



NEKST + OSA =

SUKCES

CEL: Budowa antyplagiatu



Framework

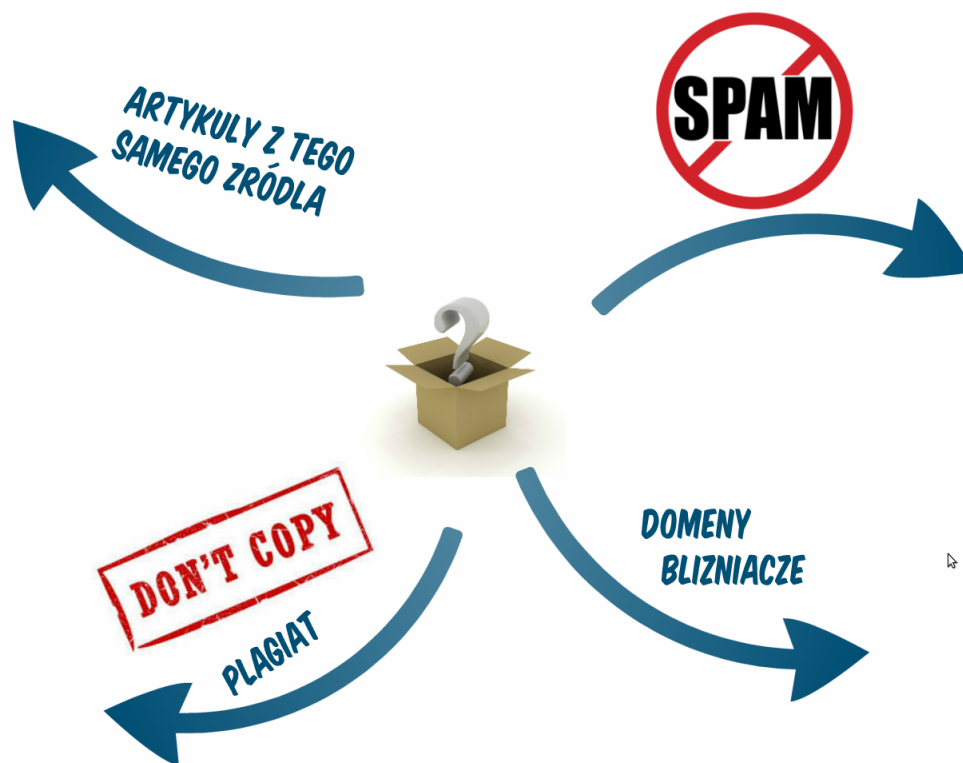


MAPREDUCE

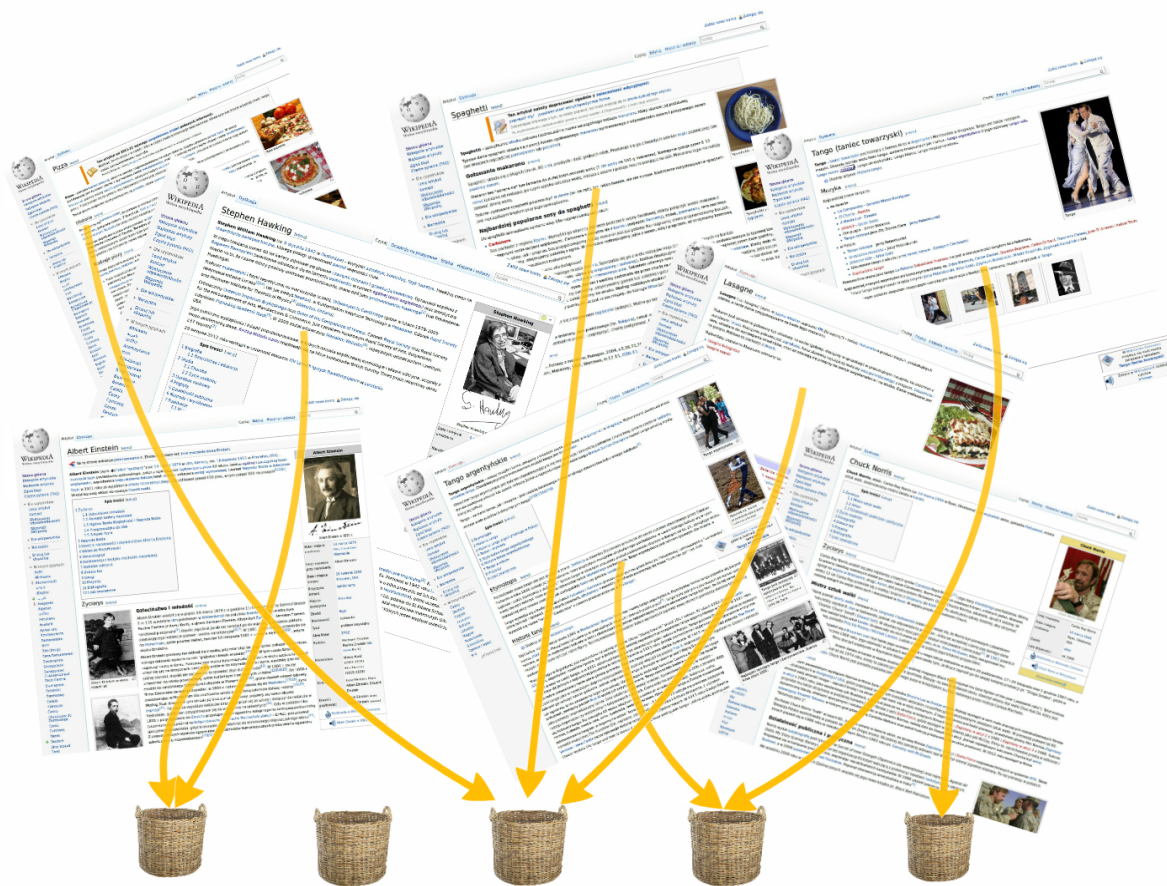
DDB

HDFS

