Zastosowanie sztucznych sieci neuronowych w informatyce

Wykład 4. Klasyfikacja danych tekstowych

dr inż. Katarzyna Poczęta

Politechnika Świętokrzyska
Wydział Elektrotechniki Automatyki i Informatyki

Kielce, 2024

Plan wykładu

1. Dane tekstowe
   - Tokenizacja danych tekstowych
   - Wektoryzacja danych tekstowych
   - Redukcja wymiarowości
2. Klasyfikacja danych tekstowych
   - Struktura i uczenie
   - Testowanie

Dane wejściowe

1. Why learn perl python ruby if the company is using c c or java as the application language i wonder why would a c c java developer want to learn a dynamic language...
2. How do i connect to a database and loop over a recordset in c whats the simplest way to connect and query a database for a set of records in c
3. Codility absolute thistinct count from an array so i took the codility interview test yesterday and was informed today that i failed unfortunately i wasnt given any other information...

Dane wejściowe

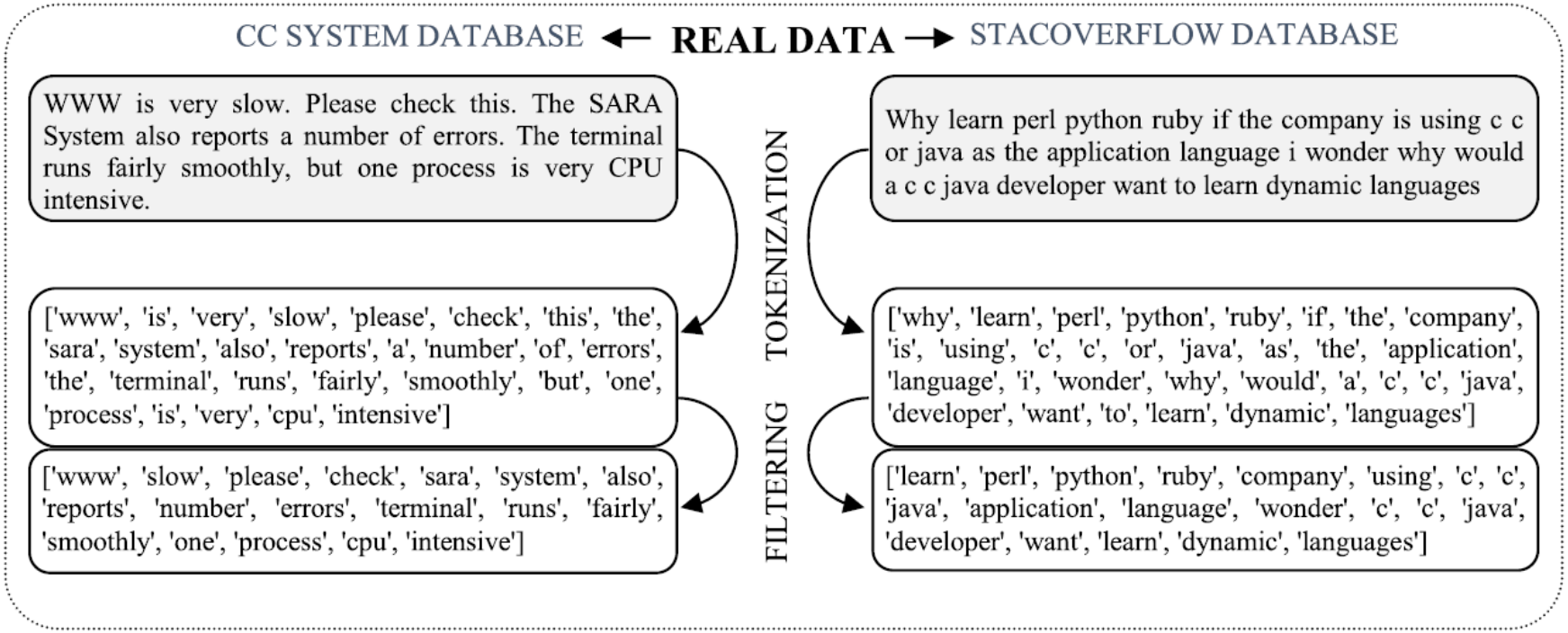
From: hamachi@adobe.com (Gordon Hamachi)
Subject: Re: Mercury Villager Minivan -- good buy?
Organization: Adobe Systems Incorporated
Distribution: usa
Lines: 38

I came across this interesting information in my local public library while researching minivans. It is the dealer price and the retail price for a minivan I am thinking about purchasing. Someone told me that the number for base price was slightly lower than the current price, but this should still give you some idea about pricing and how much you can negotiate.

Mercury Villager GS

	Dealer	Retail
Base Price	$14688	16504
Air Conditioning	729	857
Rear Defroster	143	168
Calif. Emissions	87	102
7 Passenger Seating	282	332
AM/FM Radio (no cassette)	STD	STD
Automatic Transmission	STD	STD
Anti-lock brakes	STD	STD
Destination	540	540

Tokenizacja danych tekstowych



Tokenizacja danych tekstowych

```
Organization: Arizona State University
From: Eric Davis <ICEND@ASUACAD.BITNET>
Subject: Re: HELP - 3DS
Lines: 11

In article <C70zv4.9Hq@ddtopper.Dundee.NCR.COM>, stephenc says:
>
>In 3D Studio, is there any way to create refraction, diffraction etc ?
>
>I want to simulate such things as glass lenses, bottles etc.

There might be an IPAS routine that does that,but I can'r be sure.  Another
way to do it is to render the scene without the glass object and save the
image.  Then assign that image to your glass object as a reflection.  It will
take a lot of adjusting for position and size of the reflection, but that's
the only thing I can think of.
```

Tokeny: [’organization’, ’arizona’, ’state’, ’university’, ’eric’, ’davis’, ’icend’, ’asuacad’, ’bitnet’, ’subject’, ’re’, ’help’, ’3ds’, ’lines’, ’11’, ’article’, ’c70zv4’, ’9hq’, ’ddtopper’, ’dundee’, ’ncr’, ’com’, ’stephenc’, ’says’, ’3d’, ’studio’, ’way’, ’create’, ...]

Unigramy, bigramy

'state', 'state edu', 'state university', 'writes', 'writes article', 'world', 'world nntp', 'world organization'

Stemming

Polega na uproszczeniu wybranego wyrazu do części odpornej na odmianę (przez przyimki, rodzaje itp.) - tzw. "rdzenia".

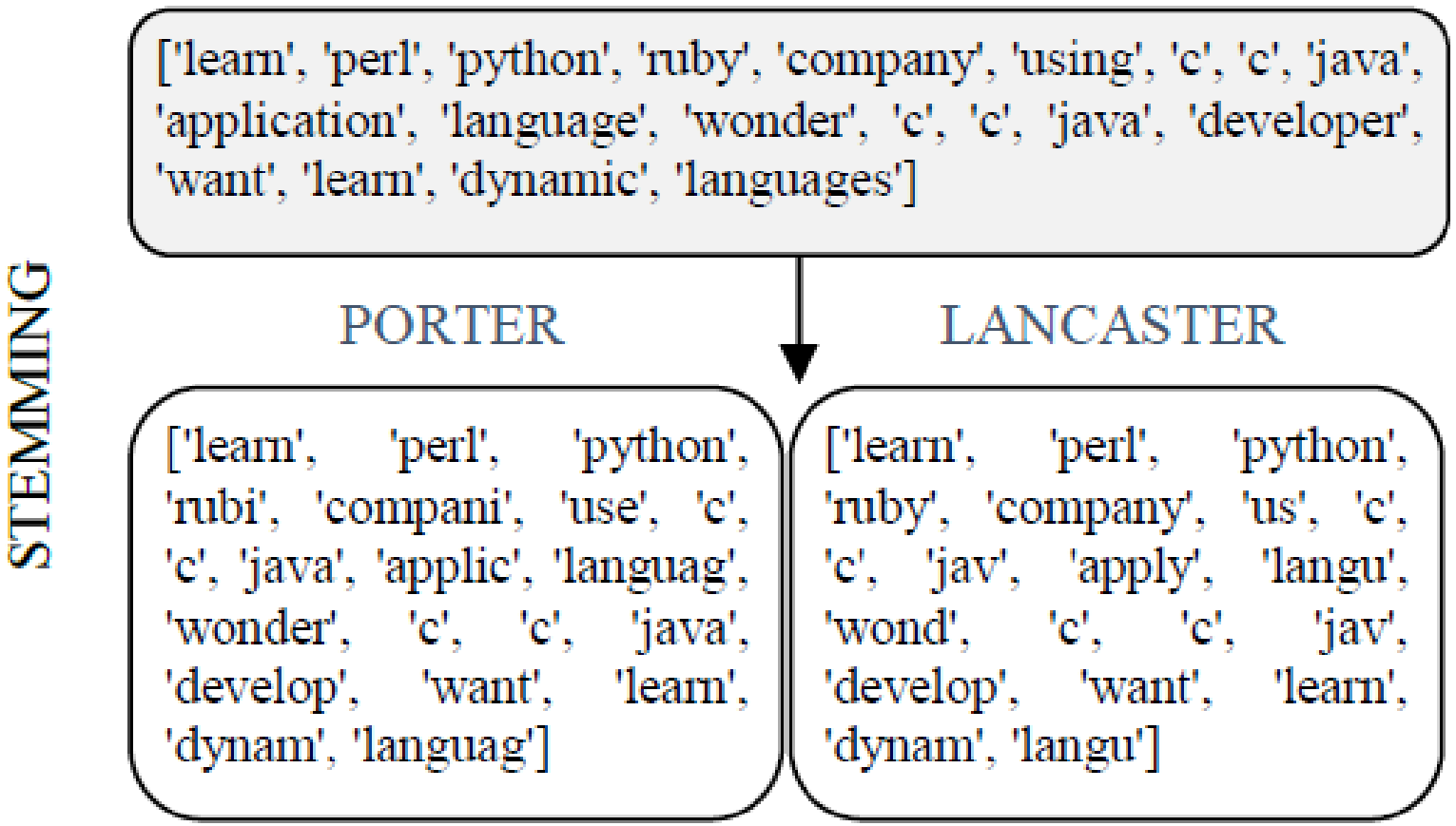
Algorytmy:

- algorytm Portera,
- algorytm Lancaster.

