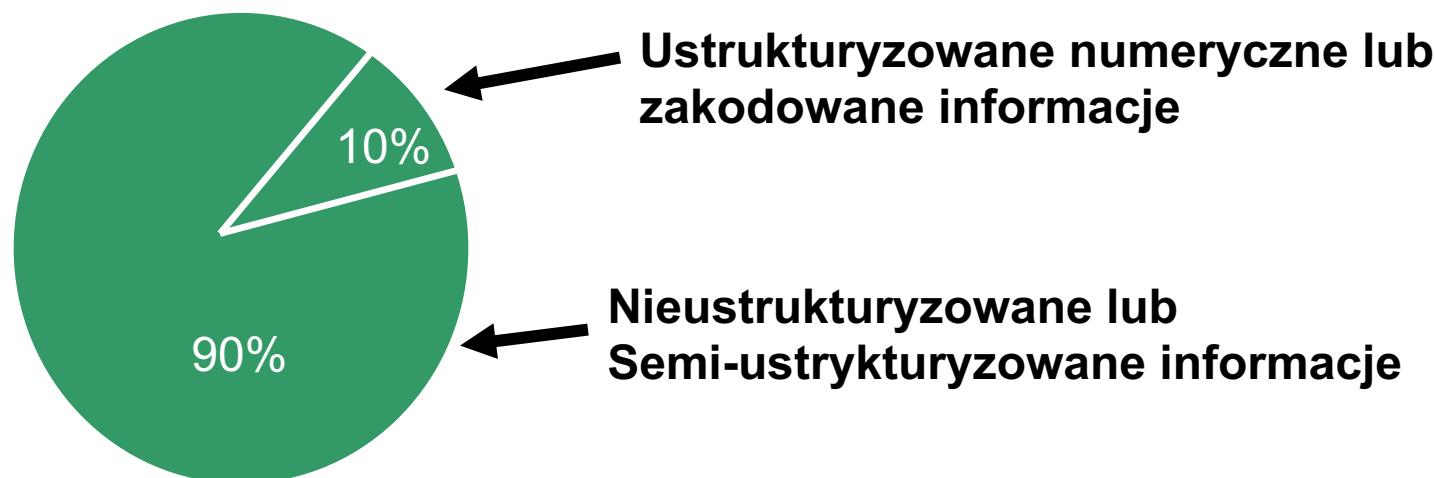


# Text Mining – trochę teorii na start ;)

Autorzy: Liliana Pięta, Rafał K. Wojdan

- Różne źródła podają, że ok 90% światowych zasobów stanowią dane niestrukturyzowane
- Procesy biznesowe w coraz większym stopniu bazują na analizach przepływających informacji, co wymaga przejścia z prostego wyszukiwania danych w dokumentach do odkrywania wiedzy w nich ukrytej



- Strony internetowe
- Emaile
- Dokumentacja techniczne
- Dokumenty korporacyjne
- Książki
- Elektroniczne biblioteki
- Reklamacje
- Fora internetowe
- Media społecznościowe
- Tytuły ofert - ecommerce

- Klasyfikacja news'ów, stron internetowych na podstawie ich treści
  - Kategoryzacja emaili
  - Organizacja repozytoriów zawierających metadane o dokumentach w celu ich przeszukiwania
  - Klastrowanie dokumentów lub stron internetowych
  - Systemy rekomendacyjne treści lub wykorzystujące dane tekstowe
  - NER (Named Entity Recognition) wykrywanie encji np. marki, kolory, typy ubrań
- <https://aclanthology.info/pdf/D/D11/D11-1144.pdf>

- Analiza sentymentu (Facebook, Twitter)
- Predykcja wartości akcji na giełdzie
- Predykcja churn'u
- Predykcja (podpowiedzi) kolejnego słowa
- Predykcja frazy do wyszukania (suggester)

- Bardzo duża liczba możliwych wymiarów (w dużej mierze rzadkich) – wszystkie słowa i frazy danego języka
- Pytanie o reprezentację danych - BOW vs CBOW?
- Skomplikowane i subtelne zależności pomiędzy pojęciami
- Dwuznaczność i zależność od kontekstu – Apple (firma) czy apple (owoc)
- Brak, a nawet zmienna struktura
- Normalizacja
- Ogromne bazy tekstów – jaką próbkę zastosować
- Bardzo zaszumione dane – błędy w pisowani