Plagiaty a duplikaty



Detekcja duplikatów

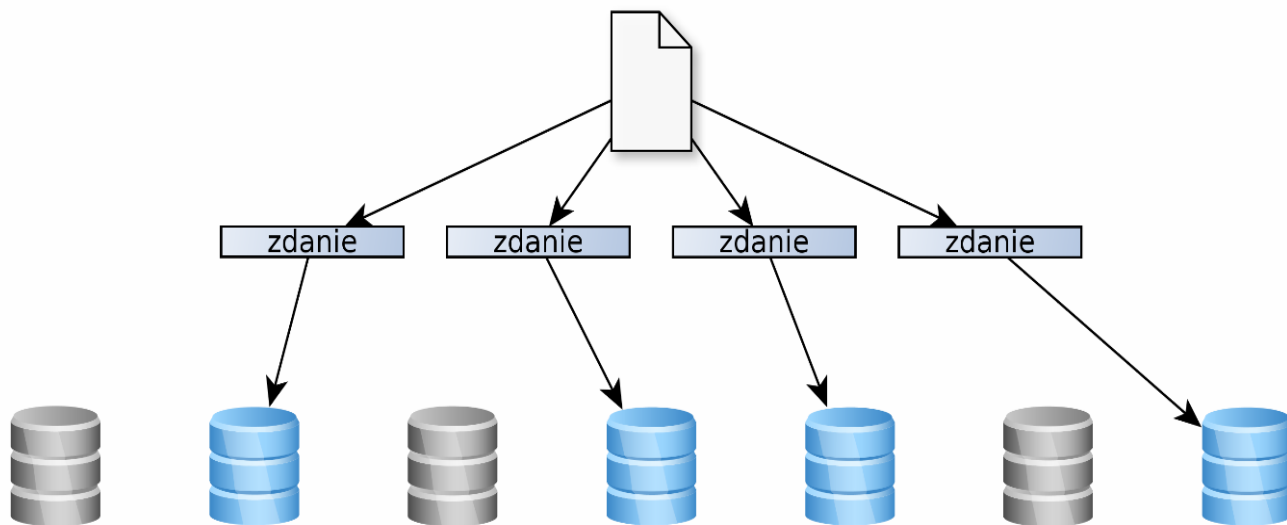


Detekcja duplikatów



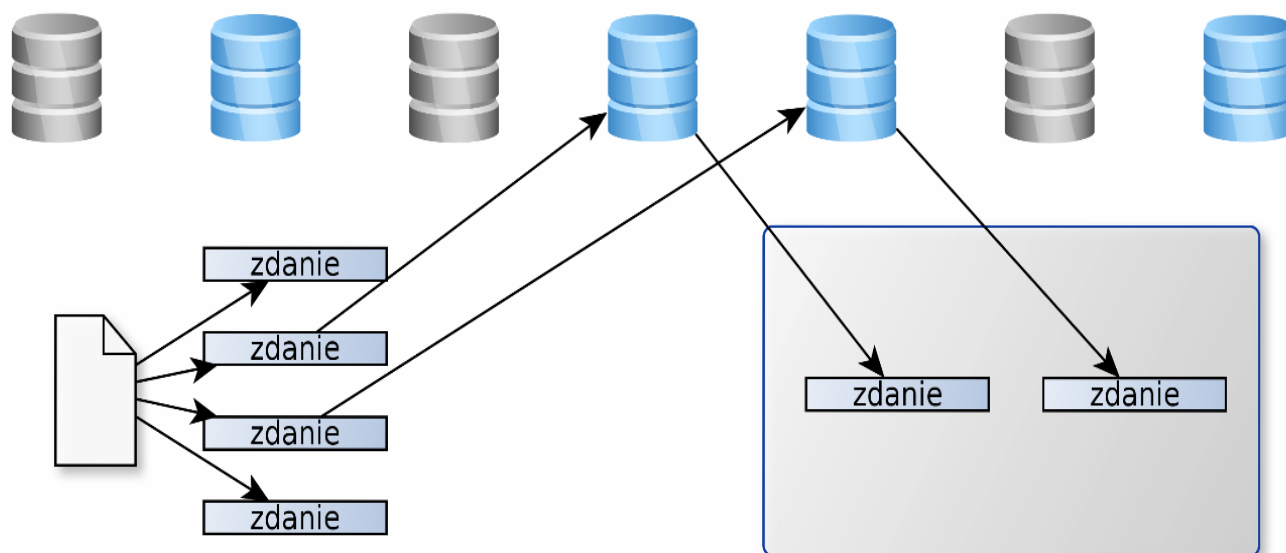
Internetowy antyplagiat

ZDANIA JAKO KANDYDACY!



