News Selector

czyli jak automatycznie podsumować wiadomości z polskich serwisów informacyjnych

Jakub Kubajek

Spis treści

Wstęp

Motywacja

- Keynote speech **Lynn Cherny** podczas PyData w Warszawie: *Tl;dr: Summarisation*

Motywacja

- Keynote speech **Lynn Cherny** podczas PyData w Warszawie: *Tl;dr: Summarisation*
- Setki artykułów dziennie

Charakterystyka modelu

Charakterystyka

- Uczenie nienadzorowane

Charakterystyka modelu

Charakterystyka

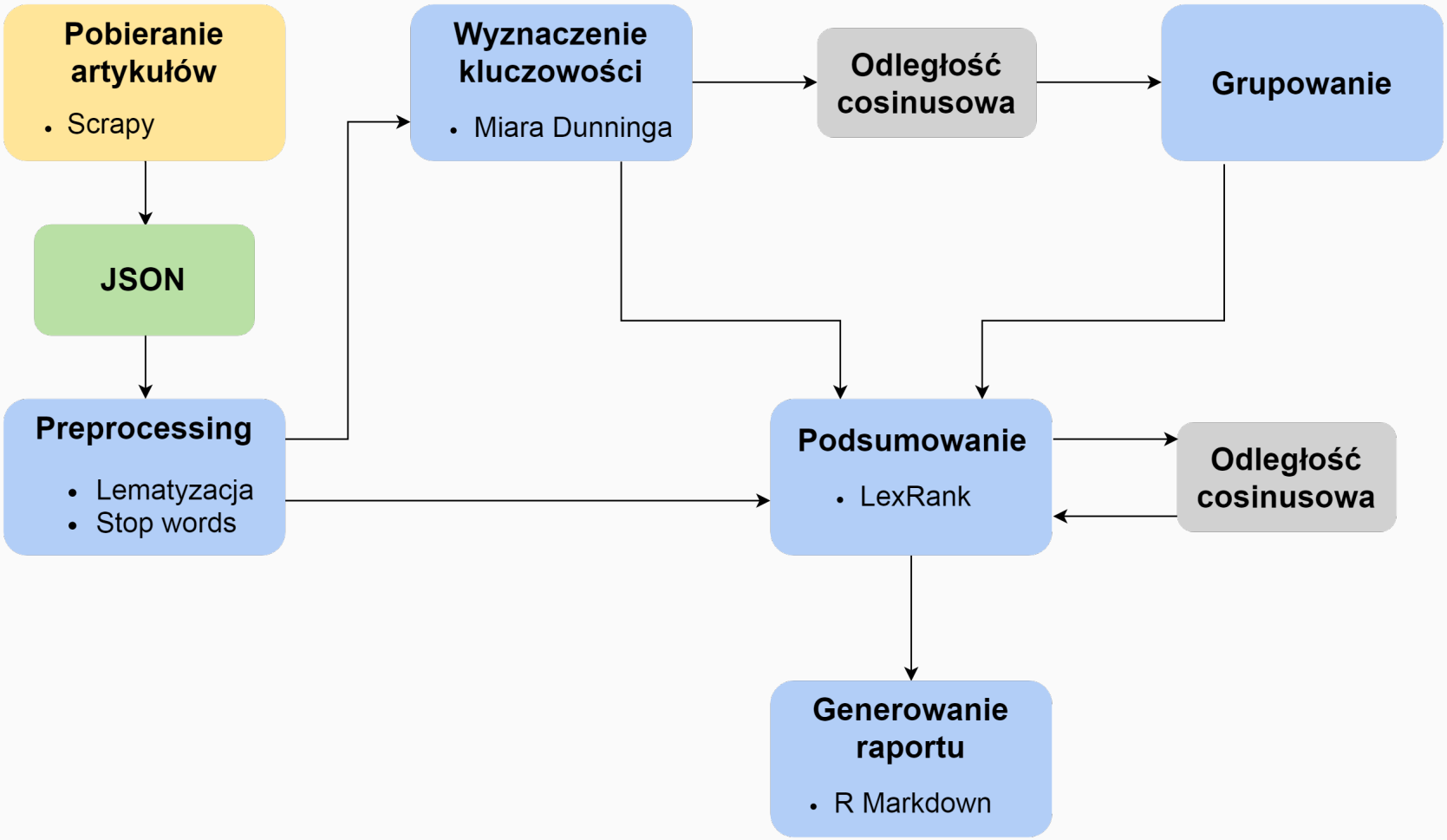
- Uczenie nienadzorowane
- Podejście słownikowe - *Morfeusz*