Biblioteki:

- pystempel (język polski),
- nltk.stem (język angielski).

Stemming


['learn', 'perl', 'python', 'ruby', 'company', 'using', 'c', 'c', 'java', 'application', 'language', 'wonder', 'c', 'c', 'java', 'developer', 'want', 'learn', 'dynamic', 'languages']
STEMMING
PORTER
LANCASTER
['learn', 'perl', 'python', 'rubi', 'compani', 'use', 'c', 'c', 'java', 'applic', 'languag', 'wonder', 'c', 'c', 'java', 'develop', 'want', 'learn', 'dynam', 'languag']
['learn', 'perl', 'python', 'ruby', 'company', 'us', 'c', 'c', 'jav', 'apply', 'langu', 'wond', 'c', 'c', 'jav', 'develop', 'want', 'learn', 'dynam', 'langu']

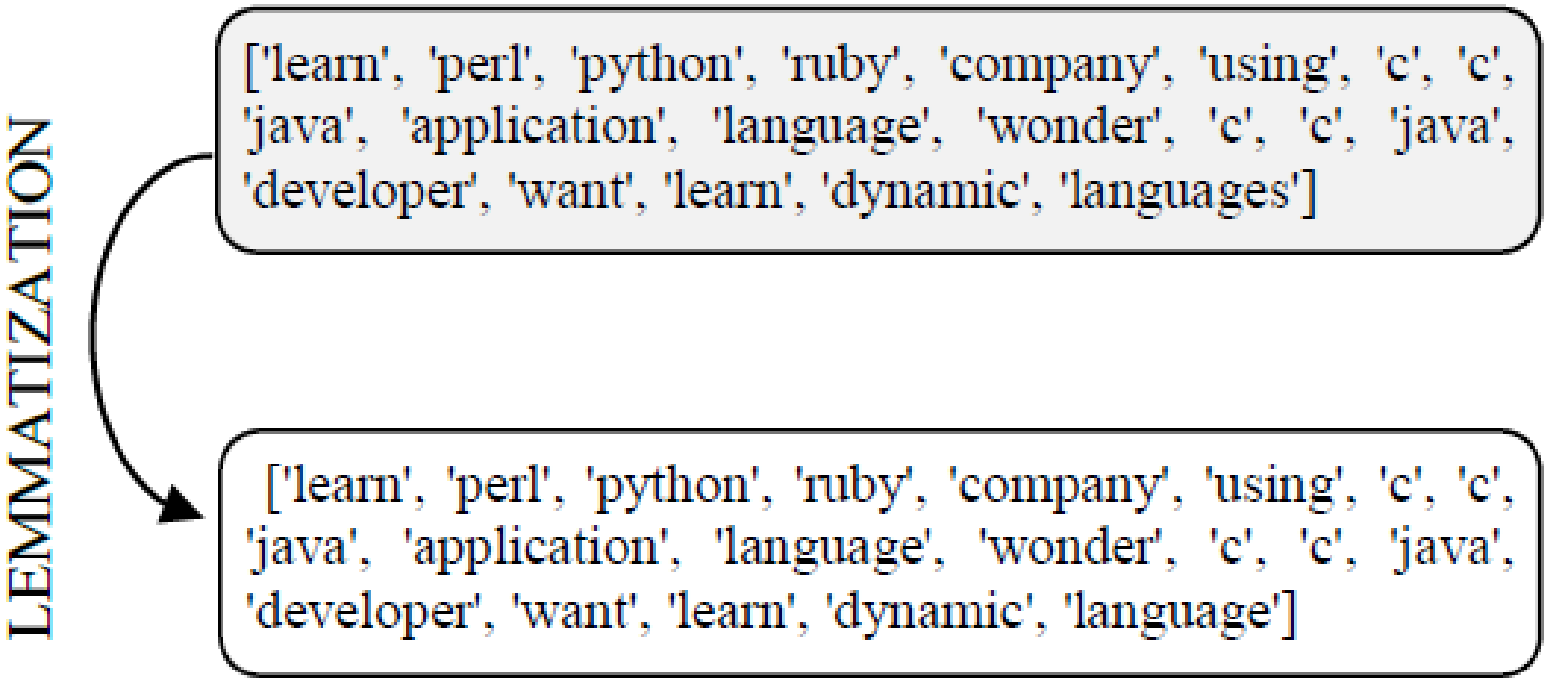
Lematyzacja

Polega na redukcji synonimów/grup słów będących odmianami danego wyrazu do jednej formy językowej.

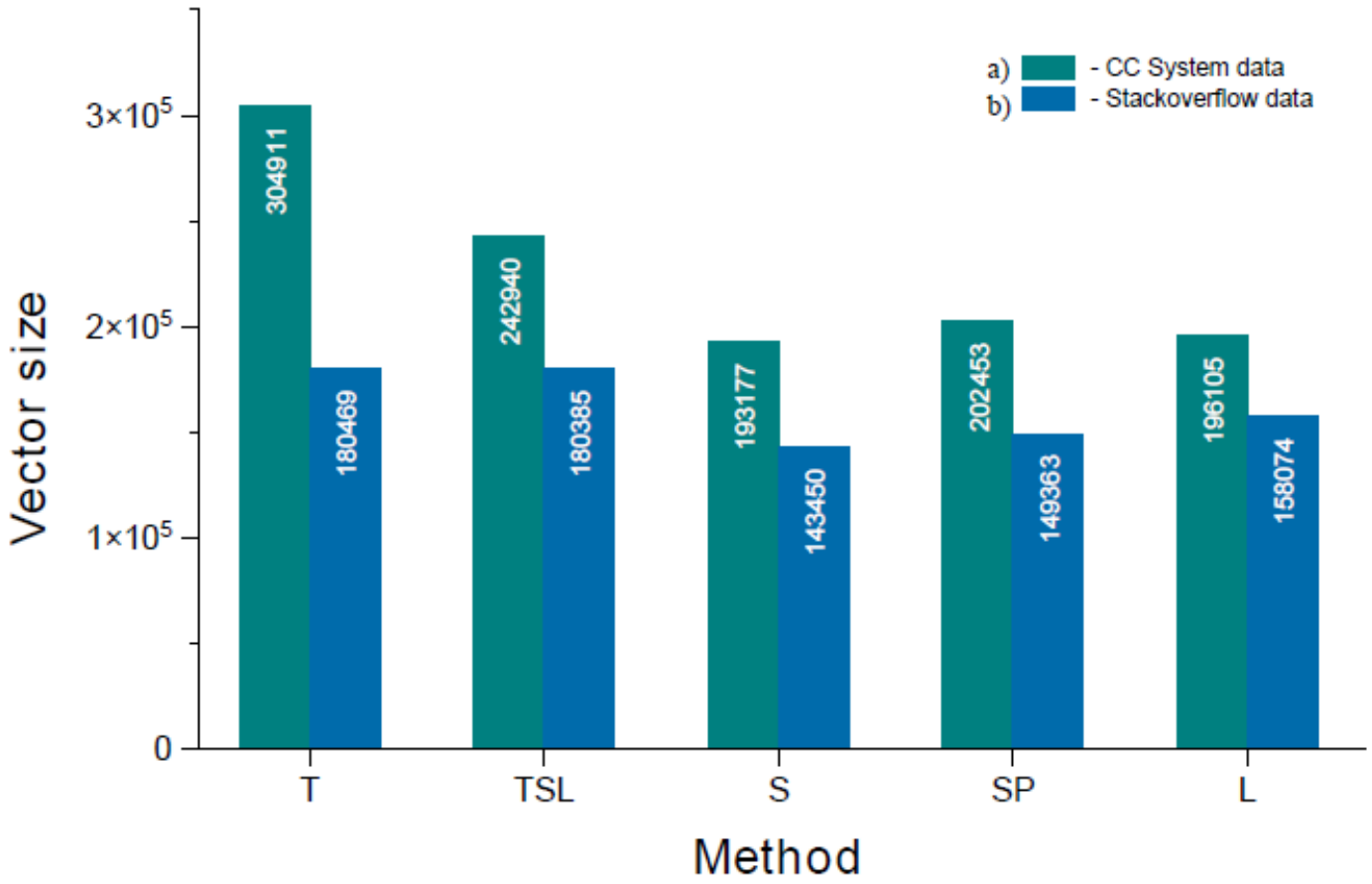
Biblioteki:

- morfeusz2 (język polski),
- nltk Wordnet lemmatizer (język angielski).

Lematyzacja



Stemming/Lematyzacja



Legend: T - Tokenization; TSL - Tokenization/Stoplist; S - Stemming;
SP - Stemming with Polimorf/Porter; L - Lemmatization

Wektoryzacja danych tekstowych

Polega na zamianie danych tekstowych na numeryczne wektory.

Najprostsza metoda Count Vectorization polega na wyznaczeniu sumy wystąpień danego słowa w dokumencie.