Internetowy antyplagiat

WERYFIKACJA ZDANIA?



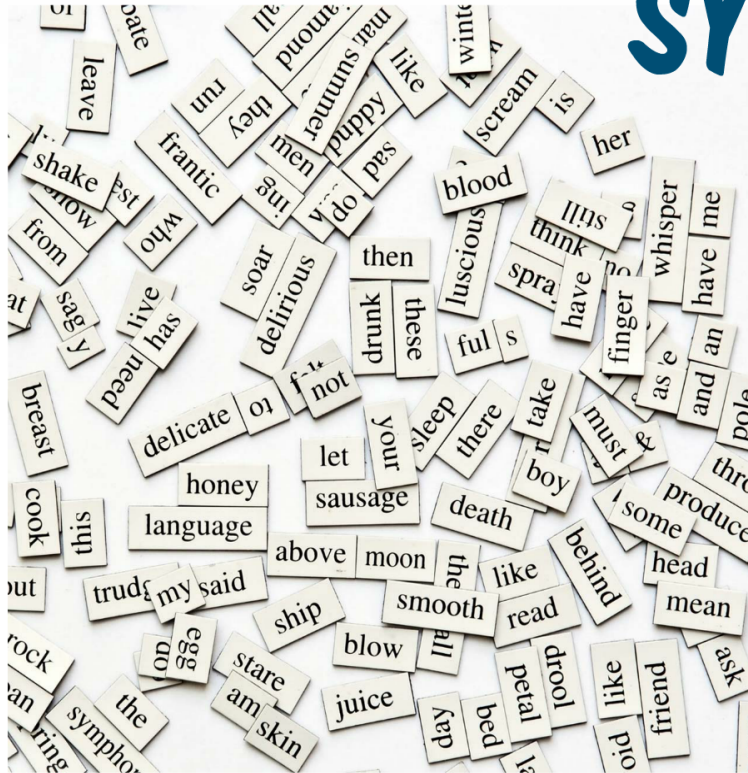
PRZETWARZANIE JĘZYKA

Walka z plagiatem to próba wykrycia metod jego ukrywania:

- **białe znaki**
- **podstawienia czcionek**
- **zmiana szyku zdań**
- **stosowanie synonimów**
- **przeformatowanie tekstu**

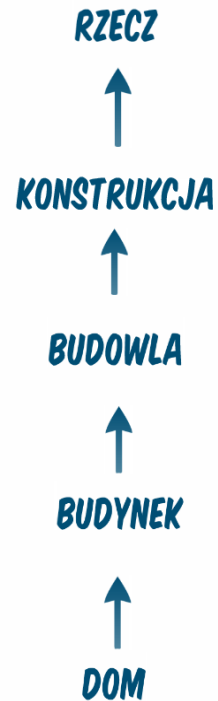
Synonymy

SYNONIMY



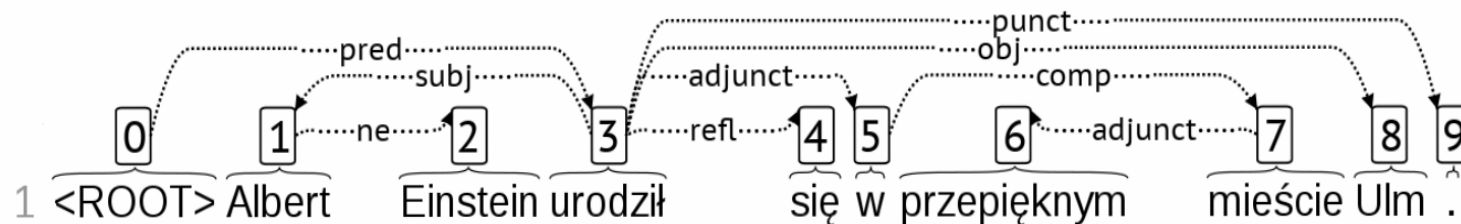
Hiperonimia

HIPERONIMY



Parser zależnościowy

MALT



1. Albert Einstein urodził się w przepięknym mieście Ulm .

Parser zależnościowy

KOREFERENCJA

KOREFERENCJA

Koreferencja występuje gdy dwa lub więcej wyrażen odnosi się do tej samej osoby lub rzeczy.

Komercyjne użycie



OSA

Otwarty System Antyplagiatowy

OSA - wersja lokalna

Dwa etapy:

- generowanie potencjalnych kandydatów (szybkie, ale możliwe fałszywe trafienia)
- dokładniejsze porównanie prac na ograniczonym zbiorze

Porównanie TFV

1. Przygotowanie tekstu wejściowego:
tokenizacja i lematyzacja
2. Podzielenie tekstu na fragmenty
o określonej liczbie słów
(przykład: 500 słów długość, 100 zakładki)
3. Zliczenie częstości słów dla fragmentów
4. Porównanie wektorów z bazą referencyjną

Mnożenie TFV

1. Dwa wektory częstości na wejściu
2. Dla każdego słowa, które znajduje się w obu zliczamy w którym z wektorów wystąpiło więcej razy

$$\frac{2 \cdot (equal + \max(left, right))}{equal + left + right} - 1$$

Mnożenie TFV

1. Dwa wektory częstości na wejściu
2. Dla każdego słowa, które znajduje się w obu bierzemy jego częstość i obliczamy:

$$\frac{\sum_i freq_L(i) \cdot freq_R(i) \cdot weight(i)}{\sqrt{(\sum_i freq_L(i)^2 \cdot weight(i)) \cdot (\sum_i freq_R(i)^2 \cdot weight(i))}}$$

Mnożenie TFV

1. Dwa wektory częstości na wejściu
2. Dla każdego słowa, które znajduje się w obu bierzemy jego częstość i obliczamy:

$$\frac{\sum_i \min(freq_L(i), freq_R(i))}{\min(\sum_i freq_L(i), \sum_i freq_R(i))}$$

Magiczna funkcja

Po wyliczeniu wcześniejszych miar używamy funkcji do obliczenia ostatecznego wyniku:

$$1 - \left(1 - \left(1 - 2 * \frac{\text{acos}(\text{angle})}{\pi} \right)^q \right)^{1/q}$$

Szczegółowe porównanie

- wykonujemy dla wcześniej ograniczonego zbioru dokumentów
- porównujemy “szkielety” dokumentów, czyli tablice lemmów słów
- zwykłe wyszukiwanie wspólnych podciągów