### **Ustrukturyzowane dane:**

Transakcje

Cechy klienta

Wyniki

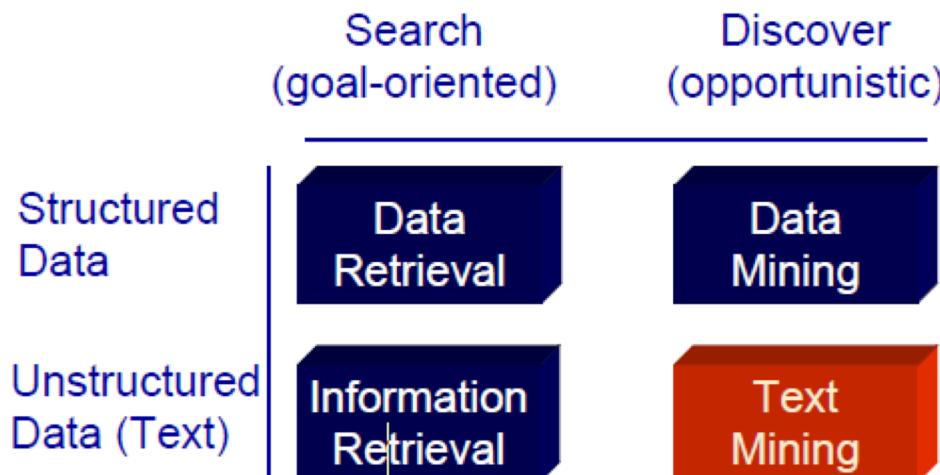
### **Semi-Ustrykturyzowane dane:**

Wszystkie “nierelacyjne” dane np. XML, JSON

### **Nieustrukturyzowane**

Wszelkie teksty

### "Search" versus "Discover"



# Przygotowanie danych

### Tokenizacja

- Przekształcenie strumienia znaków w słowa
- W języku angielskim spacja jest separatorem słów
- Wyrażenie mogą zawierać znaki specjalne, co wymaga uwzględnienia w regułach tokenizacji
- Jeden algorytm nie zadziała wszędzie np. spacje nie są dobrym pomysłem dla:  
Chińskiego, Japońskiego, Niemieckiego

### Normalizacja tekstu

- Zamiana wszystkich znaków na małe litery
- Oczyszczenie z literówek
- Poprawa błędów w pisowni
  - Algorytmy Soundex czy dystans edycyjny
  - Słowniki
  - Wykorzystanie części mowy i kontekstu by jak najlepiej zgadywać jakie to słowo

# Text analytics

## PARSOWANIE

### Określanie części mowy (POS)

- Przydatne w identyfikacji pojęć takich jak osoba, miejsce, tytuł
- Wiele często używanych słów jest wieloznaczna
- Reguły logiczne oraz rachunek prawdopodobieństwa pozwala usunąć tę wieloznaczność
- Znaczniki części mowy:
  - Oparte na regułach
  - Stochastyczne (Hidden Markov Models trenowane na zestawie oznaczonych słów)
  - Kombinacja dwóch powyższych np. „Brill” tagger

- Example of stochastic tagging
  - NNP VBZ VBN TO VB NR  
Secretariat is expected to race tomorrow
  - NNP VBZ VBN TO NN NR  
Secretariat is expected to race tomorrow
  - $P(NN|TO) = 0.00047$
  - $P(VB|TO) = 0.83 \rightarrow \text{"race" is most likely a verb}$

### **Sprowadzanie do podstawowej formy fleksyjnej (Stemming)**

- Ujednolica słowa np. czasowniki sprowadza do bezokolicznika
- „Zwija” formy mnogie do pojedynczej
- Normalizuje czasowniki pod względem formy czasowej
- Usuwa afiksy (przedrostki, przyrostki)

**Stemming** - sprowadzenie wyrazu do rdzenia poprzez usunięcie formy fleksyjnej

**Lematyzacja** - sprowadzenie grupy wyrazów do ich nieodmienionej formy np. bezokolicznik, mianownik (uwzględnia kontekst)

### Przykłady:

- 1) Better-> Lematyzacja: good
- 2) Walking -> Stemmer: Walk, usuwamy -ing; Lematyzacja: Walk, to bezokolicznik
- 3) Meeting - Stemmer: meet, Lematyzacja: meet, uwzględnia czy czasownik czy rzeczownik

**Stop words** – słowa występujące bardzo często, ale nie niosące ze sobą żadnej wartości informacyjnej np. i, że, a itd..