Wektoryzacja danych tekstowych - Count Vectorizer

Example text: *WWW is very slow. Please check this. The SARA System also reports a number of errors. The terminal runs fairly smoothly, but one process is very CPU intensive*

Dictionary

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
www	is	very	slow	please	check	this	the	SARA	system	also	reports	a	number	of	effors	terminal	runs	farly	smoothly	but	one	procesess	very	CPU	intensive

Building the vector

WWW is very slow. Please check this. The SARA System also reports a number of errors. The terminal runs fairly smoothly, but one process is very CPU intensive

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Wektoryzacja danych tekstowych - TF-IDF Vectorizer

Algorytm TF-IDF pozwala wektoryzowac dane na podstawie częstotliwości występowania oraz znaczenia danego słowa w całym zbiorze badanych dokumentów.

$$TF - IDF(w, d) = TF(w, d) \cdot IDF(w) \tag{1}$$

gdzie: w – słowo; d – dokument; TF(w, d) – częstotliwość występowania; IDF(w) – odwrócona częstotliwość występowania:

$$IDF(w) = log(\frac{DN}{DNT(w)}) \tag{2}$$

gdzie: DN – liczba dokumentów/rekordów; $DNT(w)$ – liczba dokumentów w których występuje słowo.

Wektoryzacja danych tekstowych - TF-IDF Vectorizer

Example document: *d ={'WWW is very slow. Please check this. The SARA system also reports a number of errors. The terminal runs fairly smoothly, but one process is very CPU intensive', 'System error. Unfortunately, the problem with report no. 2 returned again. The report showed different numbers of calls last week - different now.'}*

STEP 1 — **Dictionary**

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
www	is	very	slow	please	check	this	the	SARA	system	also	reports	a	number	of	effors	terminal	runs	farly	smoothly	but	one	procesess	very	CPU	intensive	error	unfortunately	problem	with	report	no	returned	again	showed	different	numbers	calls	last	week	now

STEP 2 — **Building the vector**

WWW is very slow. Please check this. The SARA System also reports a number of errors. The terminal runs fairly smoothly, but one process is very CPU intensive

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	**1**	**1**	**1**	**1**	**1**	**1**	**4**	**1**	**2**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**

System error. Unfortunately, the problem with report no. 2 returned again. The report showed different numbers of calls last week - different now

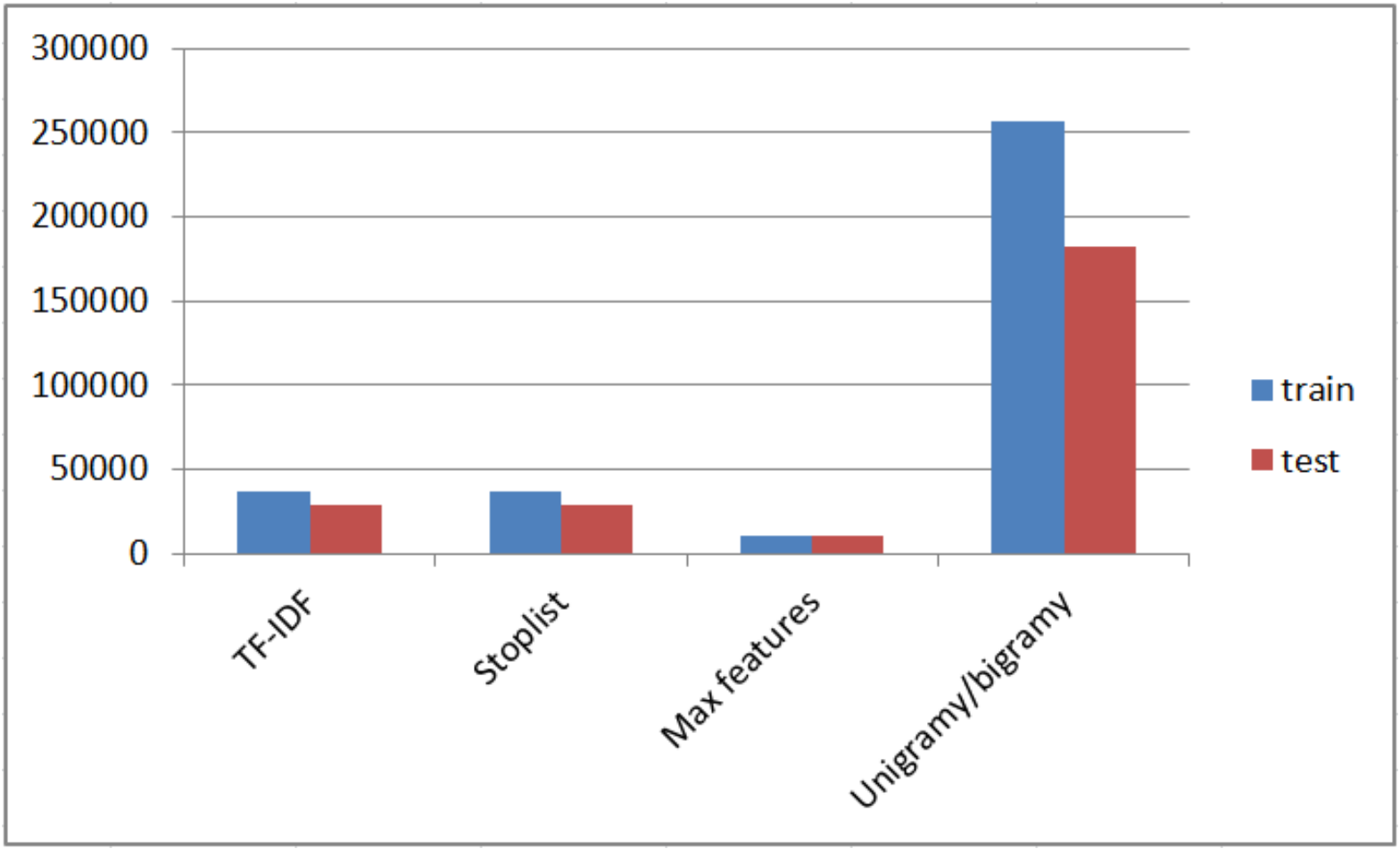
STEP 3 — **Computing TF-IDF for each word from dictionary:** DN=2;
0: w='www'; DNT(www)=1: IDF(www)=log(2/1)=0,3 | 9: w='system'; DNT(system)=2; IDF(system)=log(2/2)=0

0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0	0.3	0.3	0.3	0	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0	0.3	0.3	0.3	0.3	0	0.3	0.3	0.3	0.3	0.3

Wektoryzacja danych tekstowych - TF-IDF Vectorizer

```
# Wektoryzacja
vectorizer = TfidfVectorizer(max_features=1000, stop_words="english", ngram_range=(1,2))
x_train = vectorizer.fit_transform(data_train.data).toarray()
x_test = vectorizer.fit_transform(data_test.data).toarray()
```

Wektoryzacja danych tekstowych - TF-IDF Vectorizer



Metody transformacji przestrzeni liniowej

umożliwiają odwzorowanie wektorów z jednej przestrzeni wektorowej na wektory w innej przestrzeni.

- LSA, (Latent Semantic Analysis),
- RP (Random Projection),
- PCA (Principal Component Analysis),
- ICA (Independent Component Analysis).

Klasyfikacja

Przyporządkowanie rekordu do predefiniowanych klas.

- binarna (dwuklasowa) - np. czy mail to spam czy nie. Jedno wyjście ANN.
- wieloklasowa (multiclass) - przyporządkowująca rekordom jedną z wielu klas. Każde wyjście ANN to osobna klasa.
- wieloetykietowa (multilabel) - jeden rekord może należeć do wielu klas.

