Przetwarzanie języka naturalnego/03

2024.03.21

Krzysztof Misztal misztal.edu.pl

Spis treści

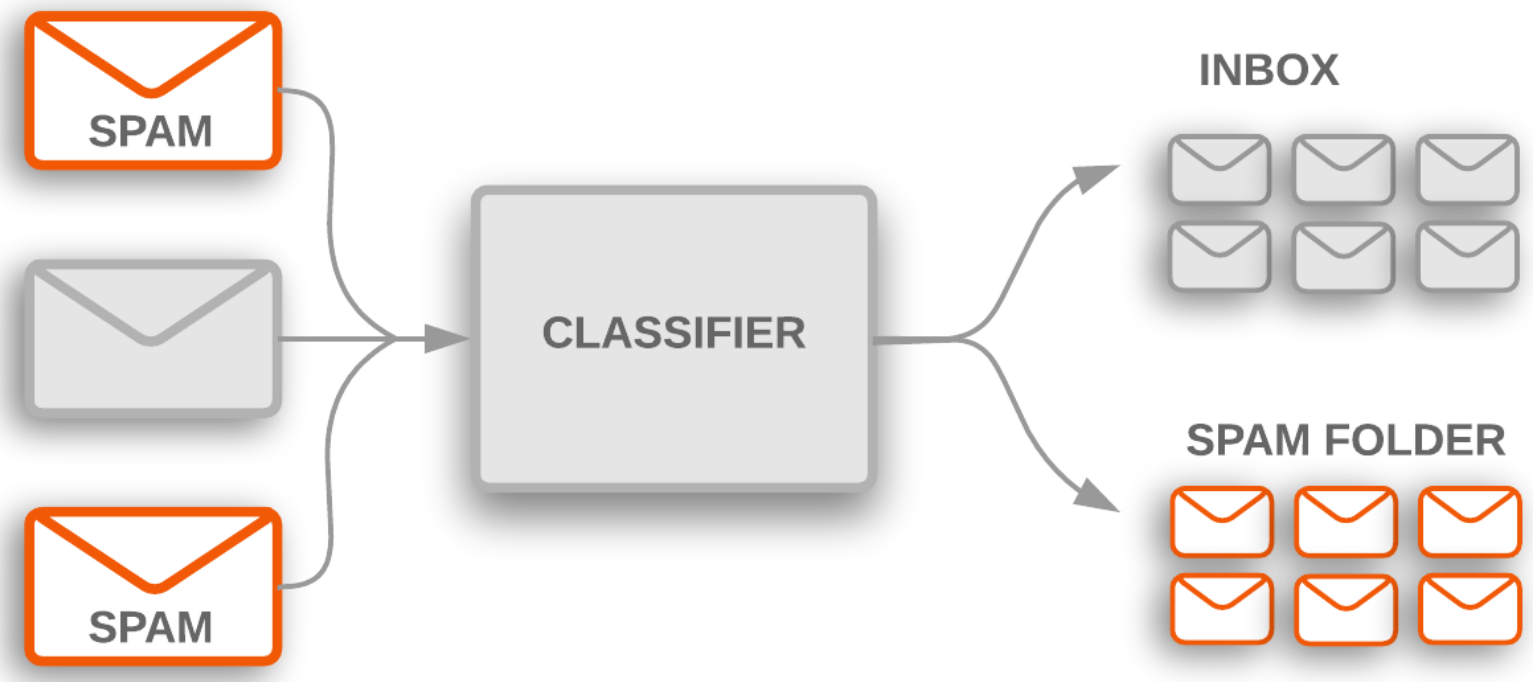
Spis treści

Klasyfikacja tekstu