Usunięcie takich wyrazów zmniejsza liczbę wymiarów

**Podpowiedź ;)**

Stop words = Total words - Start words

### Słowniki i leksykony

- Bardzo przydatne, lecz czasochłonne
- Pozwalają ograniczyć zakres analizowanych słów
- Pozwalają skupić się tylko na wybranych atrybutach
- Rozwiążują problem synonimów, skrótów czy akronimów

**Bag of Words Model** – reprezentacja tekstu opisująca występowanie określonych wyrazów/grup wyrazów w dokumencie.

### 1. Zebranie danych

*It was the best of times,  
it was the worst of times,  
it was the age of wisdom,  
it was the age of foolishness.*

*Charles Dickens*

### 2. Stworzenie słownika

- “it”
- “was”
- “the”
- “best”
- “of”
- “times”
- “worst”
- “age”
- “wisdom”
- “foolishness”

Liczba dokumentów: 4  
Słownik: 10 słów  
Korpus: 24 słowa

*It was the best of times,*  
*it was the worst of times,*  
*it was the age of wisdom,*  
*it was the age of foolishness.*

### 3. Wektory dokumentów

Dokument 1

*It was the best of times,*

“it” = 1

“was” = 1

“the” = 1

“best” = 1

“of” = 1

“times” = 1

“worst” = 0

“age” = 0

“wisdom” = 0

“foolishness” = 0

*It was the best of times,*

*it was the worst of times,*

*it was the age of wisdom,*

*it was the age of foolishness.*

1 [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]

### 3. Wektory dokumentów

- 1 "it was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0]
- 2 "it was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
- 3 "it was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]

*It was the best of times,*  
*it was the worst of times,*  
*it was the age of wisdom,*  
*it was the age of foolishness.*

### 4. Term-document matrix

Słowo/dokument	Dokument 1	Dokument 2	Dokument 3	Dokument 4
it	1	1	1	1
was	1	1	1	1
the	1	1	1	1
best	1	0	0	0
of	1	1	1	1
times	1	1	0	0
worst	0	1	0	0
age	0	0	1	1
wisdom	0	0	1	0
foolishness	0	0	0	1

*It was the best of times,  
it was the worst of times,  
it was the age of wisdom,  
it was the age of foolishness.*

**Częstość występowania terminu w dokumencie (Term Frequency – TF)** – oznacza liczbę wystąpień danego terminu w ramach dokumentu

$$TF(t) = (\text{Liczba wystąpień słowa (t) w dokumencie}) / (\text{Łączna liczba słów w dokumencie}).$$

**Ogólne założenia:**

- Im częściej dane słowo występuje tym większe ma znaczenie
- Każde wystąpienie jest niezależnym zdarzeniem

**Częstość występowania terminu w dokumentach (Document Frequency – DF)** – oznacza liczbę dokumentów, w której występuje dany termin

### Ogólne założenia:

- Rzadziej występujące słowo w różnych dokumentach, jest bardziej związane z danym dokumentem i niesie większą wartość deskrytywną na temat jego treści
- Terminy występujące w wielu dokumentach to „pospolite” słowa

**Odwrotna częstotliwość występowania terminu w dokumentach (Inverse Document Frequency – IDF)** – bardziej intuicyjna interpretacja, im większe tym lepiej, zamiast im mniejsze tym lepiej

Zwykle stosuje się logarytm dla tej odwrotności, gdyż daje lepsze wyniki

$\text{IDF}(t) = \log_e(\text{Liczba dokumentów} / \text{Liczba dokumentów, w których występuje słowo } (t))$ .