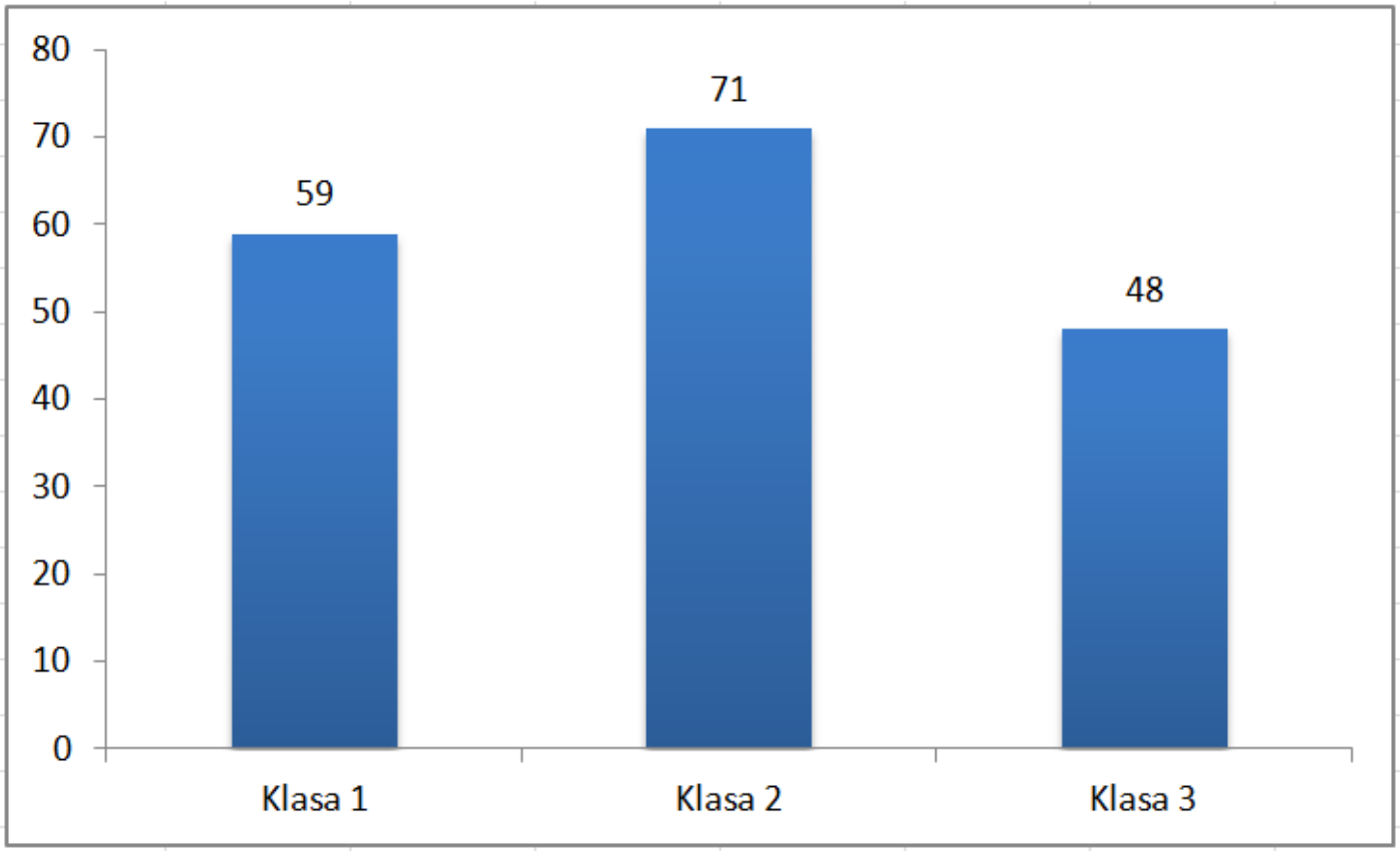
Kodowanie danych

- binarna (dwuklasowa) - 0, 1.
- wieloklasowa (multiclass) - [0,1,0,0,0].
- wieloetykietowa (multilabel) - [0,1,0,1,1].

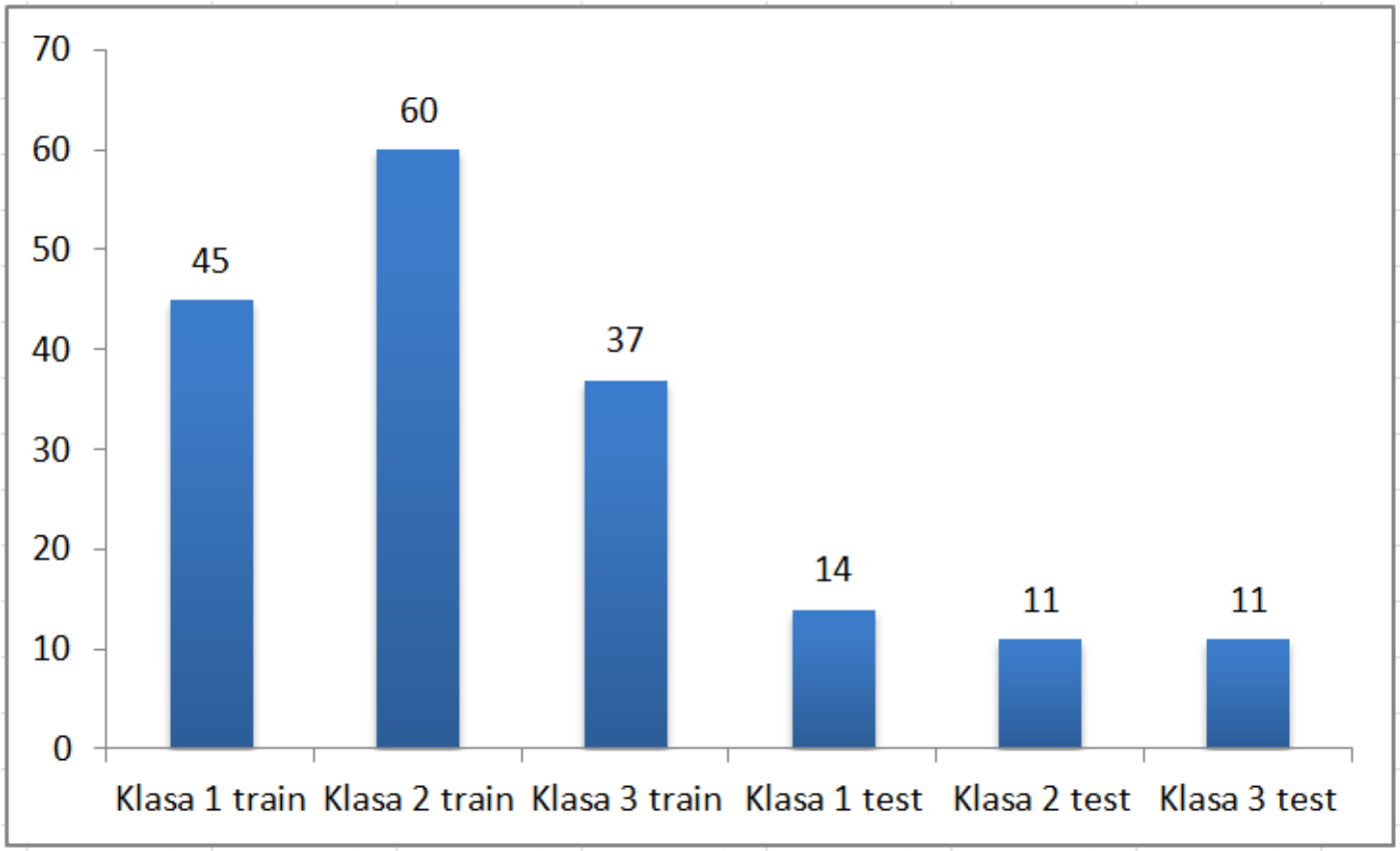
Przykładowe klasy

1	Why learn perl python ruby if the company is using c c or java as the application language i wonder why would a c c java developer want to learn a dynamic language...	C#, Java, Python
2	How do i connect to a database and loop over a recordset in c whats the simplest way to connect and query a database for a set of records in c	C#
3	Codility absolute thistinct count from an array so i took the codility interview test yesterday and was informed today that i failed unfortunately i wasnt given any other information...	C#, C++, Java, Python