Sformułowanie problemu

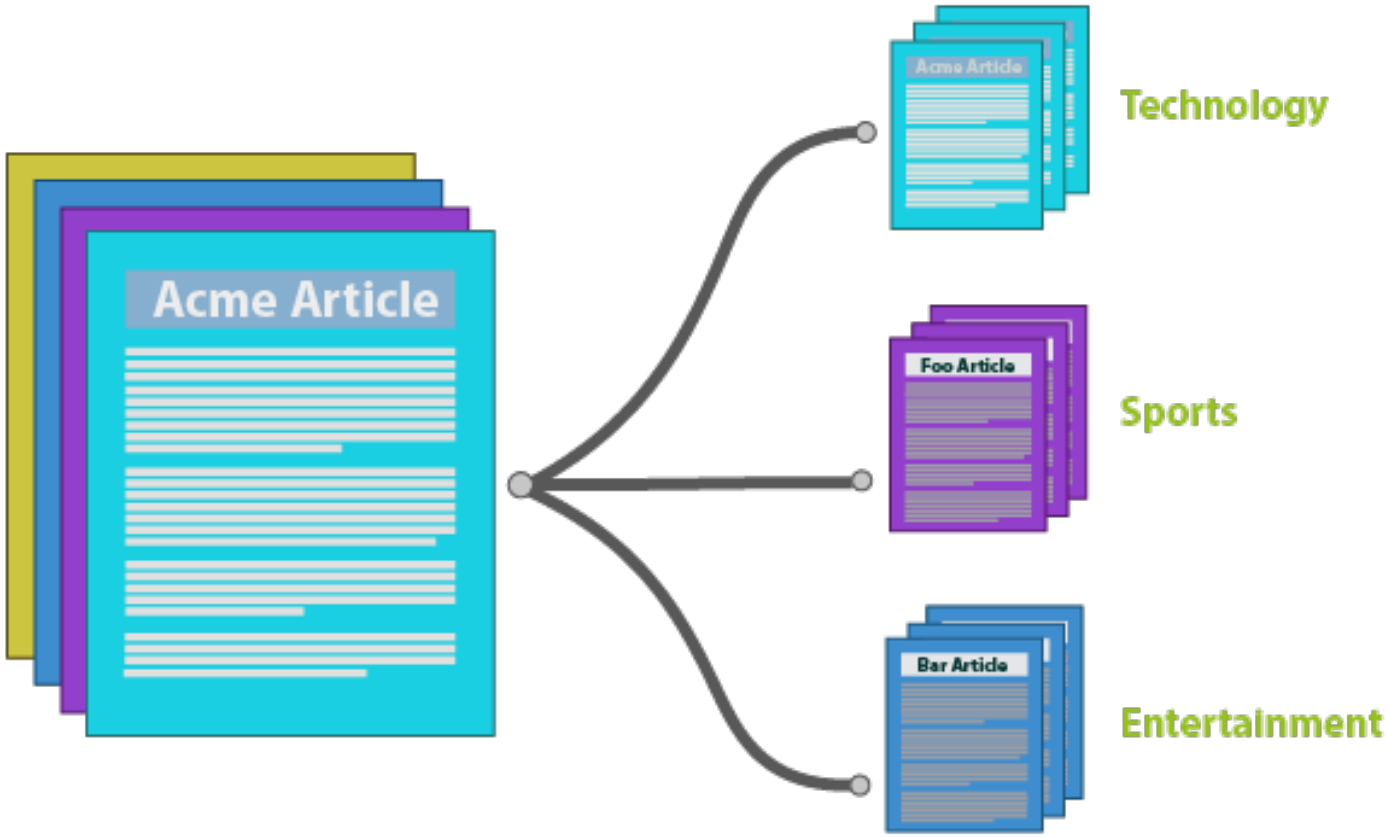
Dla zadanego dokumentu d oraz ustalonego zbioru klas $C = \{c_1, \ldots, c_k\}$ zwróć odpowiednia klasę $cl(d) \in C$.