Charakterystyka modelu

Charakterystyka

- Uczenie nienadzorowane
- Podejście słownikowe - *Morfeusz*
  - Sprowadzanie słów do lematu

Schemat modelu



Pobieranie artykułów

Scrapy

Scrapy

- Asynchroniczność
- 10x szybszy niż *rvest*
- Obsługa błędów

Wybrane portale informacyjne

GAZETA.PL

dziennik.pl

Kluczowość słów

Miara Dunninga

Informacje ogólne

- Test ilorazu wiarygodności (*Likelihood ratio test)*
- Brak założenia o normalności rozkładu
- Odpowiednie dla zjawisk o małej liczbie wystąpień

Miara Dunninga

Informacje ogólne

- Test ilorazu wiarygodności (*Likelihood ratio test)*
- Brak założenia o normalności rozkładu
- Odpowiednie dla zjawisk o małej liczbie wystąpień

Wzór

$$\lambda = \frac{L\left(p, k_0, n_0\right) L\left(p, k_1, n_1\right)}{L\left(p_0, k_0, n_0\right) L\left(p_1, k_1, n_1\right)} \tag{1}$$

$$L\left(p, k, n\right) = p^k \left(1-p\right)^{n-k} \tag{2}$$

gdzie, $p = \frac{k_0+k_1}{n_0+n_1}$, $p_0 = \frac{k_0}{n_0}$, zaś $\boldsymbol{-2log\lambda}$ ma rozkład χ^2 z jednym stopniem swobody. k_0 to liczba wystąpień danego słowa z wykluczeniem badanego dnia, k_1 to liczba wystąpień danego słowa w badanym dniu.

Implementacja w modelu

Modyfikacja miary Dunninga

- Mnożenie statystyki przez -1, gdy $p_1 < p_0$
- Brak logarytmizacji zer - $max(eps,\, p)$

Implementacja w modelu

Modyfikacja miary Dunninga

- Mnożenie statystyki przez -1, gdy $p_1 < p_0$
- Brak logarytmizacji zer - $max(eps,\, p)$

Kryteria wyboru słów

- Wartość statystyki χ^2 większa od 10
- Liczba wystąpień większa od wartości 90 percentyla

Grupowanie

Podobieństwo cosinusowe

Opis

- Miara podobieństwa dwóch wektorów
- Nie uwzględnia kolejności - *bag of words*
- Przyjmuje wartości z zakresu $[0, 1]$

Podobieństwo cosinusowe

Opis

- Miara podobieństwa dwóch wektorów
- Nie uwzględnia kolejności - *bag of words*
- Przyjmuje wartości z zakresu $[0, 1]$

Wzór

$$cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Grupowanie hierarchiczne

Podział

- Aglomeracyjny - zaczynamy od n grup
- Deglomeracyjny - zaczynamy o 1 grupy

Grupowanie hierarchiczne

Podział

- Aglomeracyjny - zaczynamy od n grup
- Deglomeracyjny - zaczynamy o 1 grupy

Metody łączenia w grupy

- Pojedyncze połączenie - $D(X, Y) = min\{d(x, y) : x \in X,\ y \in Y\}$
- Kompletne połączenie - $D(X, Y) = max\{d(x, y) : x \in X,\ y \in Y\}$
- Średnie połączenie - $D(X, Y) = mean\{d(x, y) : x \in X,\ y \in Y\}$

Grupowanie hierarchiczne

Grupowania w modelu

- Aglomeracyjne

Grupowanie hierarchiczne

Grupowania w modelu

- Aglomeracyjne
- Macierz podobieństwa między słowami (grupami) - wspólne występowanie w akapitach