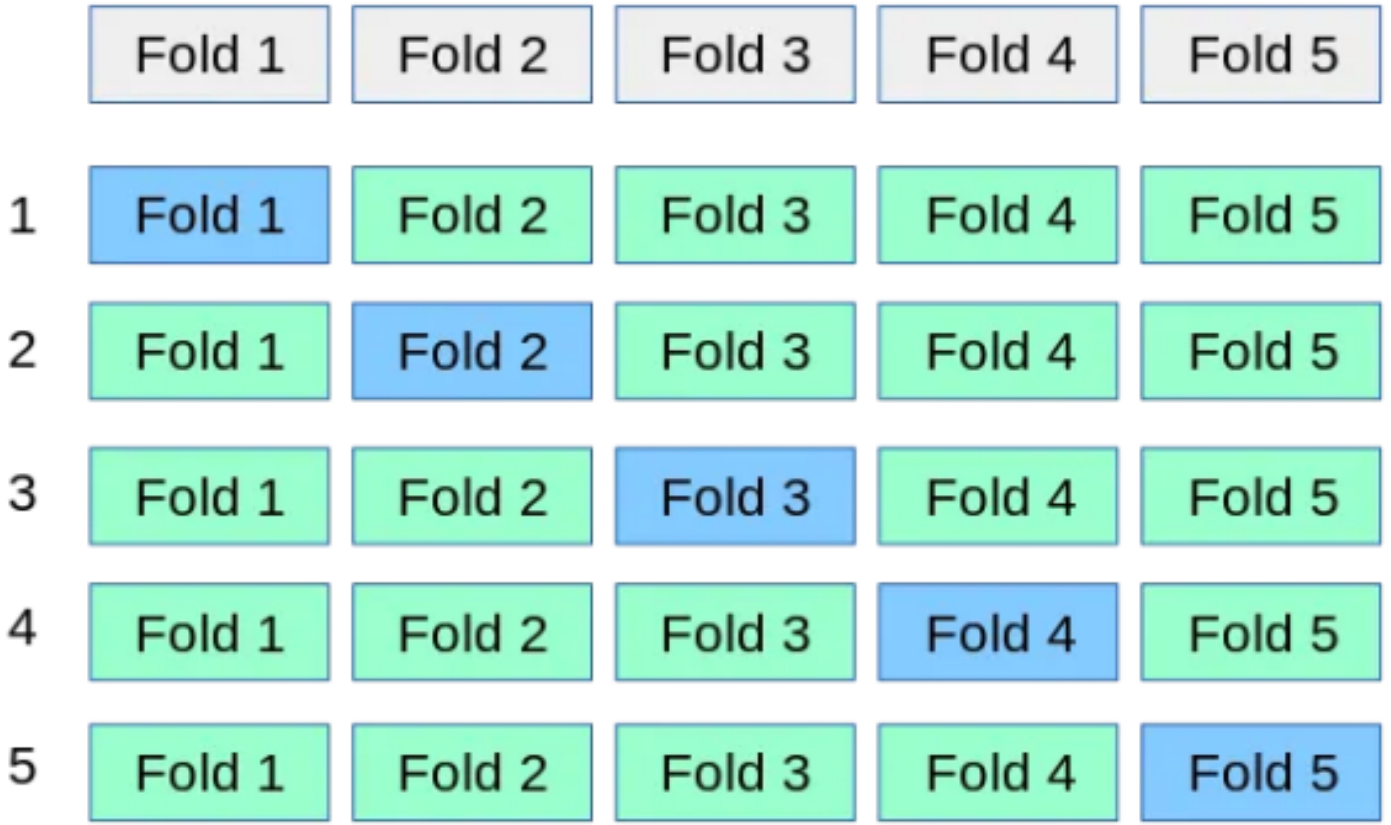
Zbalansowanie danych



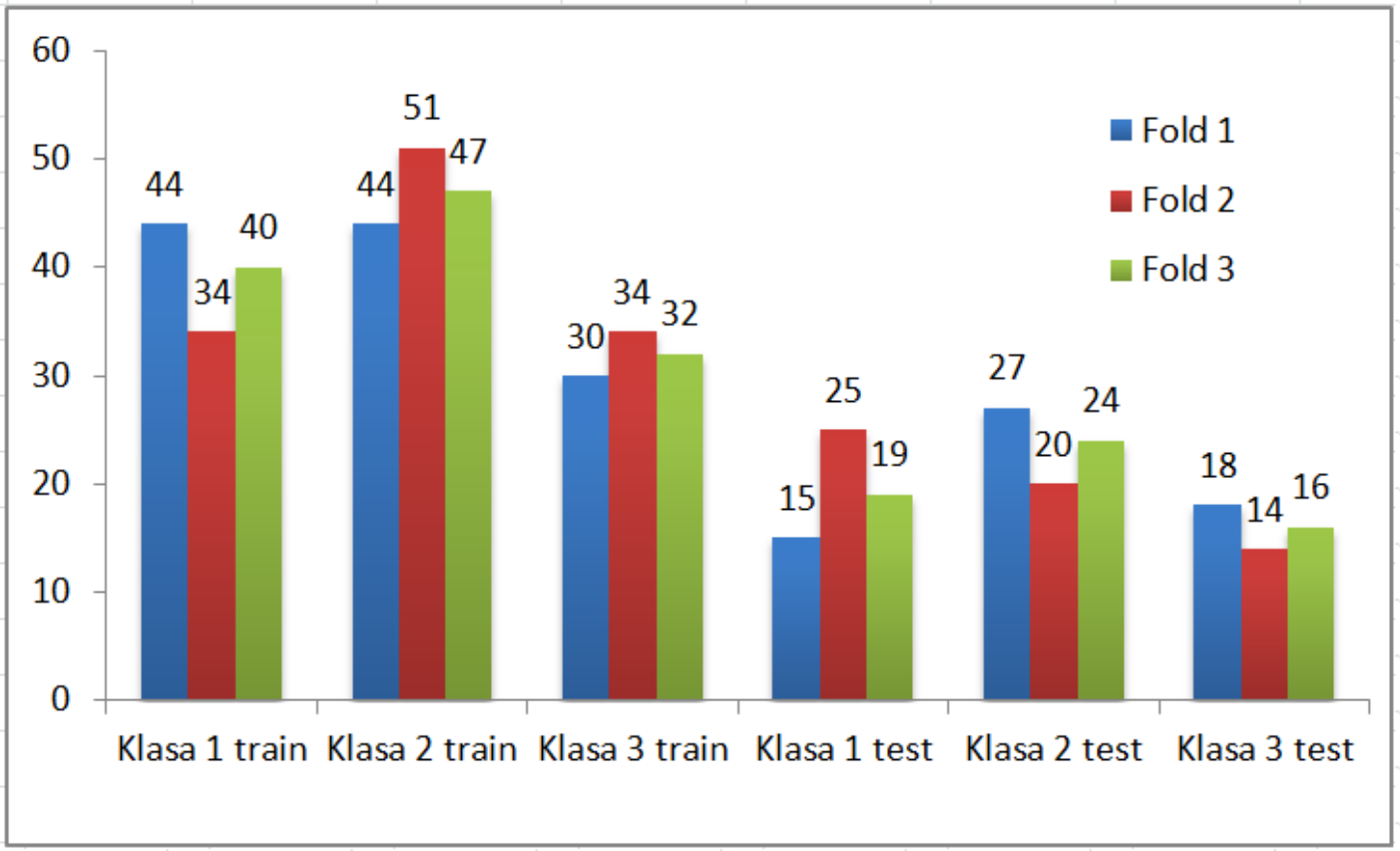
Podział na dane uczące/testowe



K-fold cross validation



K-fold cross validation



Klasyfikacja

?