Detekcja spamu



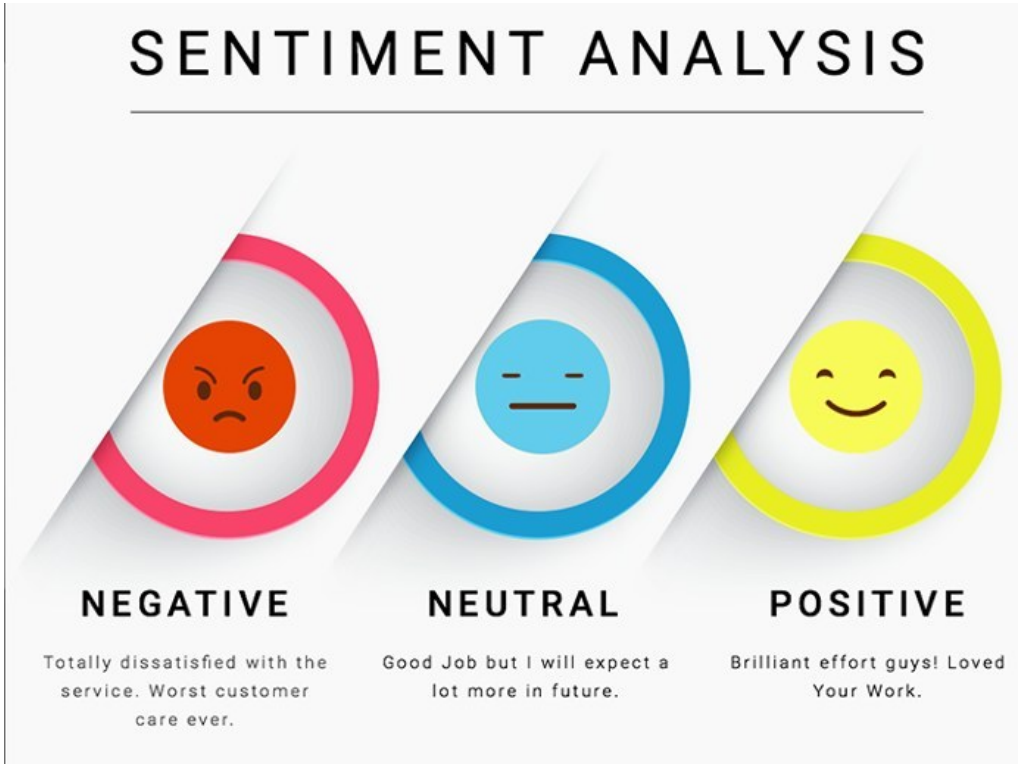
https://developers.google.com

Klasyfikacja artykułów



https://towardsdatascience.com

Analiza sentymentu



https://www.kdnuggets.com

Klasyfikacja tekstu - przykłady

- Analiza sentymentu
- Identyfikacja języka
- Identyfikacja autora
- Identyfikacja płci
- Detekcja spamu
- ...

Podejscie statystyczne (uczenie nadzorowane)

Wejście

- zbiór uczący $T = \{(d_i, c_i)\}_{i \in I} \subset D \times C$

Wyjście

- klasyfikator $cl : D \rightarrow C$

Podstawowa reprezentacja danych

Bag of words

Bag of Words


research
project
work
students
graduate
undergraduate
one
class
astronomy
student
extremely
first
school
data
level
several
working
program
telescope
time
high
year
skills
began
already
radio
However
undergraduates
department
ability
think
also
excellent
physics
summer
within
much
problem
since
galaxies
new
course
course
second
scientific
truly
application
exceptional
images
now
many
two
remarkable
support
past
courses
strong
faculty
met
years
field
known
initiative
took
stellar
make
drive
best
highly
intellectual
possible
interesting
carried
without
leadership
always
researcher
gone
given

Bag of Words

Bag of words (BoW)

Very good drama although it appeared to have a few blank areas leaving the viewers to fill in the action for themselves. I can imagine life being this way for someone who can neither read nor write. This film simply smacked of the real world: the wife who is suddenly the sole supporter, the live-in relatives and their quarrels, the troubled child who gets knocked up and then, typically, drops out of school, a jackass husband who takes the nest egg and buys beer with it. 2 thumbs up… very very very good movie.

('the', 8),
(',', 5),
('very', 4),
('.', 4),
('who', 4),
('and', 3),
('good', 2),
('it', 2),
('to', 2),
('a', 2),
('for', 2),
('can', 2),
('this', 2),
('of', 2),
('drama', 1),
('although', 1),
('appeared', 1),
('have', 1),
('few', 1),
('blank', 1)
…..

https://medium.com

Spis treści

Naive Bayes

$$cl(d) = \arg\max_{c \in C} P(c|d)$$

$$P(c|d)$$

- c – to decyzja dla danego dokumentu
- d – to dokument, czyli zbiór specjalnej konfiguracji cech
- Oszacowanie tego pradwopodobieństwa jest trudne, bo wymaga obliczenia wartość dla koniunkcji cech – w rzeczywistości nie jest to obserwowalne w danych testowych nigdy lub prawie nigdy (bo implikowałoby to obecność w danych treningowych obiektu identycznego z testowym, co zdarza się bardzo rzadko).

