



Mašinsko učenje 2024

Zadatak 3

Sadržaj



Zadatak 2 - Rekapitulacija



Zadatak 3

Zadatak 2 - Rekapitulacija

Zadatak 2 - Rekapitulacija

- Procenat uspešnosti: **65%** (22/34).
- Najveće preklapanje izvornih kodova prema alatu za detekciju plagijata: **27%**.
- Najbolji rezultati po terminima:

Termin	Tim	RMSE
Ponedeljak - G4	tim1_24	1919.49
Utorak - G5	placeholder	2261.20
Utorak - G3	tim_10	1803.90
Četvrtak - G2	tim13_24	2221.03
Petak - G1	Tehno trube	1955.56

Zadatak 2 - Rekapitulacija

- Dobre stvari (na nivou generacije):
 - Implementacija algoritama
 - Rad sa trening skupom podataka
 - Računanje metrike
 - Prpratni izveštaji.
- Stvari koje mogu biti bolje (na nivou generacije):
 - Selekcija obeležja.

Zadatok 3

Zadatak 3

- Klasifikacija:
 - Koristeći tekstove strofa pesama domaćih i regionalnih muzičkih izvođača (kolona **strofa**), identifikovati žanr svake pesme (kolona **žanr**):
 - folk
 - pop
 - rock
 - Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije **mikro f1 mera** (eng. *micro f1 score*) veća od 0.70.
 - Zadatak se rešava upotrebom isključivo jednog klasifikatora.
 - Rok za izradu zadatka je **08.05.2024. u 23:59h**.

Zadatak 3

- Klasifikacija:
 - Dozvoljeni klasifikatori za Zadatak 3:
 - Logistička regresija
 - **Perceptron**
 - Naivni Bajes
 - Mašine potpornih vektora (SVM).
 - Instalirane biblioteke za Zadatak 3:
 - NumPy
 - Pandas
 - SciPy
 - scikit-learn.

Zadatak 3

- Sledeći termin vežbi (odbrana Zadatka 3 i predstavljanje Zadatka 4):

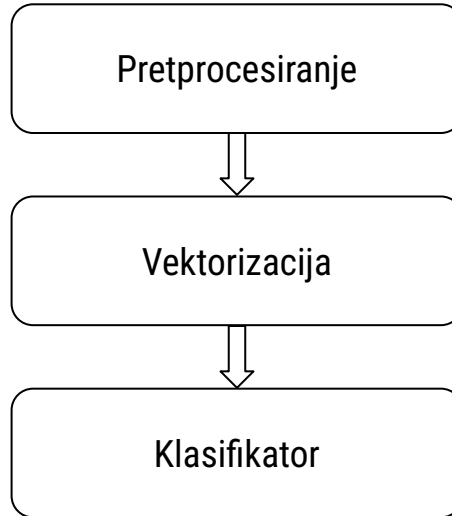
Termin	Datum
Ponedeljak - G4	13.05.2024.
Utorak - G5	14.05.2024.
Utorak - G3	14.05.2024.
Četvrtak - G2	16.05.2024.
Petak - G1	17.05.2024.

Zadatak 3

- **scikit-learn** biblioteka:
 - Instalacija
 - Docs.
- Izdvojeno:
 - Selekcija modela
 - Metrike.

Zadatak 3

- Koraci kod klasifikacije teksta:



Zadatak 3

- Koraci kod klasifikacije teksta:
 - **Pretprocesiranje:**
 - Transformacija ulaznog teksta:
 - Svođenje teksta na mala ili velika slova
 - Uklanjanje znakova interpunkcije
 - Uklanjanje reči bez značenja (eng. *stopwords*)
 - ...
 - **Sav tekstualni ulaz (i trening i test) mora proći kroz isto pretprocesiranje.**

Zadatak 3

- Koraci kod klasifikacije teksta:
 - **Vektorizacija:**
 - Pretvaranje teksta u numerički oblik kako bi ga klasifikator mogao bolje razumeti i obraditi
 - Svaki tekst se pretvara u vektor numeričkih vrednosti koje predstavljaju određene karakteristike ili attribute tog teksta
 - Najpoznatiji vektorizatori:
 - Bag of Words
 - TF-IDF

Zadatak 3

- Koraci kod klasifikacije teksta:

- **Vektorizacija:**

- Bag of Words:

Document D1	<i>The child makes the dog happy</i> the: 2, dog: 1, makes: 1, child: 1, happy: 1
Document D2	<i>The dog makes the child happy</i> the: 2, child: 1, makes: 1, dog: 1, happy: 1



	child	dog	happy	makes	the	BoW Vector representations
D1	1	1	1	1	2	[1,1,1,1,2]
D2	1	1	1	1	2	[1,1,1,1,2]

Zadatak 3

- Koraci kod klasifikacije teksta:
 - Vektorizacija:
 - TF-IDF:

TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Term frequency

Number of times term t appears in a doc, d

Inverse document frequency

$\log \frac{1 + n}{1 + \text{df}(d, t)}$

n ← # of documents

$\text{df}(d, t)$ ← Document frequency of the term t

Zadatak 3

- Koraci kod klasifikacije teksta:
 - **Vektorizacija:**
 - scikit-learn:
 - Izdvajanje osobina iz teksta:
 - Bag of Words
 - TF-IDF.
 - **Vektorizator obučen na trening skupu se primenjuje i na trening i na testni skup:**

```
vectorizer = CountVectorizer()
X_train = vectorizer.fit_transform(train_corpus)
X_test = vectorizer.transform(test_corpus)

ili

vectorizer = CountVectorizer()
vectorizer.fit(train_corpus)
X_train = vectorizer.transform(train_corpus)
X_test = vectorizer.transform(test_corpus)
```


Zadatak 3

- Koraci kod klasifikacije teksta:
 - Klasifikator:
 - Treniranje i evaluacija klasifikatora
 - Dozvoljeni klasifikatori za Zadatak 3 u scikit-learn:
 - Logistička regresija i Perceptron
 - Naivni Bajes
 - Mašine potpornih vektora.

Zadatak 3

- Kao meru performansi modela u ovom zadatku imamo mikro f1 meru (eng. *micro f1 score*).
- Ova metrika se, kao i većina metrika klasifikacije, izvodi iz matrice konfuzije (eng. *confusion matrix*):

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Type I error
(false positive)



Type II error
(false negative)

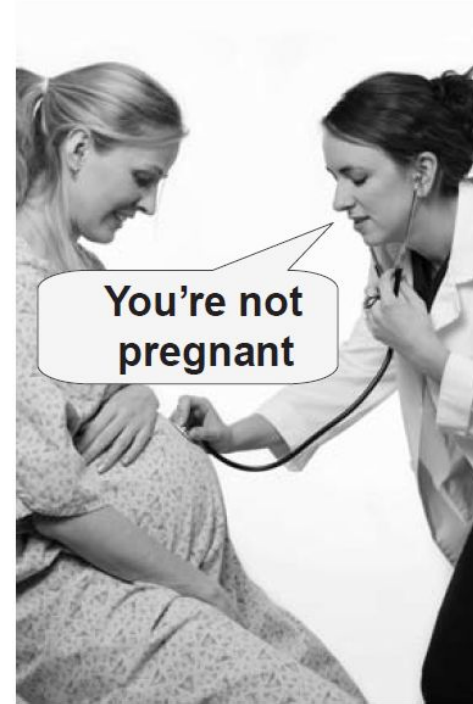


Figure 3.1 Type I and Type II errors

Zadatak 3

- **Precision** - procenat relevantnih (tačnih) među prediktovanim:
 - $P = TP / (TP + FP)$
- **Recall** - procenat relevantnih (tačnih) koje su prediktovane:
 - $R = TP / (TP + FN)$
- **F1 score** (aka ***F - measure***) - harmonijska sredina **Precision** i **Recall**:
 - $F1 = 2 * (P * R) / (P + R)$
- **Micro F1 score** - računa globalne **TP**, **FN** i **FP**:
 - `sklearn.metrics.f1_score(y_true, y_pred, average='micro')`
- Prilikom treninga, od pomoći može biti i `classification_report`.

Zadatak 3

- Saveti za rešavanje zadatka:
 - Podsetiti se gradiva sa predavanja
 - Uraditi eksplorativnu analizu podataka
 - Isprobati više operacija za pretprocesiranje teksta
 - Isprobati više vektORIZATORA i obratiti pažnju kako se radi vektorizacija
 - Isprobati više klasifikatora i analizirati njihovo ponašanje po klasama.