Błąd

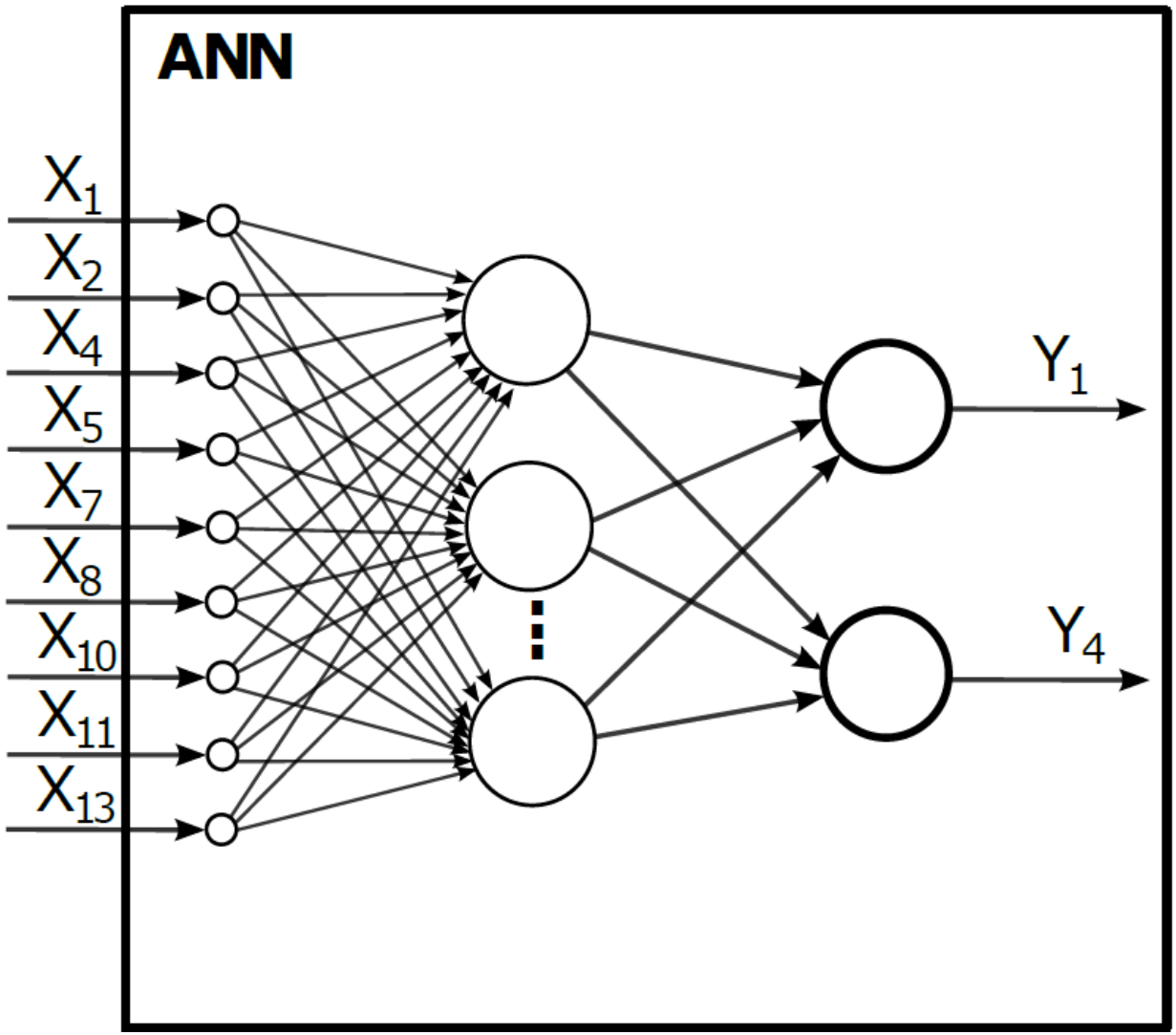
Usługa

Połączenie

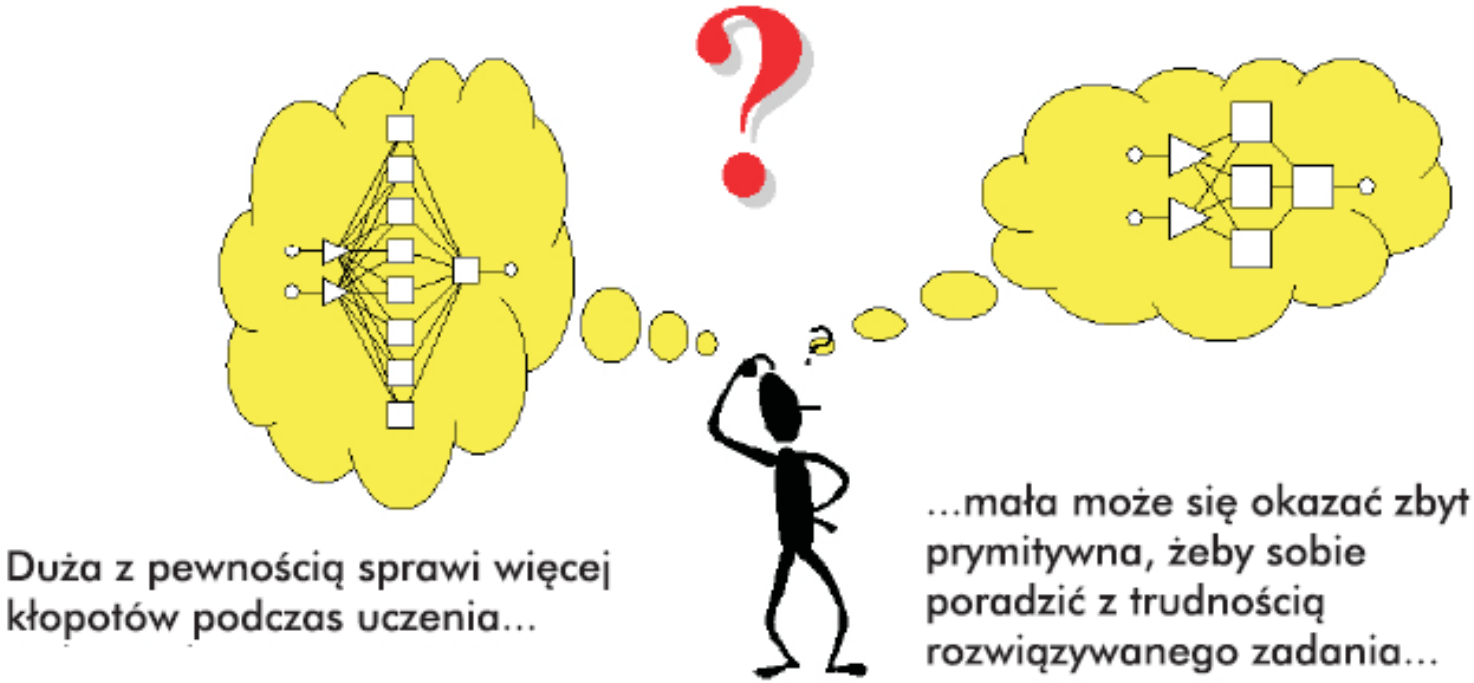
Reklamacja

Produkt

Sztuczna sieć neuronowa (SSN)



Jaka struktura



Sztuczna sieć neuronowa (SSN)

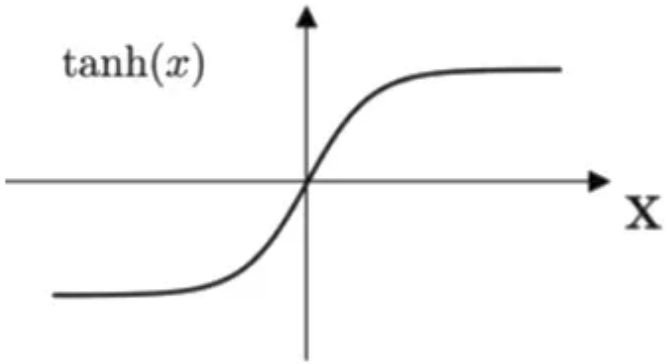
- rodzaj sztucznej sieci neuronowej,
- struktura sieci (liczba warstw, neuronów),
- rodzaj funckji aktywacji,
- algorytm uczenia, parametry uczenia,
- dane uczące i walidujące.

Sztuczne sieci neuronowe

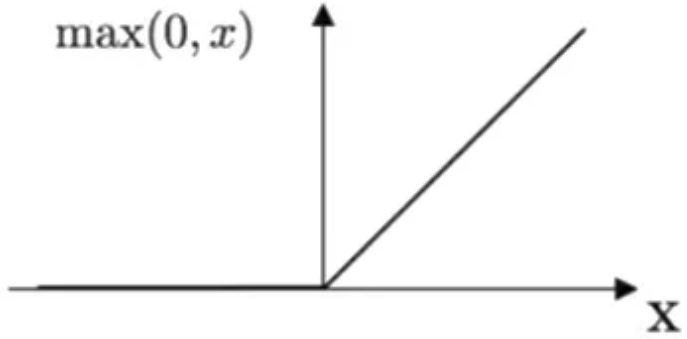
```
#Inicjalizacja modelu sztucznej sieci neuronowej
model = tf.keras.Sequential()
model.add(tf.keras.layers.Dense(10, input_dim=n_features, activation='relu'))
model.add(tf.keras.layers.Dense(5, activation='relu'))
model.add(tf.keras.layers.Dense(units = n_classes, activation = 'softmax'))
model.summary()
```

Funkcje aktywacji

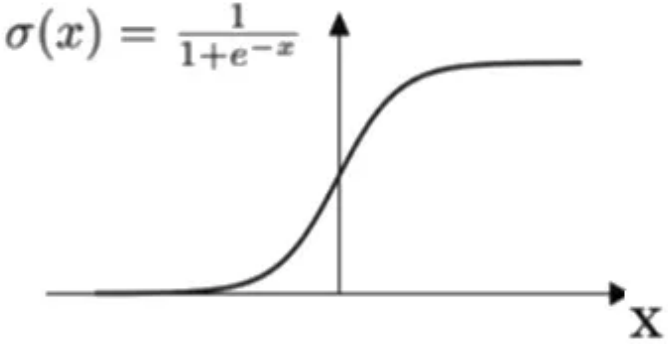
Tanh



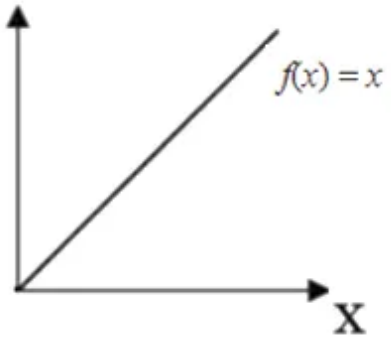
ReLU



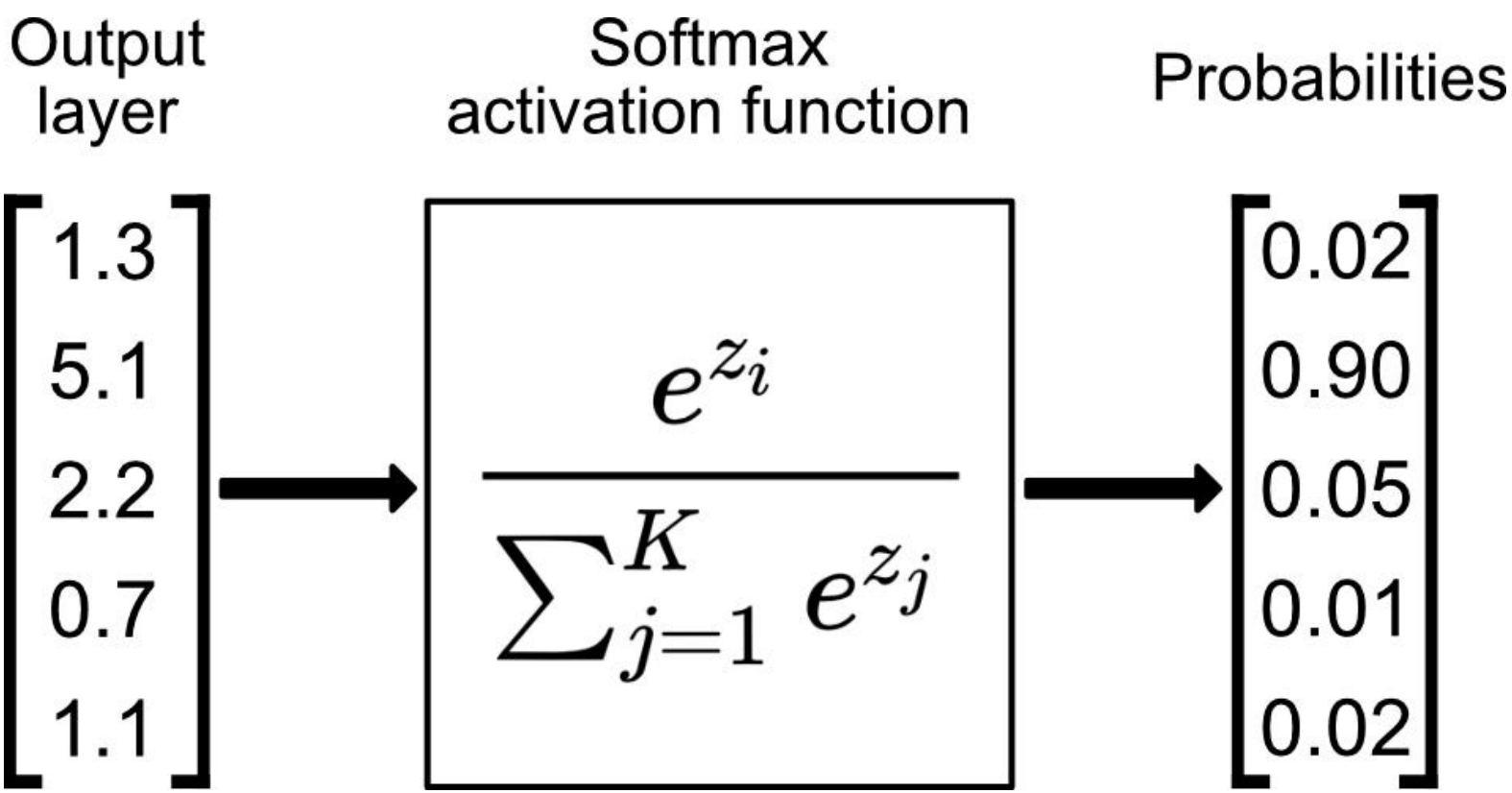
Sigmoid



Linear



Funkcje aktywacji



Algorytmy uczenia

- SGD
- RMSprop
- Adam
- AdamW
- Adadelta
- Adagrad
- Adamax
- Adafactor
- Nadam
- Ftrl
- Lion
- Loss Scale Optimizer

SGD

Metoda wstecznej propagacji błędu z momentem.