Naive Bayes

$$cl(d) = \arg\max_{c \in C} P(c|d)$$

Naive Bayes

$$cl(d) = \arg\max_{c \in C} P(c|d)$$

z Twierdzenia Bayesa wiemy, że

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Naive Bayes

$$
\begin{aligned}
cl(d) &= \arg\max_{c \in C} P(c|d) \\
&= \arg\max_{c \in C} \frac{P(d|c)P(c)}{P(d)} \\
&= \arg\max_{c \in C} P(d|c)P(c)
\end{aligned}
$$

$P(d)$ zostało pominięte ponieważ jest ono stałe dla tego wyrażenia, tzn. tak naprawdę porównujemy $\left\{ \frac{P(d|c_1)P(c_1)}{P(d)}, \ldots, \frac{P(d|c_k)P(c_k)}{P(d)} \right\}$

Naive Bayes

Każdy dokument d reprezentujemy jako zbiór jego cech, tzn.

$$\phi(d) = (f_1, \ldots, f_{k_d}),$$

np. przez słowa w nim zawarte, tj. $\phi(d) = (w : w \in d)$, zatem dostajemy

$$cl(d) = \arg\max_{c \in C} P(f_1, \ldots, f_{k_d}|c)P(c)$$

ale pojawia się problem z oszacowaniem $P(f_1, \ldots, f_{k_d}|c)$, bo mamy $\prod_i |F_i||C|$ możliwych zdarzeń, gdzie $F_i = \{\phi_i(d) : d \in D\}$.

Naive Bayes – założenie

Niezależność zmiennych – "Naive assumption"

Załóżmy, ze prawdopodobieństwo obserwacji poszczególnych cech są niezależne przy danej klasie c

$$P(f_1, \ldots, f_{k_d}|c) = P(f_1|c)P(f_2|c)\ldots P(f_{k_d}|c) = \prod_i^{k_d} P(f_i|c)$$

Naive Bayes – forma końcowa

$$cl(d) = \arg\max_{c \in C} P(c) \prod_{f \in \phi(d)} P(f|c)$$

Naive Bayes – forma końcowa [NLP - Bag of Words]

$$cl_{NB}(d) = \arg\max_{c \in C} P(c) \prod_{w_i \in d} P(w_i|c)$$

Naszymi cechami mogą być wystąpienia słów z Bag of words.

Spis treści

Naive Bayes - uczenie

Skąd wziąć $P(w_i|c)$ oraz $P(c)$?

Naive Bayes – $P(c)$

Mamy skończenie wiele klas (zazwyczaj bardzo mało), wiec można przyjąć estymator największej wiarygodności (MLE)

$$\hat{P}(c) = \frac{|\{(d_j, c_j) \in T : c_j = c\}|}{|T|}$$

Naive Bayes – $P(c)$

Analogicznie

$$\hat{P}(w_i|c) = \frac{count(w_i, c)}{\sum_w count(w, c)}$$

Naive Bayes MLE

$$cl_{NB}(d) = \arg\max_{c \in C} P(c) \prod_{w_i \in d} P(w_i|c)$$

$$cl_{NB_{MLE}}(d) = \arg\max_{c \in C} \frac{|\{(d_j, c_j) \in T : c_j = c\}|}{|T|} \prod_{w_i \in d} \frac{count(w_i, c)}{\sum_w count(w, c)}$$

Do własnego obejrzenia

https://www.dailymotion.com/video/x2x5u4o
https:
//www.youtube.com/watch?gl=SN&hl=fr&v=1KOUpkU2cME

Naive Bayes MLE – przykład

Pozytywne

- I like dogs
- John like dogs

Negatywne

- I hate dogs
- John hate dogs

Naive Bayes MLE – przykład

Pozytywne

- I like dogs
- John like dogs

Negatywne

- I hate dogs
- John hate dogs

$$\hat{P}(positive) = \hat{P}(negative) = 0.5$$

$$cl_{NB_{MLE}}(\text{"I like John"}) = positive$$

$$cl_{NB_{MLE}}(\text{"I hate John"}) = negative$$

Naive Bayes - fail case

Co jeśli natkniemy się na takie słowo $w \in d$, że dla, poprawnej dla d, klasy c, że $\forall(d_i, c) \in T : w \notin d_i$?