Grupowanie hierarchiczne

Grupowania w modelu

- Aglomeracyjne
- Macierz podobieństwa między słowami (grupami) - wspólne występowanie w akapitach
- Aktualizacja macierzy podobieństwa po stworzeniu nowej grupy:

Grupowanie hierarchiczne

Grupowania w modelu

- Aglomeracyjne
- Macierz podobieństwa między słowami (grupami) - wspólne występowanie w akapitach
- Aktualizacja macierzy podobieństwa po stworzeniu nowej grupy:
  - Sumowanie wektorów dystrybucji połączonych grup

Grupowanie hierarchiczne

Grupowania w modelu

- Aglomeracyjne
- Macierz podobieństwa między słowami (grupami) - wspólne występowanie w akapitach
- Aktualizacja macierzy podobieństwa po stworzeniu nowej grupy:
  - Sumowanie wektorów dystrybucji połączonych grup
  - Wyznaczenie podobieństwa między nową grupą, a pozostałymi grupami

Grupowanie hierarchiczne

Grupowania w modelu

- Aglomeracyjne
- Macierz podobieństwa między słowami (grupami) - wspólne występowanie w akapitach
- Aktualizacja macierzy podobieństwa po stworzeniu nowej grupy:
  - Sumowanie wektorów dystrybucji połączonych grup
  - Wyznaczenie podobieństwa między nową grupą, a pozostałymi grupami
- Zakończenie grupowania

Grupowanie hierarchiczne

Grupowania w modelu

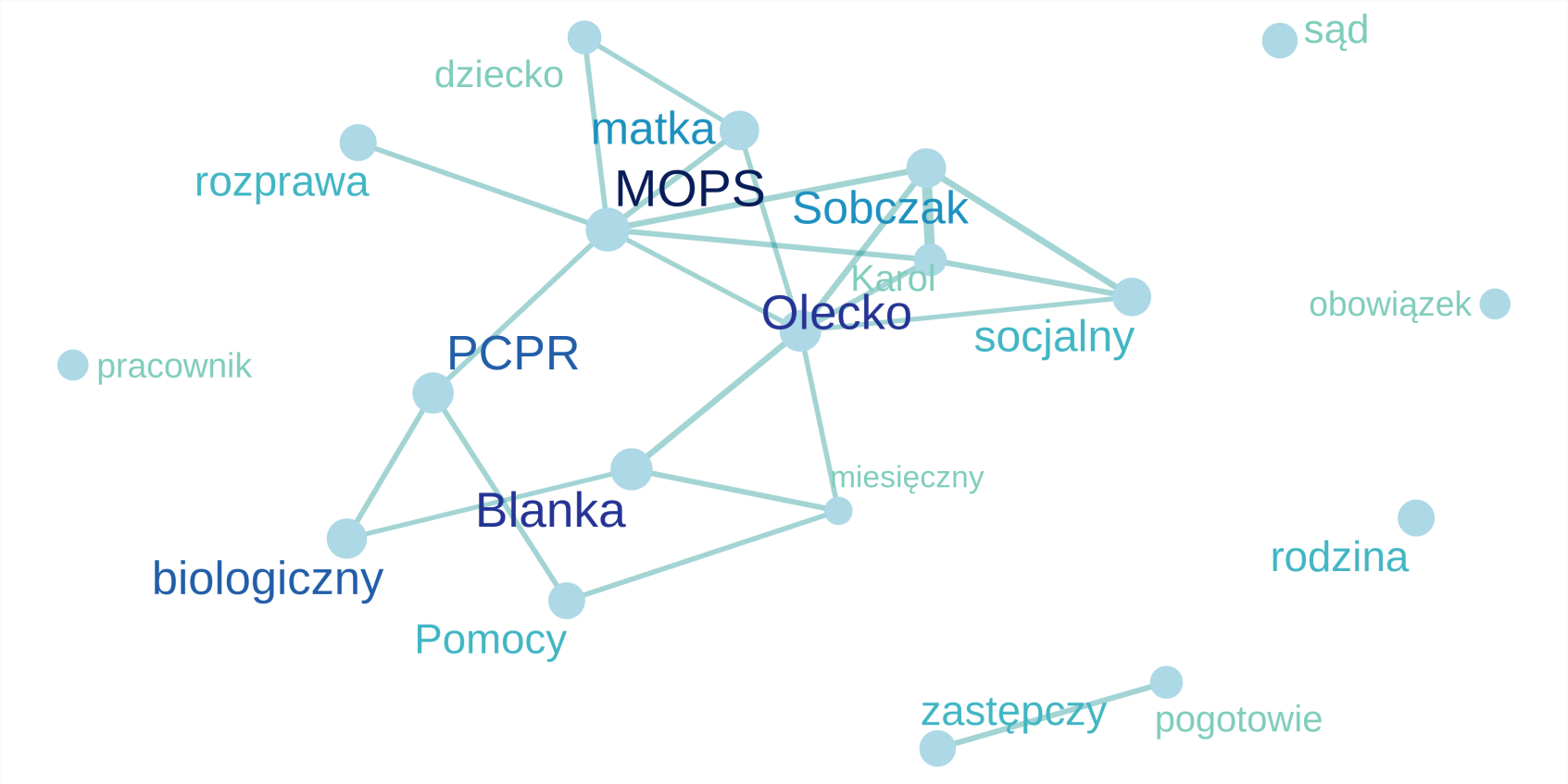
- Aglomeracyjne
- Macierz podobieństwa między słowami (grupami) - wspólne występowanie w akapitach
- Aktualizacja macierzy podobieństwa po stworzeniu nowej grupy:
  - Sumowanie wektorów dystrybucji połączonych grup
  - Wyznaczenie podobieństwa między nową grupą, a pozostałymi grupami
- Zakończenie grupowania
  - Minimalna wartość podobieństwa

