* dla konkretnego dokumentu *d* w kontekście całego korpusu.

$$TFIDF(t,d) = TF(t,d) \cdot IDF(t)$$

Wagi macierzy T-DM przykład

Korpus

D1: **Kot lubi mleko i kot lubi ryby.**
D2: **Kot lubi mleko.**
D3: **Pies lubi kości.**

Po lematyzacji

D1: **kot lubić mleko kot lubić ryba**
D2: **kot lubić mleko**
D3: **pies lubić kość**

Słownik korpusu

kot pies lubić mleko ryba kość

Binary Weight

Dokument	kot	pies	lubić	mleko	ryba	kość
D1	1	0	1	1	1	0
D2	1	0	1	1	0	0
D3	0	1	1	0	0	1

Term Frequency

Dokument	kot	pies	lubić	mleko	ryba	kość
D1	2	0	2	1	1	0
D2	1	0	1	1	0	0
D3	0	1	1	0	0	1

Document Frequency

Termin	DF
kot	2
pies	1
lubić	3
mleko	2
ryba	1
kość	1

Inverse Document Frequency

Termin	IDF	
kot	log(3/2)	0,176
pies	log(3/1)	0,477
lubić	log(3/3)	0
mleko	log(3/2)	0,176
ryba	log(3/1)	0,477
kość	log(3/1)	0,477

TF-IDF

Dokument	kot	pies	lubić	mleko	ryba	kość
D1	2 · 0,176	0 · 0,477	2 · 0	1 · 0,176	1 · 0,477	0 · 0,477
D2	1 · 0,176	0 ·0,477	1 · 0	1 · 0,176	0 · 0,477	0 · 0,477
D3	0 · 0,176	1 ·0,477	1 · 0	0 · 0,176	0 · 0,477	1 · 0,477

Pipeline statystycznej reprezentacji tekstu w NLP

Tekst przykładowy		**Koty uwielbiają myszy, a psy przepadają za królikami, ale koty też czasem lubią myszami się bawić.**
Tokenizacja Zachowanie interpunkcji i podział na tokeny/słowa		['Koty', 'uwielbiają', 'myszy', ',', 'a', 'psy', 'przepadają', 'za', 'królikami', ',', 'ale', 'koty', 'też', 'czasem', 'lubią', 'myszami', 'się', 'bawić', '.']
Lematyzacja Sprowadzenie słów do form podstawowych		['kot', 'uwielbiać', 'mysz', ',', 'a', 'pies', 'przepadać', 'za', 'królik', ',', 'ale', 'kot', 'też', 'czasem', 'lubić', 'mysz', 'się', 'bawić', '.']
Stoplista Typowe słowa mało informacyjne w danym języku: oraz interpunkcja		['kot', 'uwielbiać', 'mysz', 'pies', 'przepadać', 'królik', 'kot', 'lubić', 'mysz', 'bawić']
Słownik Zbiór unikalnych tokenów po lematyzacji i usunięciu stop listy		['kot', **'uwielbiać'**, 'mysz', 'pies', 'przepadać', 'królik', **'lubić'**, 'bawić']
Tezaurus Ujednolicenie synonimów w słowniku - jedna reprezentacja dla synonimów		['kot', **'lubić'**, 'mysz', 'pies', 'przepadać', 'królik', 'bawić']
Reprezentacja tekstu BoF – dla słownika policzone wystąpienia słów		['kot': 2, **'lubić': 2**, 'mysz': 2, 'pies': 1, 'przepadać': 1, 'królik': 1, 'bawić': 1]

Model tematyczny

Model tematyczny (Topic Model) to rodzaj modelu statystycznego (dokładnie modelu probabilistycznego) wykorzystanego do reprezentacji tekstu, która opisuje dokumenty poprzez ukryte tematy (latent topics) zamiast pojedynczych słów. W tym sensie stanowi wyższy poziom statystycznej reprezentacji tekstu niż np. Bag-of-Words czy macierz dokument–termin.

Modelowanie tematyczne (Topic Modelling) jest techniką z dziedziny przetwarzania języka naturalnego (NLP) i uczenia maszynowego, która służy do budowania modelu tematycznego – automatycznego odkrywania abstrakcyjnych „tematów” w dużych zbiorach tekstów, bez konieczności ich ręcznego czytania, grupując je na podstawie współwystępowania słów.