Naive Bayes - fail case

Co jeśli natkniemy się na takie słowo $w \in d$, że dla, poprawnej dla d, klasy c, że $\forall(d_i, c) \in T : w \notin d_i$?

$$cl_{NB_{MLE}}(\text{"I like cats"}) = ???$$

α-Laplacian smoothing

$$\hat{P}_\alpha(w_i|c) = \frac{count(w_i, c) + \alpha}{\sum_{w \in W}(count(w, c) + \alpha)}$$

$$\frac{count(w_i, c) + \alpha}{\sum_{w \in W} count(w, c) + \alpha|W|}$$

Naive Bayes - fail case

positive

- I like

negative

- I hate

$$\hat{P}_1(positive|\text{"I like dogs"}) =$$

$$= \hat{P}_1(positive)\hat{P}_1(\text{I}|positive)\hat{P}_1(\text{like}|positive)\hat{P}_1(\text{dogs}|positive)$$

$$= \frac{1}{2}\frac{1+1}{2+3}\frac{1+1}{2+3}\frac{0+1}{2+3} = \frac{4}{250} = 0.016$$

$$\hat{P}_1(negative|\text{"I like dogs"}) =$$

$$= \hat{P}_1(negative)\hat{P}_1(\text{I}|negative)\hat{P}_1(\text{like}|negative)\hat{P}_1(\text{dogs}|negative)$$

$$= \frac{1}{2}\frac{1+1}{2+3}\frac{0+1}{2+3}\frac{0+1}{2+3} = \frac{2}{250} = 0.008$$

$$cl_{NB_{MLE}}(\text{"I like dogs"}) = positive$$

Naive Bayes - fail case 2

Dla dużych zbiorów danych $\hat{P}(w_i|c) \rightarrow 0$, więc przy ograniczonej dokładności obliczeń $\hat{P}(c|d) = 0$ dla każdego c.

Log-likelihood

$$
\begin{aligned}
\arg\max_{c \in C} P(c|d) &= \arg\max_{c \in C} \log P(c|d) = \\
&= \arg\max_{c \in C} \log \left(P(c) \prod P(f_i|c) \right) = \\
&= \arg\max_{c \in C} \left[\log(P(c)) + \sum \log(P(f_i|c)) \right]
\end{aligned}
$$

W praktyce

- Stwórz zbiór uczący T oraz wybierz $\alpha \in (0, \infty)$
- Dla każdej klasy c_i
  - Policz $\hat{P}(c_i)$ z MLE
  - Skonkatenuj wszystkie dokumenty d, takie że $(d|c_i) \in T$ i nazwij d_{c_i}
  - Dla super-dokumentu d_{c_i} policz wygładzone estymatory dla kazdego słowa tego dokumentu i przypisz ich logarytmy do odpowiednich $log(\hat{P}_\alpha(w_i|c_i))$
- Przy klasyfikacji $d = (w_1, ..., w_k)$ wybierz klasę c która maksymalizuje $log(\hat{P}(c)) + \sum_{i=1}^{k} log(\hat{P}_\alpha(w_i|c))$