### TF-IDF – połączenie obu

#### Interpretacja:

Rzadkie słowa o dużej częstotliwości są najprawdopodobniej istotnym pojęciem dla analizy danej kolekcji dokumentów

$$\text{TF-IDF} = \text{TF}(t) * \text{IDF}(t)$$

### TF matrix

Słowo/miara	Tf – d <sub>1</sub>	Tf – d <sub>2</sub>	Tf – d <sub>3</sub>	Tf – d <sub>4</sub>
it	1/6	1/6	1/6	1/6
was	1/6	1/6	1/6	1/6
the	1/6	1/6	1/6	1/6
best	1/6	0	0	0
of	1/6	1/6	1/6	1/6
times	1/6	1/6	0	0
worst	0	1/6	0	0
age	0	0	1/6	1/6
wisdom	0	0	1/6	0
foolishness	0	0	0	1/6

$$Tf(,,it'', d_1) = 1/6$$

$$Tf(,,it'', d_2) = 1/6$$

$$Tf(,,it'', d_3) = 1/6$$

$$Tf(,,it'', d_4) = 1/6$$

### IDF matrix

Słowo	IDF
it	0
was	0
the	0
best	0,602
of	0
times	0,301
worst	0,602
age	0,301
wisdom	0,602
foolishness	0,602

$$\text{idf}(“this”)=\log(4/4)=0$$

$$\text{idf}(“was”)=\log(4/4)=0$$

$$\text{idf}(“the”)=\log(4/4)=0$$

$$\text{idf}(“best”)=\log(4/1)=0,602$$

$$\text{idf}(“of”)=\log(4/4)=0$$

$$\text{idf}(“times”)=\log(4/2)=0,301$$

$$\text{idf}(“worst”)=\log(4/1)=0,602$$

$$\text{idf}(“age”)=\log(4/2)=0,301$$

$$\text{idf}(“wisdom”)=\log(4/1)=0,602$$

$$\text{idf}(“foolishness”)=\log(4/1)=0,602$$

*It was the best of times,  
it was the worst of times,  
it was the age of wisdom,  
it was the age of foolishness.*

### TF-IDF matrix

Słowo/miara	TF-IDF d1	TF-IDF d2	TF-IDF d3	TF-IDF d4
it	0	0	0	0
was	0	0	0	0
the	0	0	0	0
best	0,1003	0	0	0
of	0	0	0	0
times	0,0501	0,0501	0	0
worst	0	0,1003	0	0
age	0	0	0,0501	0,0501
wisdom	0	0	0,1003	0
foolishness	0	0	0	0,1003

$$\text{tf-idf}(“this”, d1) = 0,1666 * 0 = 0$$

$$\text{tf-idf}(“best”, d1) = 0,1666 * 0,602 = 0,1003$$

### Terminy wielowyrazowe (N-Gramy)

Stanowią kombinację słów

N – oznacza liczbę słów składowych wyrazów

Przykłady

2gram „vice president”

3gram „central intelligence agency”

4-gram „united states of america”

N-gramy można modelować wykorzystując prawdopodobieństwo warunkowe

### Progi dla:

**Tf** - np. min 10

**Idf** - np. min 4

**Tf-idf** - między 10,a 90 percylem

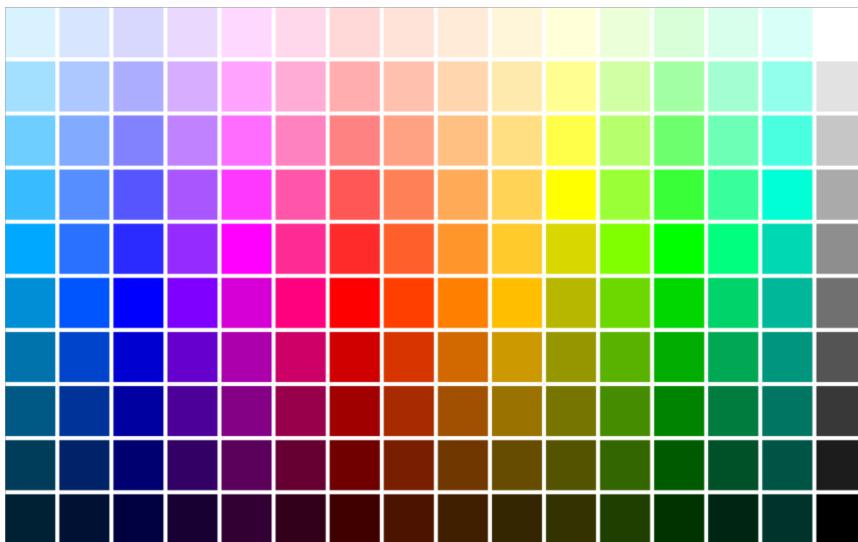
**Latent semantics analysis** – metoda bazująca na PCA i SVD, pozwalająca na podjęcie decyzji, ile pierwszych wymiarów w n-wymiarowej przestrzeni wystarczy do opisu jej wariancji (wymiary posortowane malejąco na podstawie wielkości wartości własnych)