Najpopularniejszy algorytm: **LDA – Latent Dirichlet Allocation**

Idea modelu tematycznego

$\theta_d \sim \text{Dir}(\alpha)$; $\theta_d = [p_{d1}, p_{d2}, \ldots p_{dK}]$, $d = 1 : D$

This is a long road and no one will see it. But there will be some things that will make it to the finish line this year. We expect to see a solid start to the season. d = 1

$\theta_1 = [0, 0.5, 0.5, 0]$

He lied to me and tried to convince me he knew everything I knew. Typical. That is how he always won at tennis. He admitted it. d = 2

$\theta_2 = [0, 0.2, 0, 0.8]$

There are a huge number of birds in this area. The most common bird in this area is the golden yellow and it is the only other species that has ever been photographed. d = 3

$\theta_3 = [1, 0, 0, 0]$

$\phi_k \sim \text{Dir}(\beta)$; $\phi_k = [q_{k1}, q_{k2}, \ldots q_{kV}]$, $k = 1 : K$ – liczba tematów, V – liczba słów w słowniku

SCIENCE k = 1 $\phi_1 = [0.01, 0, \ldots]$

SPORT k = 2 $\phi_2 = [0.008, 0.12,$

HOPE k = 3 $\phi_3 = [0.001, 0, \ldots]$

DECEIT k = 4 $\phi_4 = [0, 0.09, \ldots]$

- Każdy dokument *d* jest mieszanką kilku tematów; model dla każdego dokumentu określa udział tematów korpusu (rozkład tematów korpusu) θ_d
- Każdy temat o numerze *k* jest reprezentowany przez specyficzny rozkład prawdopodobieństwa słów korpusu ϕ_k

Obszar przetwarzania

Zmienne obserwowalne

- Dokumenty
- Słowa w dokumentach

Zmienne ukryte (latentne)

- Zmienne, których nie widać i które LDA próbuje odkryć poprzez:
  - tematy (z_k) przypisane do każdego słowa; każde słowo w dokumencie jest generowane z jakiegoś tematu,
  - rozkłady tematów dla dokumentów (θ_d) i słów dla tematów (ϕ_k).

Temat można interpretować (określić) na podstawie słów o najwyższym prawdopodobieństwie.

Latentne zmienne „tłumaczą” obserwowane dane, ponieważ to one modelują, skąd pochodzi każde słowo (ukryte przyczyny generowania słów).

Cel przetwarzania

Celem algorytmu LDA jest znalezienie takich rozkładów latentnych zmiennych (przypisanie tematów *z* i związanych z nimi rozkładów θ, ϕ), które najlepiej wyjaśniają obserwowane dane (dokumenty i słowa) w sensie probabilistycznym, czyli które **maksymalizują prawdopodobieństwo obserwowanych słów W w dokumentach d**:

$$P(W|\alpha,\beta) = \sum_{Z} P(W,Z|\alpha,\beta)$$

W – wszystkie obserwowane słowa w korpusie
Z – wszystkie zmienne latentne (tematy słów)
α, *β* – parametry Dirichleta dla dokumentów i tematów

LDA „szuka” takich ukrytych tematów, które najlepiej tłumaczą, dlaczego widzi się określona konkretne słowa w określonych konkretnych dokumentach

Algorytm LDA – Gibbs sampling

1. Dane wejściowe: D – korpus dokumentów (wyczyszczony), K – liczba tematów, hiperparametry α (dokument → temat) i β (temat → słowo)
2. Inicjalizacja:
   - dla każdego słowa losowe przypisania tematu (topiku)
   - obliczenie liczników (*): $n_{d,k}$ – krotność tematu k obecnego w dokumencie d, $n_{k,w}$ – krotność słowa w obecnego w temacie k, n_k – liczba słów w temacie k
3. Iteracja – dla każdego dokumentu d i każdego słowa t w d:
   - usunięcie tematu ze słowa
   - przeliczenie liczników (*) (zmniejszenie)
   - wyznaczenie prawdopodobieństwa przypisania tematu k do t-tego słowa w dokumencie d (prawdopodobieństw jest tyle ile tematów)
   - zgodnie z ww. prawdopodobieństwem wylosowanie nowego tematu dla słowa (d, t) i przypisane tego tematu do słowa
   - przeliczenie liczników (*) (zwiększenie)
   - warunek stopu

     (wyczerpanie liczby iteracji) lub (brak istotnej poprawy jakości modelu)
4. Wynik (estymacje):
   - dla każdego dokumentu d rozkład tematów $\theta_d = [p_{d1}, p_{d2}, \ldots p_{dK}]$
   - dla każdego tematu k rozkład słów $\phi_k = [q_{k1}, q_{k2}, \ldots q_{kV}]$