Grupowanie hierarchiczne

Grupowania w modelu

- Aglomeracyjne
- Macierz podobieństwa między słowami (grupami) - wspólne występowanie w akapitach
- Aktualizacja macierzy podobieństwa po stworzeniu nowej grupy:
  - Sumowanie wektorów dystrybucji połączonych grup
  - Wyznaczenie podobieństwa między nową grupą, a pozostałymi grupami
- Zakończenie grupowania
  - Minimalna wartość podobieństwa
  - Algorytm Silhoutte

Przykład


sąd
dziecko
matka
rozprawa
MOPS
Sobczak
Karol
Olecko
obowiązek
socjalny
pracownik
PCPR
miesięczny
Blanka
biologiczny
rodzina
Pomocy
zastępczy
pogotowie

Automatyczne podsumowanie

LexRank

- Algorytm grafowy, bazujący na idei algorytmu *PageRank*

LexRank

- Algorytm grafowy, bazujący na idei algorytmu *PageRank*
- Wybiera te zdania, które zawierają najwięcej odniesień (słów) do innych zdań

LexRank

- Algorytm grafowy, bazujący na idei algorytmu *PageRank*
- Wybiera te zdania, które zawierają najwięcej odniesień (słów) do innych zdań
- Połączenia między zdaniami to podobieństwo cosinusowe

LexRank

- Algorytm grafowy, bazujący na idei algorytmu *PageRank*
- Wybiera te zdania, które zawierają najwięcej odniesień (słów) do innych zdań
- Połączenia między zdaniami to podobieństwo cosinusowe
- Wagi poszczególnych słów do wyznaczenia podobieństwa - $TF \cdot IDF$

LexRank

- Algorytm grafowy, bazujący na idei algorytmu *PageRank*
- Wybiera te zdania, które zawierają najwięcej odniesień (słów) do innych zdań
- Połączenia między zdaniami to podobieństwo cosinusowe
- Wagi poszczególnych słów do wyznaczenia podobieństwa - $TF \cdot IDF$
  - TF (*Term Frequency*) - liczba wystąpień danego słowa