Przyjmuje się, że 2-50 wymiarów powinno być wykorzystywane w celach klasteryzacji, natomiast 30-200 do klasyfikacji i predykcji (Sanders, 2004)

# Text analytics

## REDUKCJA WYMIARÓW - intuicja

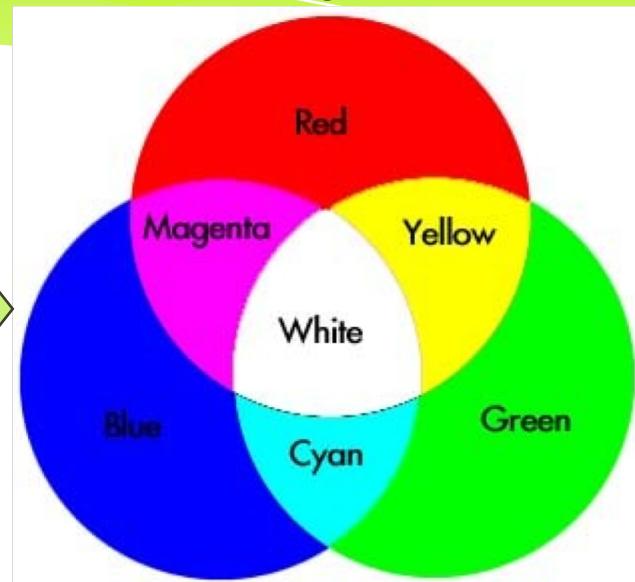
Wymiary:  
 $10 \times 16 = 160$



Wymiary:  
**12**



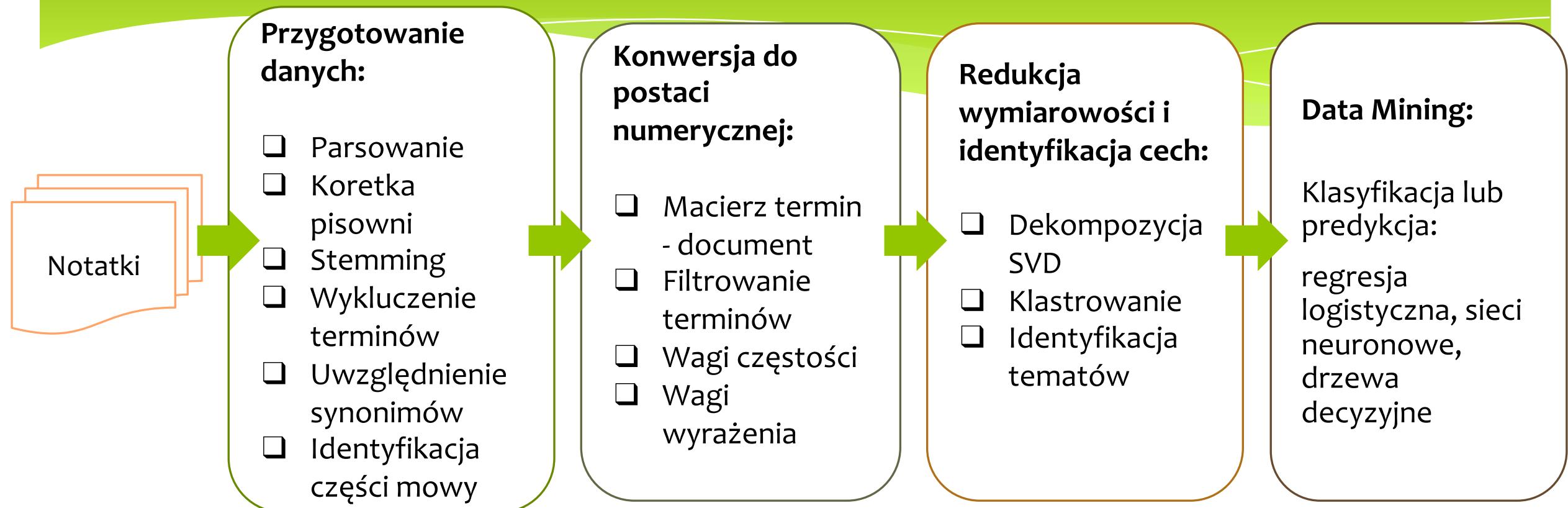
Wymiary:  
**3**



# Text mining flow in feature engineering

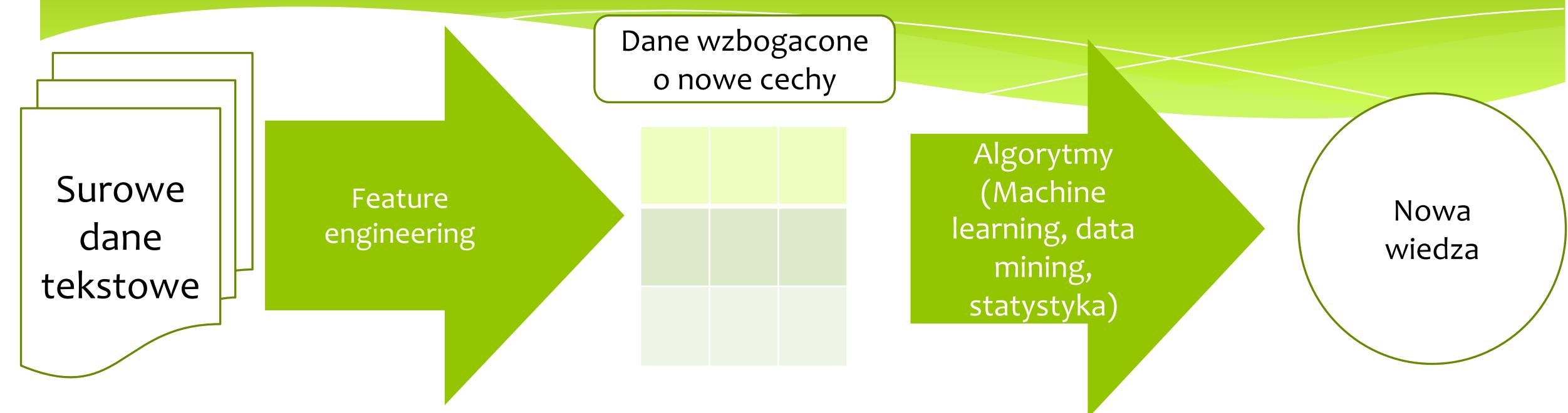
# Text mining flow

## Data flow



# Feature engineering

## MIEJSCE W PROCESIE ANALIZY DANYCH



# feature engineering

## METODY NADZOROWANE VS NIENADZOROWANE

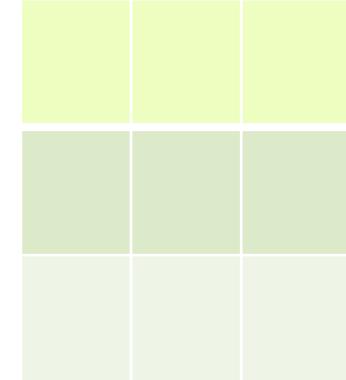
Feature engineering → kategoryzacja i ekstrakcja treści  
oraz text mining



Nadzorowany → Metody NLP,  
modele klasyfikujące

Dane wzbogacone  
o nowe cechy

Nienadzorowany → LSA, SVD,  
„tematyzacja”, klasteryzacja



# Przykłady

# Pos-tagging

## PRZYKŁAD

The screenshot shows a website for the Cognitive Computation Group. At the top, there is a navigation bar with links for News, Research, People, Software, Demos, Publications, Resources, and Schedule. Below the navigation bar, a dark header bar displays the word "Demo". The main content area features a title "Part of Speech Tagging Demo" and a subtitle "755,108 views". There are two buttons: "Download Software" and "About This Demo". A text box contains three paragraphs about security measures for the Super Bowl. At the bottom left, there is a "Submit" button.

COGNITIVE COMPUTATION GROUP

News Research ▾ People Software ▾ Demos Publications Resources ▾ Schedule ▾

Demo

Part of Speech Tagging Demo

755,108 views

If you wish to cite this work, please cite [this publication](#).

Helicopters will patrol the temporary no-fly zone around New Jersey's MetLife Stadium Sunday, with F-16s based in Atlantic City ready to be scrambled if an unauthorized aircraft does enter the restricted airspace.