Prawdopodobieństwa przypisania tematu *k* do słowa *t* w dokumencie *d*

$$P(z_{d,t} = k \mid z_{-(d,t)}, w) \propto (n_{d,k}^{-(d,t)} + \alpha) \cdot \frac{n_{k,w_{d,t}}^{-(d,t)} + \beta}{n_k^{-(d,t)} + V\beta}$$

$z_{d,t}$	temat *t*-tego słowa w dokumencie *d*
k	ukryty temat (topik) przypisywany do słowa
$z_{-(d,t)}$	zbiór wszystkich pozostałych przypisanych tematów do słów w całym korpusie, z wyłączeniem badanego słowa; na ich podstawie wyliczane są liczebności $n_{d,k}$ i $n_{k,w}$
w	korpus tekstowy; zbiór wszystkich słów we wszystkich dokumentach
$n_{d,k}^{-(d,t)}$	liczba słów w dokumencie *d* przypisanych do tematu *k* z wyłączeniem badanego słowa
$n_{d,k}^{-(d,t)}$	ile razy temat występuje w dokumencie *d* z wyłączenia badanego słowa
$n_{k,w_{d,t}}^{-(d,t)}$	ile razy badane słowo pojawia się w temacie *k* z wyłączenia bieżącego (badanego, konkretnego) wystąpienia tego słowa (*d*, *t*)
$n_k^{-(d,t)}$	ile wszystkich słów korpus przypisano do tematu *k* wykluczając bieżące wystąpienie słowa (*d*, *t*)
V	liczba słów w słowniku (liczba unikalnych słów w korpusie dokumentów)
$\alpha > 0$ hiperparametr wygładzający	kontroluje, ile tematów ma dokument; mała wartość – dokument „skupiony” z małą liczbą tematów; duża wartość – dokument ma wiele tematów, rozproszony. Wartość: zazwyczaj 50/K lub z przedziału (0, 1)
$\beta > 0$ hiperparametr wygładzający	kontroluje, ile różnych słów jest w temacie; mała wartość (np. 0,1) – temat ma małą liczbę charakterystycznych słów, duża wartość – temat ma wiele słów (może być rozmyty). Wartość: zazwyczaj: 0,01 lub z przedziału (0, 1)

Wagi tematów w korpusie

Opcja A. Średnia prawdopodobieństw po dokumentach – daje globalną popularność tematu, każdy dokument ma taka sama wagę

$$P(z_k) \approx \frac{1}{D} \sum_{d=1}^{D} p_{dk}$$

Opcja B. Średnia ważona prawdopodobieństw po dokumentach – na ważność wpływa długość dokumentu (N_d), w którym występuje temat

$$P(z_k) \approx \frac{1}{\sum_{d=1}^{D} N_d} \sum_{d=1}^{D} N_d \cdot p_{dk}$$

Opcja C. Waga słów tematu – bezpośrednie przypisanie ze słów

$$P(z_k) = \frac{liczba\ słów\ przypisanych\ do\ k}{liczba\ wszystkich\ słów}$$

z_k oznacza k-ty temat, $k = 1, \ldots, K$

Wybór liczby tematów *K*

Spójność tematyczna (Topic Coherence); mierzy, jak bardzo słowa wewnątrz jednego tematu są ze sobą powiązane semantycznie. Wartość zazwyczaj skalowana do przedziału (0, 1). Im większa tym lepsza spójność.