W praktyce

Pozytywne

- I like dogs
- John like dogs

Negatywne

- I hate dogs
- John hate dogs

	I	John	like	hate	dogs
positive	-1.25	-1.95	-1.25	-1.95	-0.85
negative	-1.25	-1.95	-1.95	-1.25	-0.85

$$\log(\hat{P}(positive)) = \log(\hat{P}(negative)) = -0.69$$

W praktyce

	I	John	like	hate	dogs	
positive	-1.25	-1.95	-1.25	-1.95	-0.85	-1.95
negative	-1.25	-1.95	-1.95	-1.25	-0.85	-1.95

$$\log(\hat{P}(positive)) = \log(\hat{P}(negative)) = -0.69$$

$$\log(\hat{P}(\text{"I like cats"}|positive)) =$$

$$\log(\hat{P}(\text{"I"}|positive))+\log(\hat{P}(\text{"like"}|positive))+\log(\hat{P}(\text{"cats"}|positive))$$

$$(-1.25) + (-1.25) + (-1.95) = -4.45$$

$$\log(\hat{P}(\text{"I like cats"}|negative)) =$$

$$\log(\hat{P}(\text{"I"}|negative))+\log(\hat{P}(\text{"like"}|negative))+\log(\hat{P}(\text{"cats"}|negative))$$

$$(-1.25) + (-1.95) + (-1.95) = -5.15$$

W praktyce

$\log(\hat{P}(\textit{positive})) = -0.69$
$\log(\hat{P}(\text{"I like cats"}|\textit{positive})) = -4.45$
$\log(\hat{P}(\textit{negative})) = -0.69$
$\log(\hat{P}(\text{"I like cats"}|\textit{negative})) = -5.15$

$$-4.45 - 0.69 > -5.15 - 0.69$$

więc

$$\textit{cl}(\text{"I like cats"}) = \textit{positive}$$

Spis treści

Naive Bayes – podsumowanie

$$cl(d) = cl(f) = \text{argmax}\, P(c) \prod P(f_i|c)$$

gdzie $\phi(d) = f = (f_1, \ldots, f_k)$

- **założenie Bag of words** – pozycja "ważnych" słów w tekście nie ma znaczenia
- **założenie warunkowej niezależności** – cechy f_i niezależnie wskazują klasę c

Klasyfikacja binarna

Założenie: $|C| = 2$

$$cl(f) = \operatorname*{argmax}_{c \in C} P(c) \prod P(f_i|c)$$

$$cl(f) = 1 \Leftrightarrow P(c_1) \prod P(f_i|c_1) > P(c_{-1}) \prod P(f_i|c_{-1})$$

$$\Leftrightarrow \frac{P(c_1) \prod P(f_i|c_1)}{P(c_{-1}) \prod P(f_i|c_{-1})} > 1$$

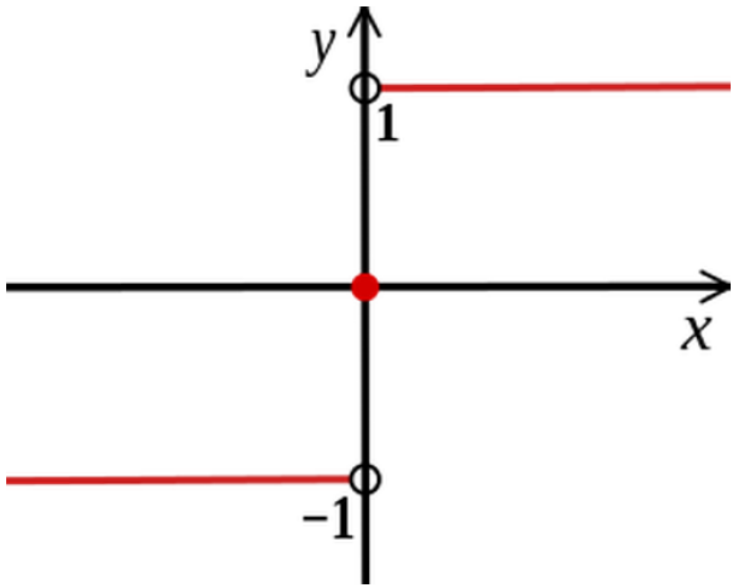
$$\Leftrightarrow \frac{P(c_1)}{P(c_{-1})} \times \frac{\prod P(f_i|c_1)}{\prod P(f_i|c_{-1})} > 1$$

$$\Leftrightarrow \log \left(\frac{P(c_1)}{P(c_{-1})} \times \frac{\prod P(f_i|c_1)}{\prod P(f_i|c_{-1})} \right) > \log(1)$$

$$\Leftrightarrow \log \left(\frac{P(c_1)}{P(c_{-1})} \right) + \sum \log \left(\frac{P(f_i|c_1)}{P(f_i|c_{-1})} \right) > 0$$