Down below, bomb-sniffing dogs will patrol the trains and buses that are expected to take approximately 30,000 of the 80,000-plus spectators to Sunday's Super Bowl between the Denver Broncos and Seattle Seahawks.

The Transportation Security Administration said it has added about two dozen dogs to monitor passengers coming in and out of the airport around the Super Bowl.

Submit

[http://cogcomp.org/page/demo\\_view/pos](http://cogcomp.org/page/demo_view/pos)

# Pos-tagging

## PRZYKŁAD c.d

The Part-of-Speech tagger has automatically labeled the input in the following way.

NNPS/ Helicopters MD/ will NN/ patrol DT/ the JJ/ temporary JJ/ no-fly NN/ zone IN/ around NNP/ New NNP/ Jersey POS/ 's NNP/ MetLife NNP/ Stadium NNP/ Sunday ,/, IN/ with NNP/ F-16s VBN/ based IN/ in NNP/ Atlantic NNP/ City JJ/ ready TO/ to VB/ be VBN/ scrambled IN/ if DT/ an JJ/ unauthorized NN/ aircraft VBZ/ does VB/ enter DT/ the VBN/ restricted NN/ airspace ./ .

IN/ Down IN/ below ,/, JJ/ bomb-sniffing NNS/ dogs MD/ will NN/ patrol DT/ the NNS/ trains CC/ and NNS/ buses WDT/ that VB/ are VBN/ expected TO/ to VB/ take RB/ approximately CD/ 30,000 IN/ of DT/ the JJ/ 80,000-plus NNS/ spectators TO/ to NNP/ Sunday POS/ 's NNP/ Super NNP/ Bowl IN/ between DT/ the NNP/ Denver NNS/ Broncos CC/ and NNP/ Seattle NNP/ Seahawks ./ .

DT/ The NNP/ Transportation NNP/ Security NNP/ Administration VBD/ said PRP/ it VBZ/ has VBN/ added IN/ about CD/ two NN/ dozen NNS/ dogs TO/ to VB/ monitor NNS/ passengers VBG/ coming RP/ in CC/ and RP/ out IN/ of DT/ the NN/ airport IN/ around DT/ the NNP/ Super NNP/ Bowl ./ .

IN/ On NNP/ Saturday ,/, NNP/ TSA NNS/ agents VBD/ demonstrated WRB/ how DT/ the NNS/ dogs MD/ can VB/ sniff RP/ out JJ/ many JJ/ different NNS/ types IN/ of NNS/ explosives ./ . RB/ Once PRP/ they VBP/ do ,/, PRP/ they VBP/ 're VBN/ trained TO/ to VB/ sit RB/ rather IN/ than NN/ attack ,/, RB/ so RB/ as RB/ not TO/ to VB/ raise NN/ suspicion CC/ or VB/ create DT/ a NN/ panic ./ .

NNP/ TSA NN/ spokeswoman NNP/ Lisa NNP/ Farbstein VBD/ said DT/ the NNS/ dogs VBP/ undergo CD/ 12 NNS/ weeks IN/ of NN/ training ,/, WDT/ which VBZ/ costs IN/ about NN/ \$200,000 ,/, NN/ factoring IN/ in NN/ food ,/, NNS/ vehicles CC/ and NNS/ salaries IN/ for NNS/ trainers ./ .

NNS/ Dogs VBP/ have VBN/ been VBN/ used IN/ in NN/ cargo NNS/ areas IN/ for DT/ some NN/ time ,/, CC/ but VBP/ have RB/ just VBN/ been VBN/ introduced RB/ recently IN/ in NN/ passenger NNS/ areas IN/ at NNP/ Newark CC/ and NNP/ JFK NNS/ airports ./ . NNP/ JFK VBZ/ has CD/ one NN/ dog CC/ and NNP/ Newark VBZ/ has DT/ a NN/ handful ,/, NNP/ Farbstein VBD/ said ./ .

[http://cogcomp.org/page/demo\\_view/pos](http://cogcomp.org/page/demo_view/pos)

# Sentiment analysis

## PRZYKŁAD

 **Sentiment Analysis** | [Information](#) | [Live Demo](#) | [Sentiment Treebank](#) | [Help the Model](#) | [Source Code](#)

Please enter text to see its parses and sentiment prediction results:

This movie doesn't care about cleverness, wit or any other kind of intelligent humor.  
Those who find ugly meanings in beautiful things are corrupt without being charming.  
There are slow and repetitive parts, but it has just enough spice to keep it interesting.