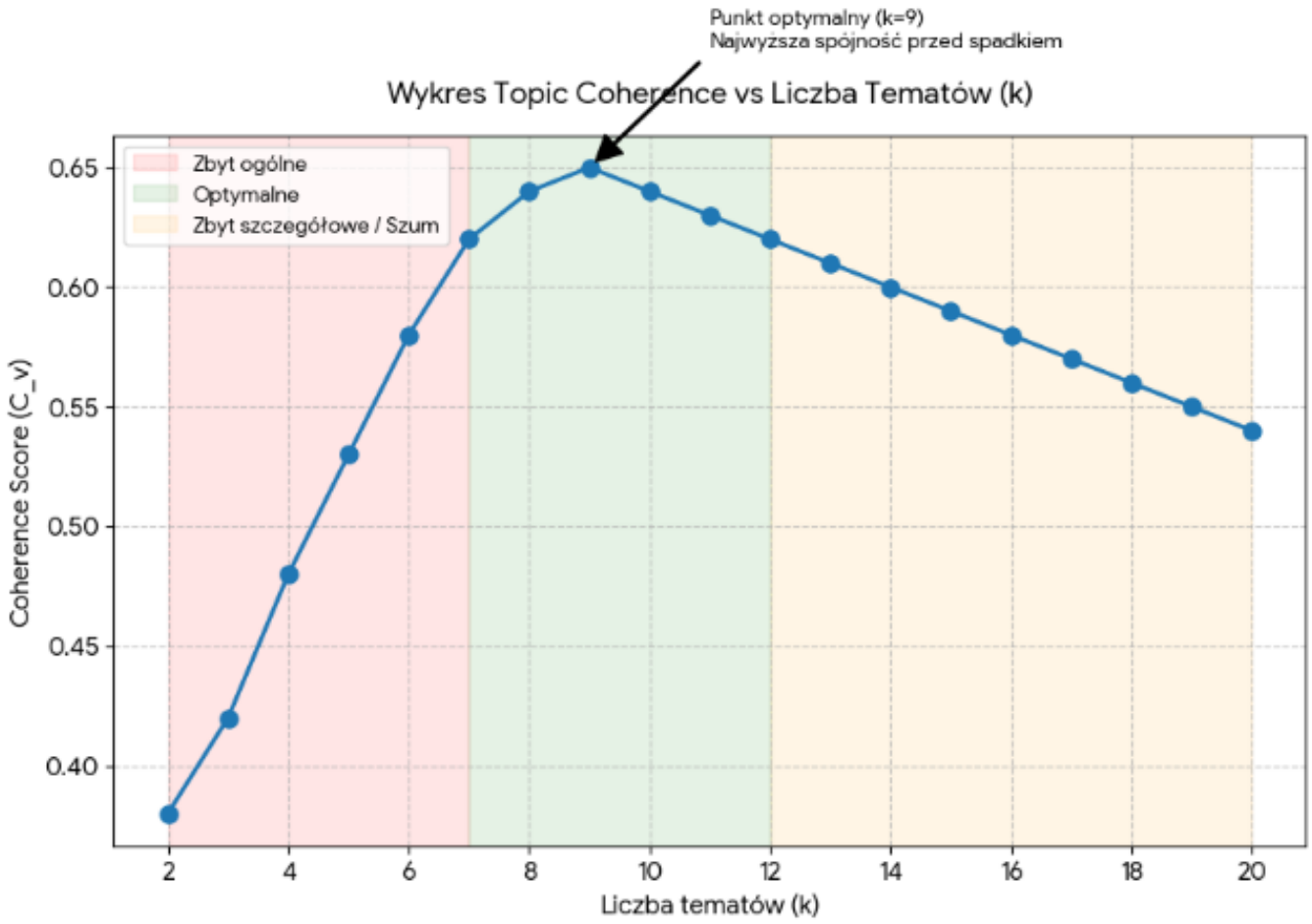
Perpleksja (Perplexity); mierzy, jak dobrze model przewiduje próbki tekstu (zaskoczenie modelu nowymi danymi); im niższa tym model lepszy.

Algorytm postępowania

1. Określić zakres K
2. Wykonać LDA dla kolejnych $k = 1, \ldots, K$
3. Dla każdego modelu policzyć (ew. zrobić wykresy)
   - spójność tematyczna (priorytet)
   - perpleksja (pomocniczo)
   - sprawdzić tematy "ręcznie"
4. Wybrać kompromis:
   - dobra spójność
   - sensowna interpretacja

Wynik: dla ustalonego k określenie etykiety (nazwy, tytułu) dla każdego tematu na podstawie najmocniejszych występujących w nim słów kluczowych.
Jeśli trudno ustalić jasną nazwę tematu, być może model jest źle dopasowany.

Wykres spójności – wybór liczby tematów k



Czasami lepiej wybrać nieco mniejszą liczbę tematów (niżej niż maksimum), która jest w pełni zrozumiała dla człowieka, niż najwyższy statystycznie punkt, który tworzy tematy trudne do zinterpretowania.

Spójność tematyczna

Topic Coherence. Określa, na ile słowa wewnątrz tematu są ze sobą powiązane semantycznie. Im wyższy wynik, tym bardziej „sensowny” jest dany temat dla osoby.