LexRank

- Algorytm grafowy, bazujący na idei algorytmu *PageRank*
- Wybiera te zdania, które zawierają najwięcej odniesień (słów) do innych zdań
- Połączenia między zdaniami to podobieństwo cosinusowe
- Wagi poszczególnych słów do wyznaczenia podobieństwa - $TF \cdot IDF$
  - TF (*Term Frequency*) - liczba wystąpień danego słowa
  - IDF (*Inverse Document Frequency*) - $IDF_i = log(\frac{D}{d_i}) + 1$, gdzie D to liczba zdań, a d_i to liczba zdań w których wystąpiło i-te słowo

LexRank

- Algorytm grafowy, bazujący na idei algorytmu *PageRank*
- Wybiera te zdania, które zawierają najwięcej odniesień (słów) do innych zdań
- Połączenia między zdaniami to podobieństwo cosinusowe
- Wagi poszczególnych słów do wyznaczenia podobieństwa - $TF \cdot IDF$
  - TF (*Term Frequency*) - liczba wystąpień danego słowa
  - IDF (*Inverse Document Frequency*) - $IDF_i = log(\frac{D}{d_i}) + 1$, gdzie D to liczba zdań, a d_i to liczba zdań w których wystąpiło i-te słowo
- Ranking to stan ustalony łańcucha Markowa zbudowanego jako ważona suma:

LexRank

- Algorytm grafowy, bazujący na idei algorytmu *PageRank*
- Wybiera te zdania, które zawierają najwięcej odniesień (słów) do innych zdań
- Połączenia między zdaniami to podobieństwo cosinusowe
- Wagi poszczególnych słów do wyznaczenia podobieństwa - $TF \cdot IDF$
  - TF (*Term Frequency*) - liczba wystąpień danego słowa
  - IDF (*Inverse Document Frequency*) - $IDF_i = log(\frac{D}{d_i}) + 1$, gdzie D to liczba zdań, a d_i to liczba zdań w których wystąpiło i-te słowo
- Ranking to stan ustalony łańcucha Markowa zbudowanego jako ważona suma:
  - Macierzy podobieństwa między zdaniami

LexRank

- Algorytm grafowy, bazujący na idei algorytmu *PageRank*
- Wybiera te zdania, które zawierają najwięcej odniesień (słów) do innych zdań
- Połączenia między zdaniami to podobieństwo cosinusowe
- Wagi poszczególnych słów do wyznaczenia podobieństwa - $TF \cdot IDF$
  - TF (*Term Frequency*) - liczba wystąpień danego słowa
  - IDF (*Inverse Document Frequency*) - $IDF_i = log(\frac{D}{d_i}) + 1$, gdzie D to liczba zdań, a d_i to liczba zdań w których wystąpiło i-te słowo
- Ranking to stan ustalony łańcucha Markowa zbudowanego jako ważona suma:
  - Macierzy podobieństwa między zdaniami
  - Macierzy *losowego przejścia* między zdaniami

Implementacja w modelu

Modyfikacje

- Wybór artykułów - minimalny odsetek słów z tematu

Implementacja w modelu

Modyfikacje

- Wybór artykułów - minimalny odsetek słów z tematu
- Wagi słów w podobieństwie cosinusowym

Implementacja w modelu

Modyfikacje

- Wybór artykułów - minimalny odsetek słów z tematu
- Wagi słów w podobieństwie cosinusowym
  - TF - liczba wystąpień danego słowa

Implementacja w modelu

Modyfikacje

- Wybór artykułów - minimalny odsetek słów z tematu
- Wagi słów w podobieństwie cosinusowym
  - TF - liczba wystąpień danego słowa
  - Zmodyfikowana miara Dunninga

Implementacja w modelu

Modyfikacje

- Wybór artykułów - minimalny odsetek słów z tematu
- Wagi słów w podobieństwie cosinusowym
  - TF - liczba wystąpień danego słowa
  - Zmodyfikowana miara Dunninga
    - $log(D_i + 1)$ dla $p_1 \geq p_0$

Implementacja w modelu

Modyfikacje

- Wybór artykułów - minimalny odsetek słów z tematu
- Wagi słów w podobieństwie cosinusowym
  - TF - liczba wystąpień danego słowa
  - Zmodyfikowana miara Dunninga
    - $log(D_i + 1)$ dla $p_1 \geq p_0$
    - 0 dla $p_1 < p_0$

