Detekcija spam veb stranica pomoću URL-a

Marina Borozan 1092/18

Uvod

- **Spam (štetne) veb stranice su veliki problem u današnje vreme, uzimajući u obzir učestalost internet aktivnosti**
- **“Spam tehnologija” napreduje I ima mogućnosti da svoje stranice visoko rangira u rezultatima pretrage**
- **Motiv za detekciju putem URL-a je smanjenje mogućnosti posećivanja stranice ukoliko se na osnovu URL-a može predvideti da je štetna**

Vrste spam-a

- **Phishing (Pecanje) - krađa identiteta (ličnih podataka), najčešće putem posebnog email-a ili chat-a**
- **Malware (Zlonameran softver) - softver koji je namenjen za nanošenje štete na računaru i računarskim mrežama**
- **Defacement -izmena izgleda(sadržaja) postojeće veb stranice**
- **Spam u užem smislu - slanje neželjenih masovnih poruka bez ikakvog kriterijuma**

Kratak opis procesa

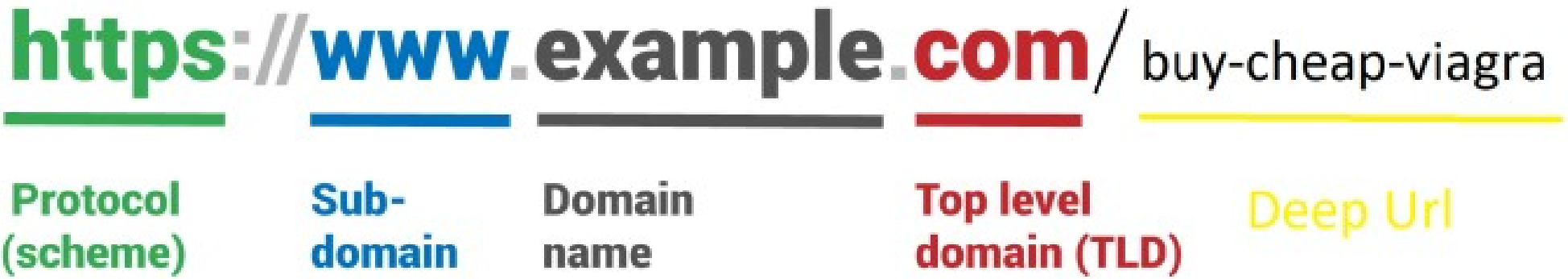
- **Obrada URL-ova radi dobijanja značajnih informacija**
- **Procesiranje tekstualnih podataka putem specijalizovanih algoritama – CountVectorizer, Multinomial Naive Bayes**
- **Klasifikacija pomoću algoritama LinearSVM, SVM sa RBF kernelom I RandomForest**