You can also upload a file (limit 200 lines):    Show trees in binary form

---

[609 Comments](#) [Recursive Deep Models for Semantic Compositionality](#) [!\[\]\(3333269b15f7b52bb7877c37f51db396\_img.jpg\) Login](#) [Sort by Best](#)

 Recommend 69  Share

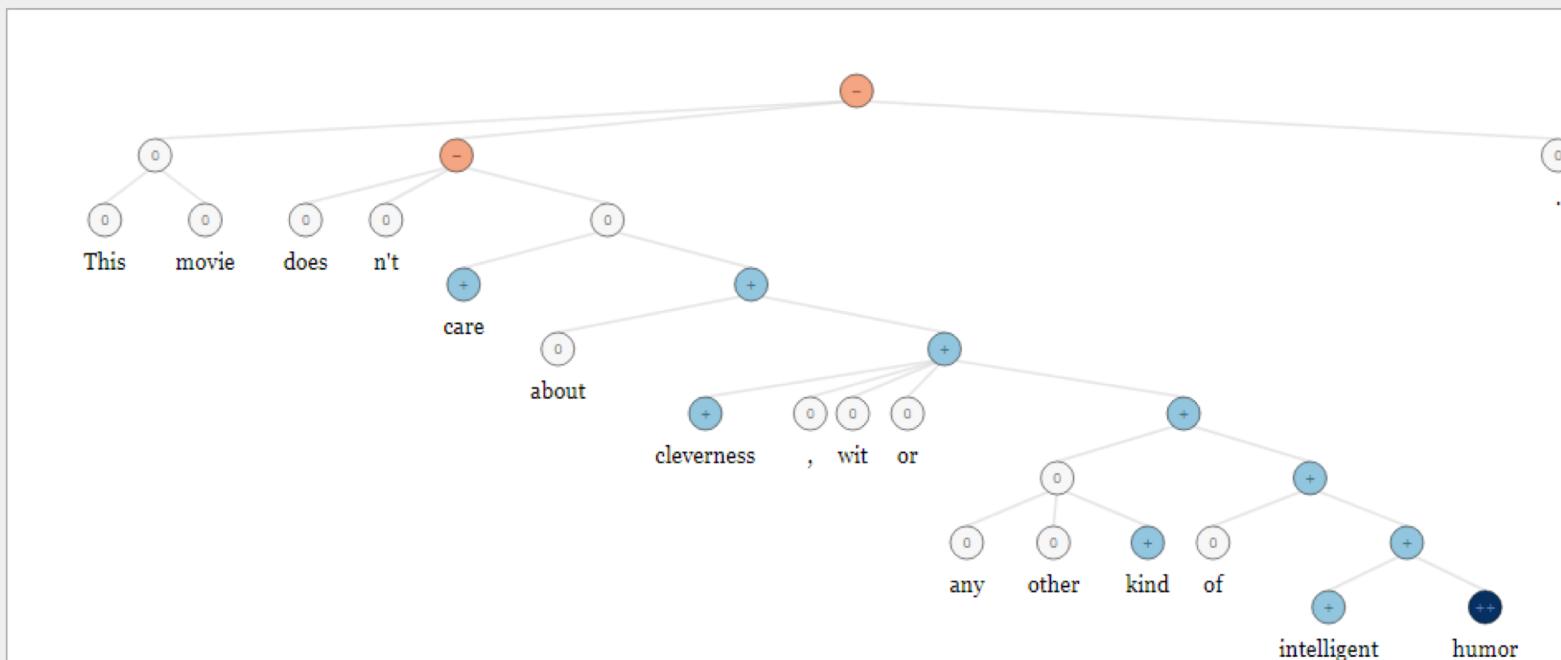
<http://nlp.stanford.edu:8080/sentiment/rntnDemo.html>

# Sentiment analysis

## PRZYKŁAD

### Sentiment Trees

You can double-click on each tree figure to see its expanded version with greater details. There are 5 classes of sentiment classification: **very negative**, **negative**, **neutral**, **positive**, and **very positive**.



<http://nlp.stanford.edu:8080/sentiment/rntnDemo.html>

Polski wordnet

<http://plwordnet.pwr.wroc.pl/wordnet/>

Polski stemmer

<https://github.com/morfologik/polimorfologik>