Implementacja w modelu

Modyfikacje

- Wybór artykułów - minimalny odsetek słów z tematu
- Wagi słów w podobieństwie cosinusowym
  - TF - liczba wystąpień danego słowa
  - Zmodyfikowana miara Dunninga
    - $log(D_i + 1)$ dla $p_1 \geq p_0$
    - 0 dla $p_1 < p_0$
  - Podobieństwo cosinusowe występowania danego słowa do występowania tematu (grupy słów)

Implementacja w modelu

Modyfikacje

- Wybór artykułów - minimalny odsetek słów z tematu
- Wagi słów w podobieństwie cosinusowym
  - TF - liczba wystąpień danego słowa
  - Zmodyfikowana miara Dunninga
    - $log(D_i + 1)$ dla $p_1 \geq p_0$
    - 0 dla $p_1 < p_0$
  - Podobieństwo cosinusowe występowania danego słowa do występowania tematu (grupy słów)
- Skalowanie wartości rankingu - $F_i = \frac{log(w_i^T)}{log(w_i)}$

Implementacja w modelu

Modyfikacje

- Wybór artykułów - minimalny odsetek słów z tematu
- Wagi słów w podobieństwie cosinusowym
  - TF - liczba wystąpień danego słowa
  - Zmodyfikowana miara Dunninga
    - $log(D_i + 1)$ dla $p_1 \geq p_0$
    - 0 dla $p_1 < p_0$
  - Podobieństwo cosinusowe występowania danego słowa do występowania tematu (grupy słów)
- Skalowanie wartości rankingu - $F_i = \frac{log(w_i^T)}{log(w_i)}$
  - gdzie F_i to współczynnik skalujący, w_i liczba słów z tematu w i-tym zdaniu, a W liczba słów w temacie

Implementacja w modelu

Modyfikacje

- Wybór artykułów - minimalny odsetek słów z tematu
- Wagi słów w podobieństwie cosinusowym
  - TF - liczba wystąpień danego słowa
  - Zmodyfikowana miara Dunninga
    - $log(D_i + 1)$ dla $p_1 \geq p_0$
    - 0 dla $p_1 < p_0$
  - Podobieństwo cosinusowe występowania danego słowa do występowania tematu (grupy słów)
- Skalowanie wartości rankingu - $F_i = \frac{log(w_i^T)}{log(w_i)}$
  - gdzie F_i to współczynnik skalujący, w_i liczba słów z tematu w i-tym zdaniu, a W liczba słów w temacie
- Pominięcie podobnych zdań ($cos(\theta) > 0,4$)

Bibliografia i

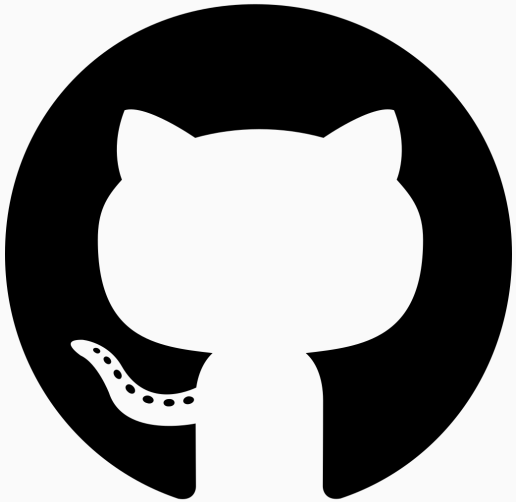

T. Dunning.
Accurate methods for the statistics of surprise and coincidence.

COMPUTATIONAL LINGUISTICS, 19(1):61–74, 1993.


G. Erkan and D. R. Radev.
Lexrank: Graph-based lexical centrality as salience in text summarization.
Journal of artificial intelligence research, 22:457–479, 2004.


L. Page, S. Brin, R. Motwani, and T. Winograd.
The pagerank citation ranking: Bringing order to the web.
Technical report, Stanford InfoLab, 1999.

GitHub



jkubajek