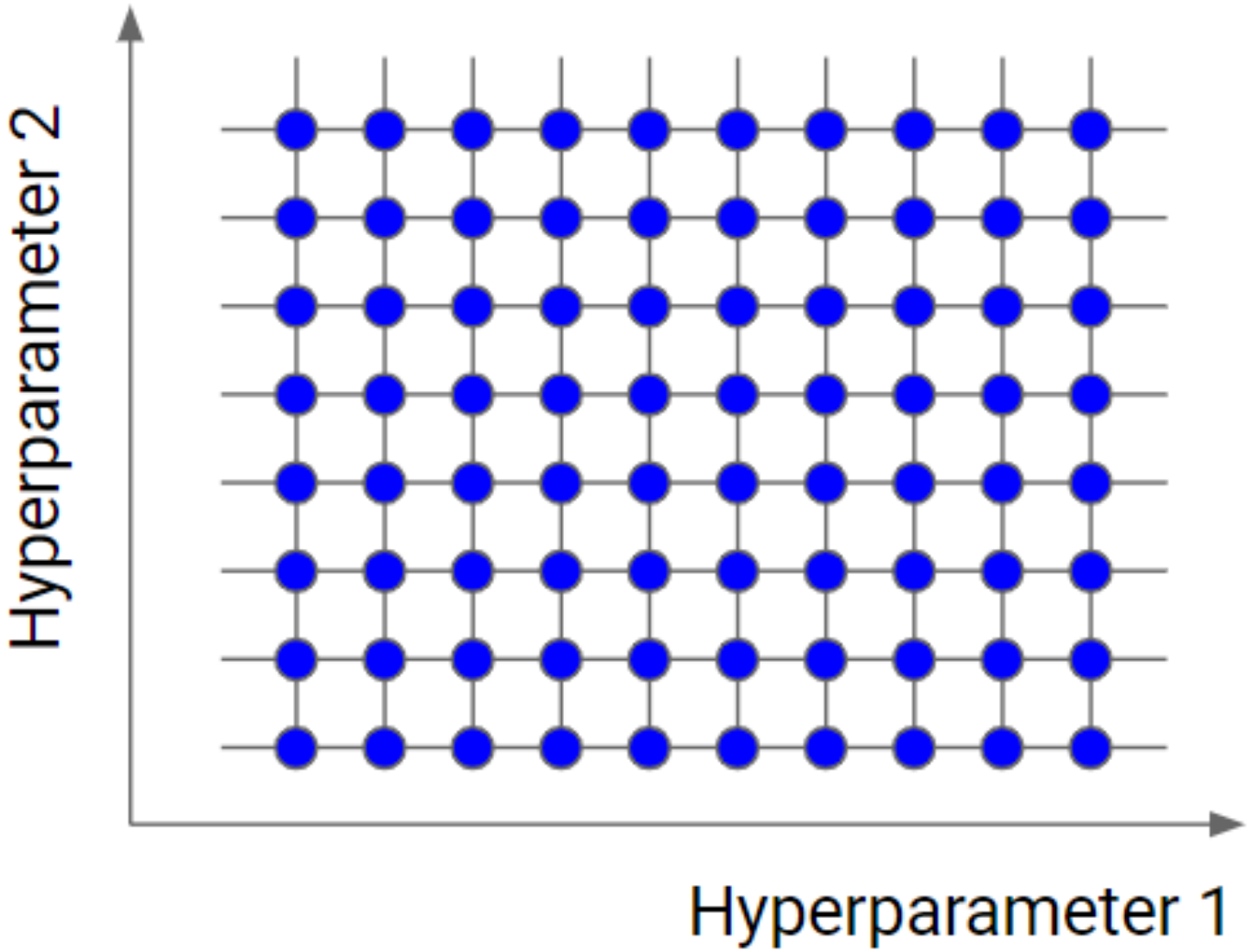
```
w = w - learning_rate * g
```

```
velocity = momentum * velocity - learning_rate * g
w = w + velocity
```

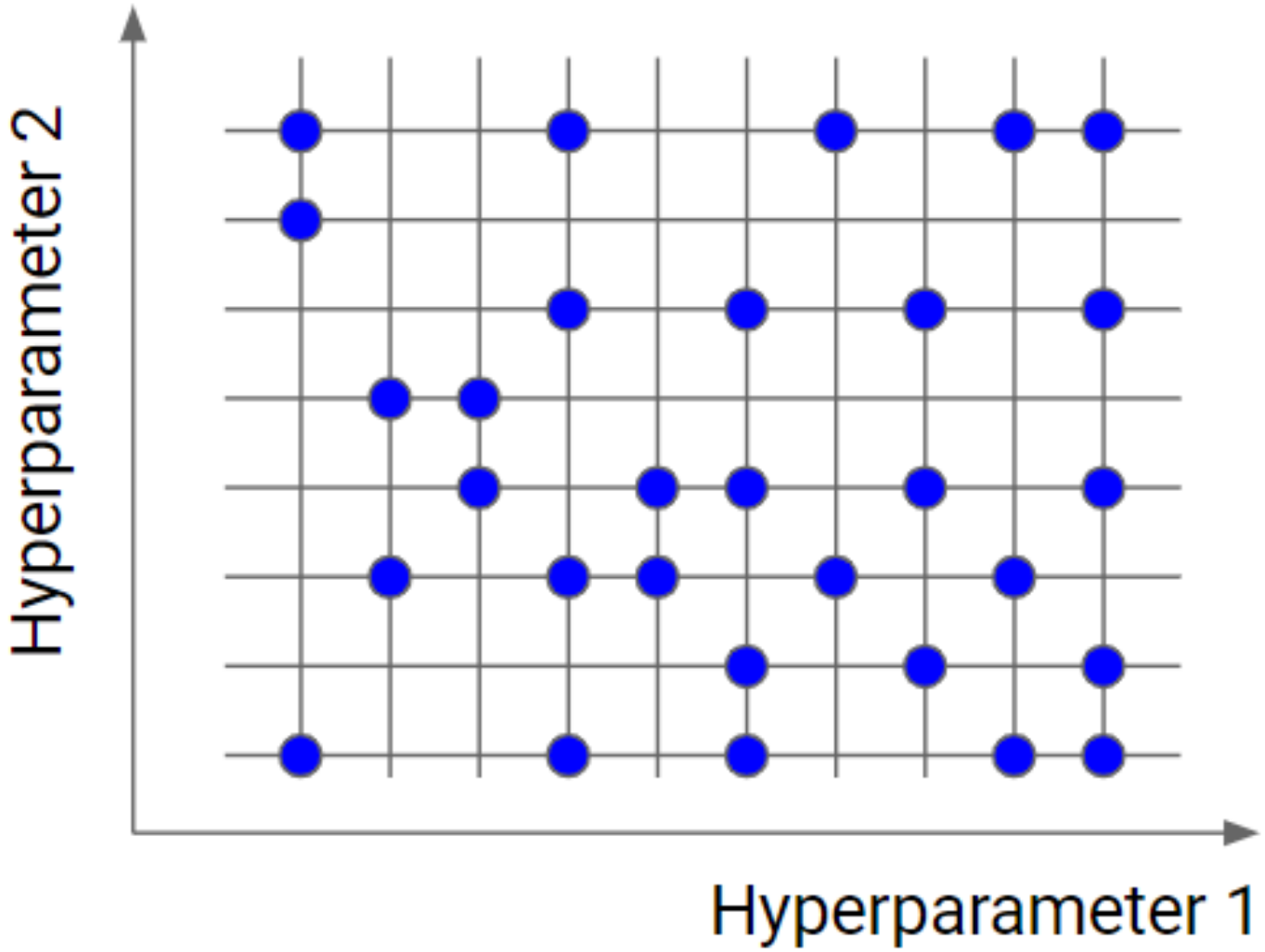
Adam

Metoda gradientowa (rozszerzenie SGD), która opiera się na adaptacyjnej estymacji momentów pierwszego i drugiego rzędu.