Podaci

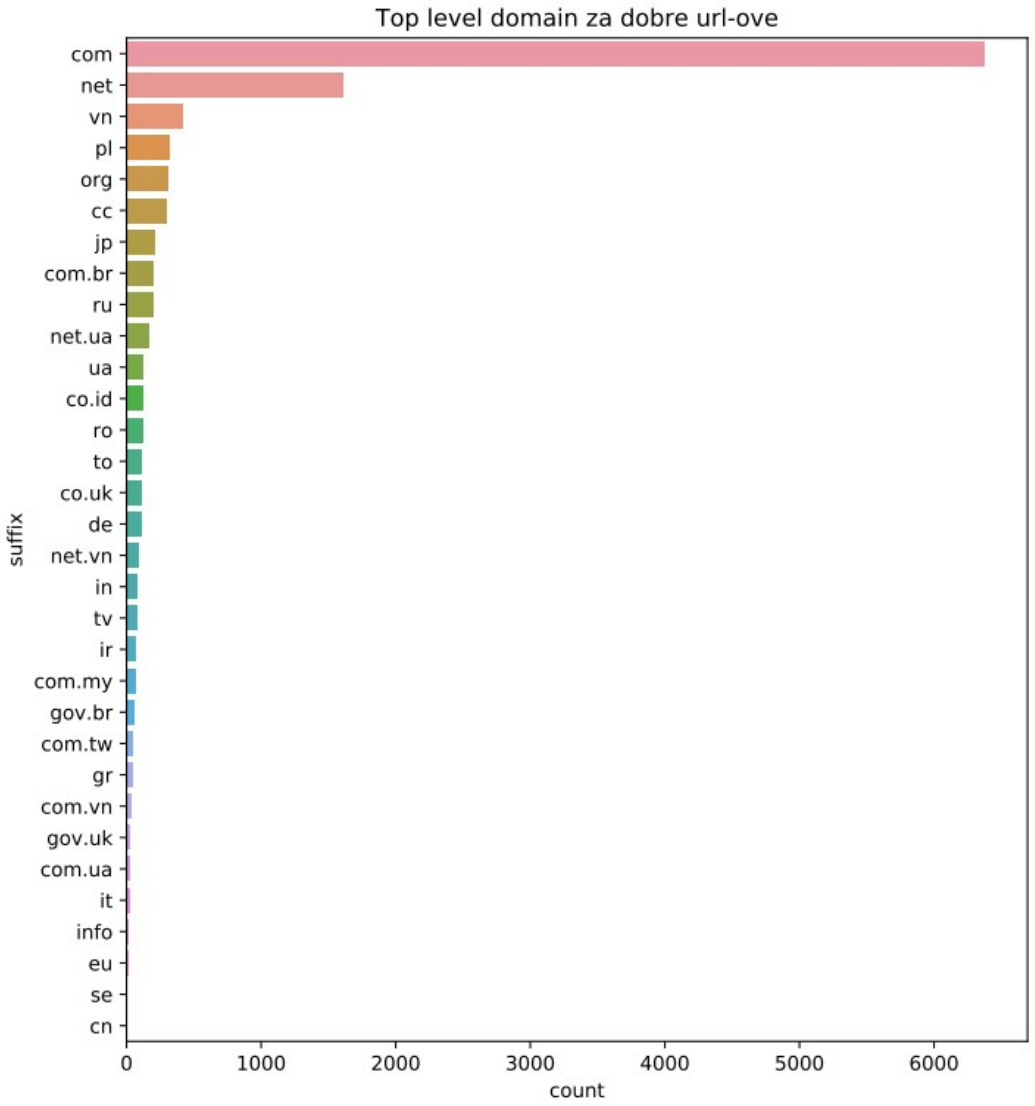
- **Set podataka čini oko 89 000 URL-ova, od čega dobri(ne štetni) čine oko 40%**
- **Podaci su dobijeni sa sajta Canadian Institute for Cybersecurity**