Algorytm obliczania C_{PMI} (Pointwise Mutual Information)

1. Segmentacja selekcja n najsilniejszych słów tematu i utworzenie z nich par „każdy z każdym”
2. Estymacja prawdopodobieństwa; dla każdego słowa i każdej pary słów:
   - $P(w_i)$ = (liczba dokumentów w korpusie zawierających słowo w_i)/ (liczba wszystkich dokumentów)
   - $P(w_i, w_j)$ = (liczba dokumentów w korpusie zawierających jednocześnie słowa w_i oraz w_j)/ (liczba wszystkich dokumentów)
3. Dla każdej pary słów (wi , wj), $i < j$, obliczenie wyniku ($Score$), używając wybranej formuły. Dla PMI:

$$Score(w_i, w_j) = PMI(w_i , w_j) = \log \frac{P(w_i,w_j)}{P(w_i)\cdot P(w_j)}$$

Często do prawdopodobieństwa w liczniku dodaje się małą stałą c, aby uniknąć log(0)

4. Coherence C_{topic} dla tematu – średnia arytmetyczna miary Coherence dla wszystkich unikalnych par słów z listy *N* najlepszych słów tematu

$$C_{topic} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j)$$

5. Coherence C_{model} dla modelu – średnia arytmetyczne miar Cohernce dla całkowitej liczby *K* tematów modelu

$$C_{model} = \frac{1}{K} \sum_{k=1}^{K} C_{topic}^{(k)}$$

Spójność tematyczna - komentarz

Model tematyczny generuje listę słów dla każdego tematu, posortowaną według ich ważności (prawdopodobieństwa wewnątrz tematu). Przy obliczaniu spójności tematu nie analizuje się wszystkich słów ze słownika, lecz tylko te z „góry” listy tematu.

- Zazwyczaj przyjmuje się $N = 10$ lub $N = 20$. Oznacza to, że ocenia się spójność tematu na podstawie jego 10 lub 20 najważniejszych słów.
- W mianowniku wzoru na agregację wartość $2/(N(N-1))$ służy do obliczenia średniej. Mianownik N(N-1) jest liczbą wszystkich możliwych par unikatowych , jakie można utworzyć z N słów.
- Jeżeli wybierze się zbyt małe N, miara może nie oddać pełni znaczenia tematu. Jeśli zbyt duże wynik może zostać zaniżony przez mniej istotne, szumiące słowa z końca listy.

Uwaga do zależności na PMI
W praktyce stosuje się znormalizowaną miarę *PMI:* $$NPMI(w_i, w_j) = \frac{\log \frac{P(w_i,w_j)}{P(w_i)\cdot P(w_j)}}{-\log P(w_i, w_j)}$$

NPMI skaluje wynik do przedziału [-1, 1], co zapobiega faworyzowaniu bardzo radkich słów
Interpretacji wartości NPMI:

- dodatnia: słowa mają tendencję do współwystępowania (spójny temat).
- bliska 0: słowa pojawiają się niezależnie od siebie.
- ujemna: słowa unikają swojego towarzystwa.

Filozofia perpleksji

Perplexity w modelowaniu tematycznym (w szczególności w LDA) to miara statystyczna określająca, jak dobrze model probabilistyczny przewiduje nowe wcześniej niewidziane dane (miara „zaskoczenia” nowymi dokumentami).

Perpleksja oblicza na podstawie modelu prawdopodobieństwo wystąpienia obserwowanych słów w zbiorze testowym, biorąc pod uwagę wywnioskowane tematy.

Wysokie prawdopodobieństwo = Niskie perpleksja (dobre dopasowanie)
Niskie prawdopodobieństwo = Wysokie perpleksja (słabe dopasowanie)

Paradoks perpleksywności

- Chociaż niższy poziom perpleksywności wskazuje na lepsze dopasowanie statystyczne, nie gwarantuje on, że zagadnienia będą zrozumiałe dla człowieka.
- Wysoce zoptymalizowane modele (niski poziom perpleksywności) mogą czasami prowadzić do bezsensownych zagadnień, które ludziom trudno opisać.
- Perpleksja mierzy moc predykcyjną, a nie spójność semantyczną.
- Należy ja traktować jako miarę pomocniczą.

Cechy i mechanizmy działania TM

Algorytm należy do obszaru uczenia nienadzorowanego.

Temat nie jest zdefiniowanym pojęciem, lecz klastrem (grupą) słów, które często pojawiają się blisko siebie. Jest reprezentowany przez specyficzny rozkład prawdopodobieństwa słów korpusu.

Zakłada się, że pojedynczy dokument (np. artykuł) nie musi dotyczyć tylko jednej kwestii, lecz może być mieszanką wielu tematów w różnych proporcjach. Model dla każdego dokumentu określa udział tematów korpusu, czyli rozkład tematów korpusu.

Współcześnie modelowanie tematyczne przeżywa renesans dzięki wykorzystaniu **dużych modeli językowych (LLM)**, które pozwalają na precyzyjniejsze i bardziej kontekstowe wyodrębnianie znaczeń z tekstu. Obecnie: BERTopic.