Metoda siatki (grid search)



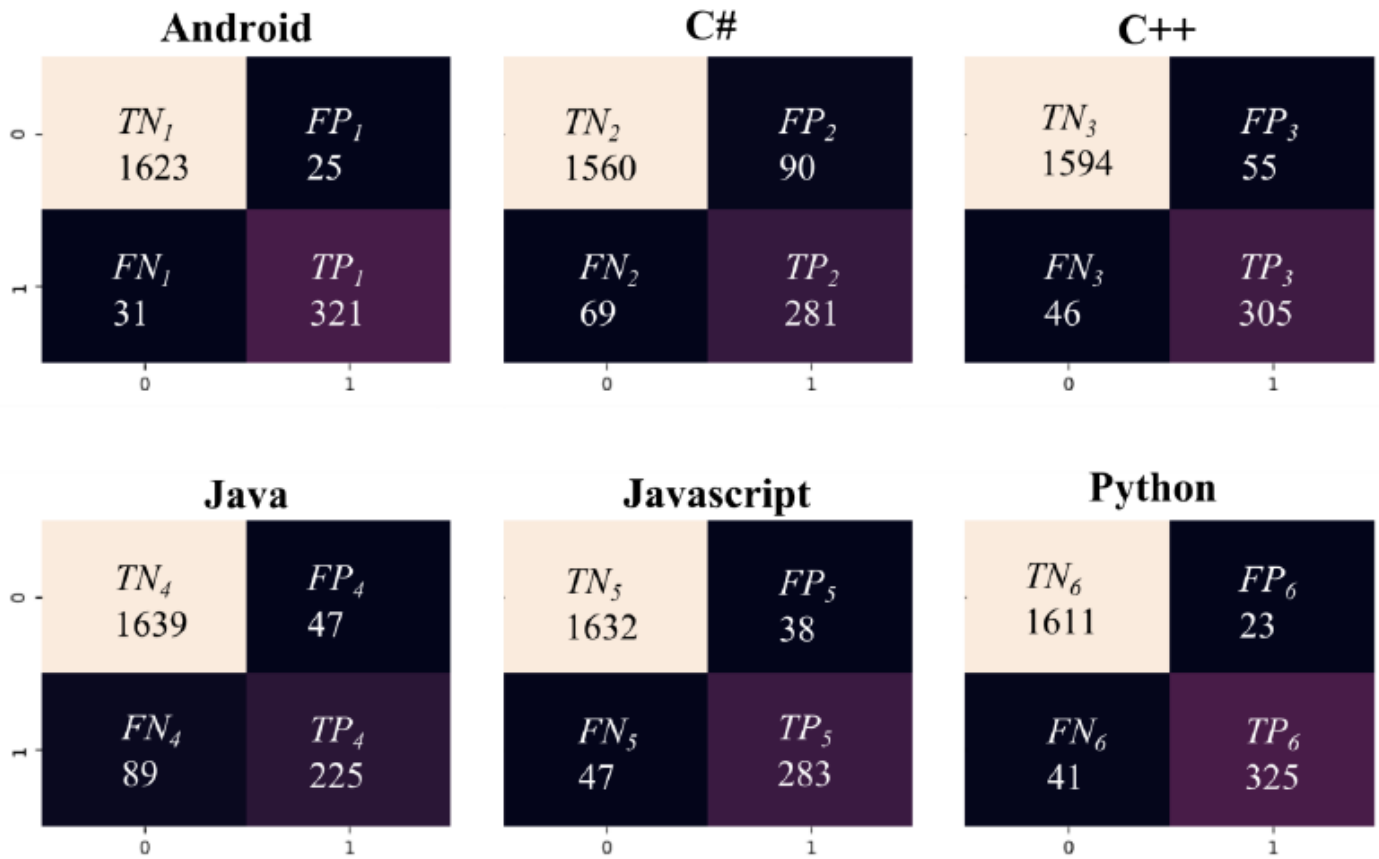
Metoda losowa (random search)



Testowanie

Jak testować?

Macierz pomyłek



Accuracy

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

Recall

$$recall = \frac{TP}{TP + FN} \tag{4}$$

Precision

$$precision = \frac{TP}{TP + FP} \tag{5}$$

F1score

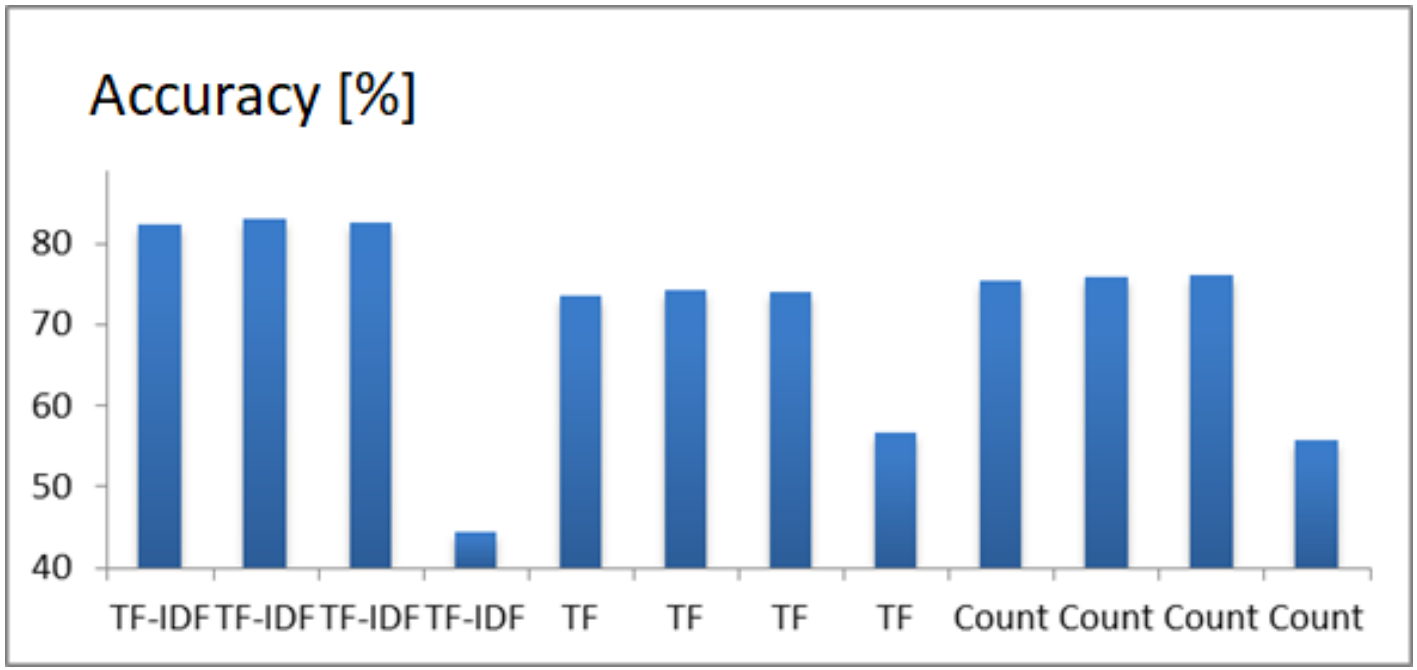
$$F1score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{6}$$

Exact match

$$\text{exact match} = \frac{1}{n}\sum_{j=1}^{n} I(h(x_j) = y_j) \cdot 100\% \tag{7}$$

gdzie: n – liczba rekordów, $h(x_j)$ – prognozowane klasy dla x_j, y_j – właściwe klasy dla x_j, $I(true) = 1$ i $I(false) = 0$.

Jak testować - wizualizacja



Jak testować

Method	CC System		Stackoverflow	
	Test accuracy [%]	Test emotica [%]	Test accuracy [%]	Test emotica [%]
DT	85.11	52.40	92.78	77.05
SVM	92.91	75.36	90.47	54.85
RF+LP	90.03	71.75	94.62	72.45
FastText	90.61	68.52	94.72	81.55
BERT	93.58	78.54	96.34	86.80
Proposed approach	93.59	79.19	95.88	83.15

Jak testować

Text	Set category	ANN classifier category
Notification with task user rule. It is not possible to end the task and send a notification e-mail if the task user is entered in the notification rule.	Incident, ECM	Incident, ECM
Revocation of a transfer of equipment. At the client's 355500, we prepared a new contract and transferred the equipment to it. Unfortunately, the client changed his mind about signing the contract. How are we going to transfer the equipment to the previous contract?	Service, SARA	Incident, SARA
Correction of the number of configured connections. Referring to our conversation, please configure the correct number of channels supported by the Helpline. There are currently 30 channels configured, which is the default value. 50 pcs is to be configured	Service, ACC	Service, ACC
Support request (# 31639) - complete Please complete the field E07 – the date of completion of a given task by a given consultant. Of course, there can be many end dates for the same task if the process has loops.	Service, ECM	Service, ECM
GDPR application - analysis of startup problems. The problem occurred on 10/12/2018 (The application was launched on that day) Conclusions: The application cannot start due to problems with connecting to the database	Incident, Systemic	Incident, Systemic

Bibliografia

- L. Rutkowski: Metody i techniki sztucznej inteligencji.Wydawnictwo naukowe PWN, Warszawa, 2009.
- R. Tadeusiewicz: Odkrywanie właściwości sieci neuronowych przy użyciu programów w języku C#. Polska Akademia Umiejętności, Kraków, 2007.
- https://isheunesu48.medium.com/cross-validation-using-k-fold-with-scikit-learn-cfc44bf1ce6 [21.03.2024]
- https://towardsdatascience.com/softmax-activation-function-explained-a7e1bc3ad60 [21.03.2024]
- Goel A., Goel A.K., Kumar A. The role of artificial neural network and machine learning in utilizing spatial information. Spat. Inf. Res. 2023;31(3):275–85. doi: 10.1007/s41324-022-00494-x.
- https://www.yourdatateacher.com/2021/05/19/hyperparameter-tuning-grid-search-and-random-search/ [22.03.2024]
- https://keras.io/api/optimizers/ [22.03.2024]

Bibliografia

- K. Poczeta, M. Płaza, T. Michno, M. Krechowicz, M. Zawadzki, A multi-label text message classification method designed for applications in call/contact centre systems, Applied Soft Computing 145, 2023.
- L. Kant, Predicting Tags for StackOverflow, 2021, URL https://www.kaggle.com/laxmimerit/predicting-tags-for-stackoverflow-deep-learning/data. (Accessed 29 November 2021)
- https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html [11.04.2024]
- G. Madjarov, D. Kocev, D. Gjorgjevikj, S. Dzeroski, An extensive experimental comparison of methods for multi-label learning, Pattern Recognition 45 (9) (2012) 3084–3104, best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011).

Dziękuję za uwagę!