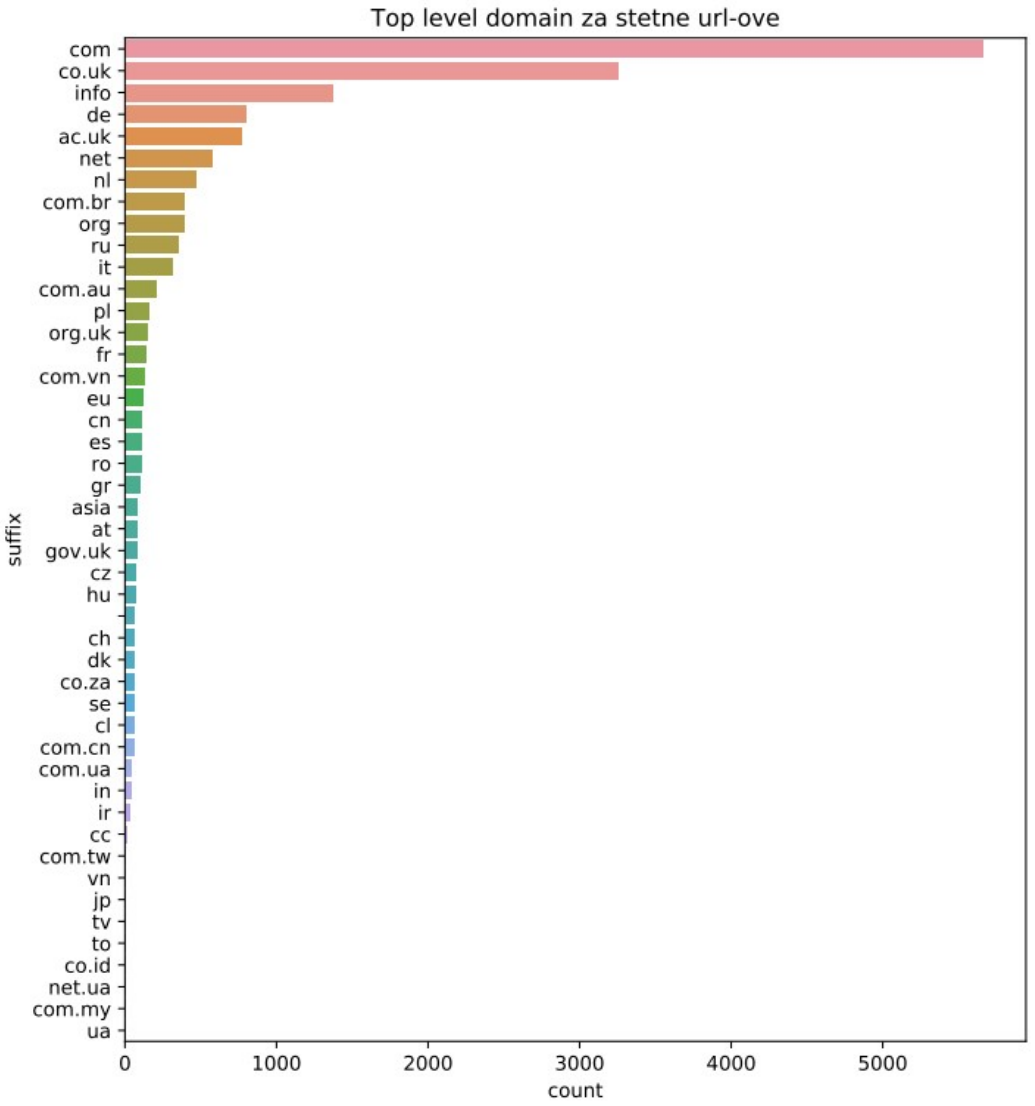
spam	53531
ne spam	35378

Procesiranje URL-a

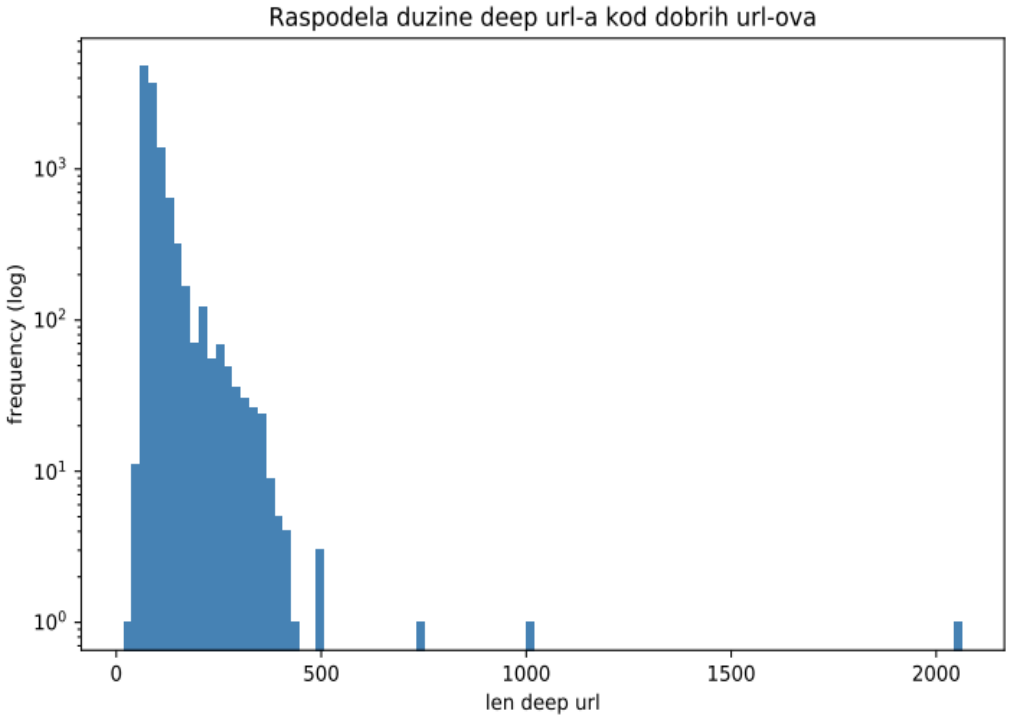


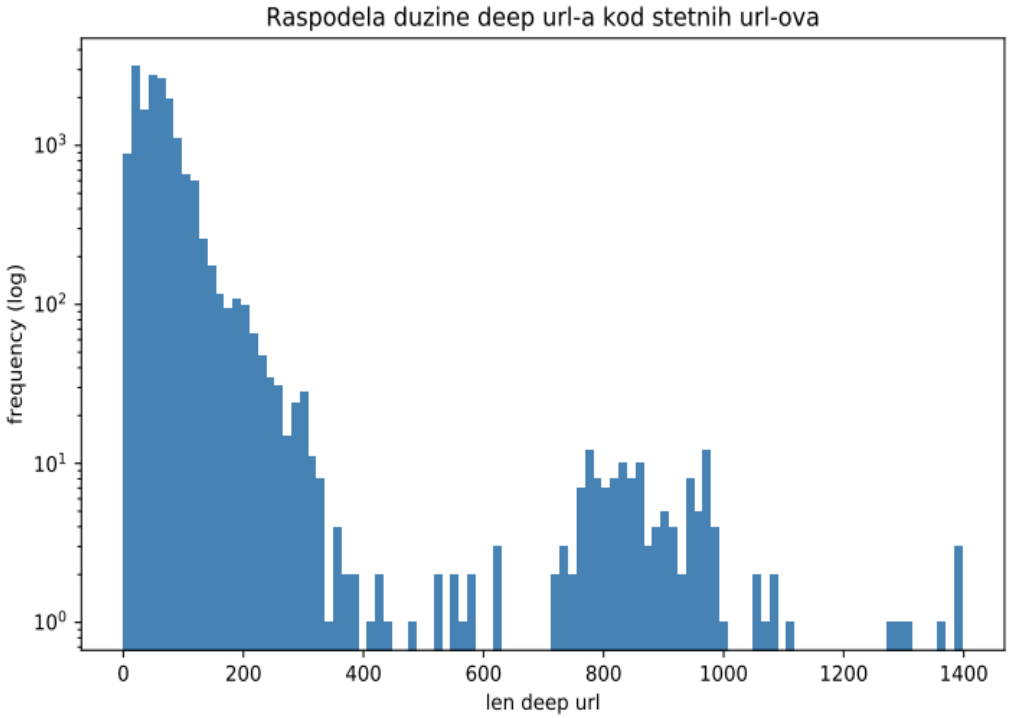
Top level domain - TLD





Deep url


Raspodela duzine deep url-a kod dobrih url-ova
frequency (log)
10^3
10^2
10^1
10^0
0
500
1000
1500
2000
len deep url


Raspodela duzine deep url-a kod stetnih url-ova
frequency (log)
10^3
10^2
10^1
10^0
0
200
400
600
800
1000
1200
1400
len deep url

Multinomial Naive Bayes

- **Deep url je parsiran na reči razdvojene delimiter-ima, regex: ‘|\.|\/|\/\/|:|-|_|%|\?|=|;|<|>|~|\$|&|\+’**
- **Reči su prvo procesirane koristeći CountVectorizer**
- **Verovatnoće predviđene MNB su korišćene kao atributi u daljoj obradi podataka**
- **Accuracy score MNB: ~0.9592**

Random Forest

- **Test set nam čini 10% seta podataka**
- **Za različite vrednosti parametra max_depth dobijamo:**

	train accuracy	test accuracy
max_depth = 10	99,23%	99,14%
max_depth = 20	99,80%	99,39%
max_depth = 30	99,96%	99,36%