NB w przypadku binarnym

$$cl(f) = \operatorname{sgn}\left[\log\left(\frac{P(c_1)}{P(c_{-1})}\right) + \sum \log\left(\frac{P(f_i|c_1)}{P(f_i|c_{-1})}\right)\right]$$



Naive Bayes - podsumowanie

- Bardzo szybki algorytm $O(k|C|)$
- Możliwe łatwe douczanie $O(k)$
- Dobrze radzi sobie z nieistotnymi cechami
- Jeśli cechy sa na prawdę niezależne jest to **optymalny** klasyfikator w sensie Bayesowskim
- Dobry punkt wyjścia dla klasyfikacji tekstu ze względu na stosunek trudność/szybkość/jakość

Spis treści

Jakość modelu

Jak ocenić jakość naszego klasyfikatora?

Jakość modelu

Wynik klasyfikacji nie zawsze jest zgodny z oczekiwaniem:

- **błąd I rodzaju** (ang. False Positive, False Accept)
- **błąd II rodzaju** (ang. False Negative, False Reject)

decyzja/sytuacja	prawda: positive	prawda:negative
model: positive	TP (decyzja poprawna)	FP (błąd I rodzaju)
model: negative	FN (błąd II rodzaju)	TN (decyzja poprawna)

Tabela: Rodzaje błędów klasyfikatora.

Jakość modelu

W ekstrakcji informacji

- positive = token zawiera dana informacje
- negative = token nie zawierający informacji

np. detekcja nazw własnych w tekście

- positive = nazwa własna
- negative = każdy inny tekst

Miary jakości

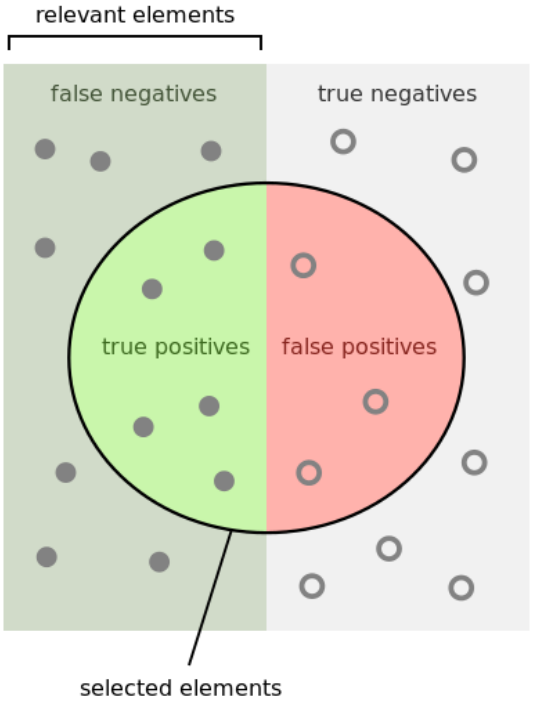
- precyzja (precision)

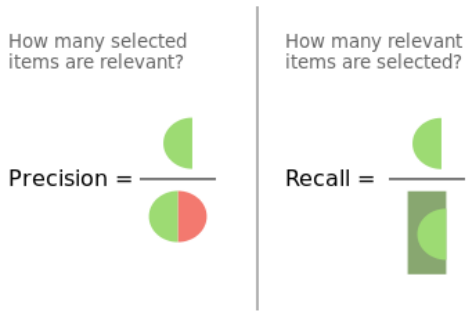
$$\frac{TP}{TP + FP}$$

- czułość (recall)

$$\frac{TP}{TP + FN}$$

`https://pl.wikipedia.org/wiki/Tablica_pomy\OT4\lek`





Do własnego obejrzenia

https://sebastianraschka.com/Articles/2014_naive_bayes_1.html
https://towardsdatascience.com/implementing-a-naive-bayes-classifier-for-text-categorizat

Dziękuję za uwagę.