Linear SVM

- **Za različite vrednosti parametra Cost dobijamo:**

	train accuracy	test accuracy
C = 1	97,24%	96,83%
C = 10	95,01%	94,53%

SVM sa RBF kernelom

- **Za različite vrednosti parametara gamma I Cost dobijamo:**

	train accuracy	test acuracy
C = 1, gamma = 0.01	98,26%	98,29%
C = 1, gamma = 100	99,79%	92,43%
C = 10, gamma = 1	99,70%	97,90%

Najbitniji atributi

- **Koristimo feature_importances_ iz RandomForest modela**

label_proba_1	0.3803
label_proba_0	0.3587
len deep url	0.1386
len of domain	0.0405
suffix_co.uk	0.0177
suffix_net	0.0119

Predviđanje bez MultinomialNB

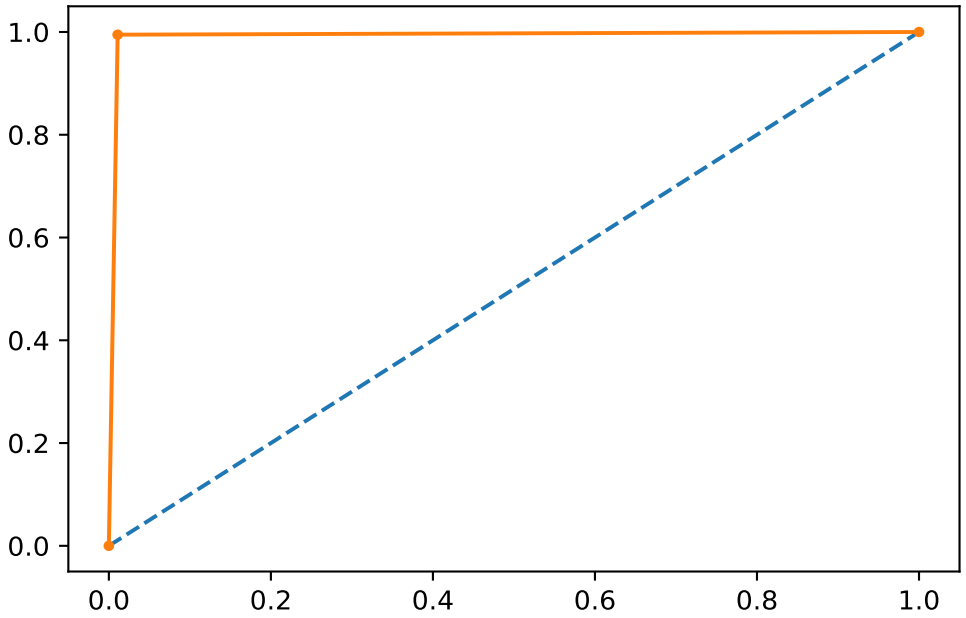
- **Isključujemo dodatne kolone sa predviđenim verovatnoćama iz MNB:**

	train accuracy	test accuracy
RandomForest	95,48%	92,96%
LinearSVM	65,35%	63,77%
RBF SVM	88,76%	87,46%

Ocena klasifikacije

Najbolji rezultat dobijamo koristeći RandomForest sa max_depth = 30

- **ROC kriva I matrica konfuzije za RF**



	klasif. ne spam	klasif. spam
nije spam	TN=1135	FP=9
spam	FN=8	TP=1664

Zaključak

- **Prikazano je nekoliko algoritama za klasifikaciju, od kojih se najbolje ponašaju RandomForest I RBF SVM**
- **Najvažniji atributi su predikcije verovatnoća iz MultinomialNB I LinearSVM se naročito loše ponaša u slučaju kada ih ne koristimo**
- **Spam je brzo rastući problem, jer se stalno otkrivaju nove tehnike I ako ih ne pratimo I analiziramo, sve ih je teže detektovati**
- **Mašinsko učenje u ovoj oblasti ima izuzetan značaj I mnogo prostora za unapređenje u rešavanju sličnih problema**

Kraj

-
-
-
- **Hvala na paznji!**