

Laboratorijska vježba br. 12 Sentimentalna Analiza Teksta Koristeći Web API Service

Uputstvo za izradu laboratorijske vježbe

Izrada laboratorijske vježbe vrši se u formi izvještaja koja je data u nastavku. Potrebno je popuniti sva polja data u izvještaju, odgovoriti na pitanja i dodati tražene slike. Nije dozvoljeno brisati postojeća, niti dodavati nova polja.

Izvještaj sa izradom laboratorijske vježbe pretvorene u PDF dokument šalje se na e-mail adresu odgovornog asistenta grupe za laboratorijske vježbe.

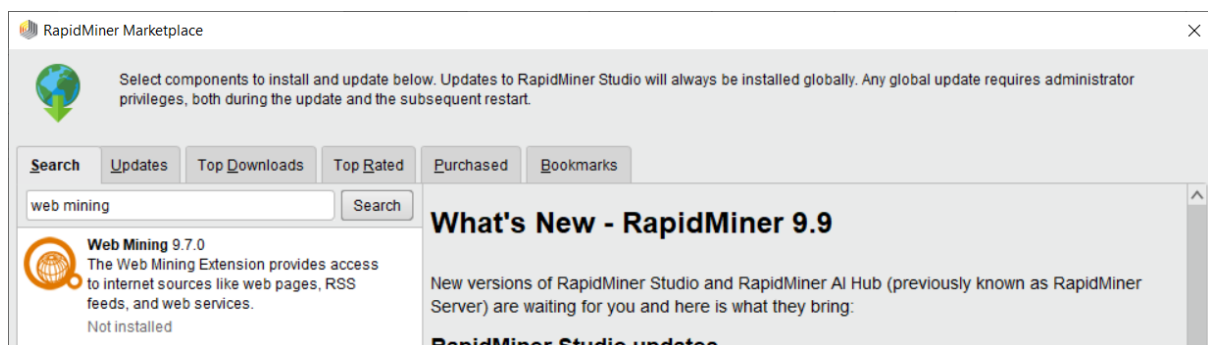
NAPOMENA: Izvještaj se radi samostalno. Rad u paru ili grupi nije dozvoljen.

Informacije o studentu

Ime i prezime: Haris Masovic
Broj indexa: 1689/17993
Grupa za laboratorijske vježbe: Utorak 17:00

Zadatak 1. (Priprema okruženja)

Pokrenuti RapidMiner okruženje, a zatim u glavnom meniju selektovati opciju **Extensions Marketplace**. U polje za pretragu unijeti pojam „web mining“, nakon čega će se prikazati ekstenzija **Web Mining**, kao što je prikazano na slici ispod. Selektovati i instalirati ekstenziju.



Na [sljedećem linku](#) izvršiti registraciju i pretplatiti se na besplatni web API servis **Text analysis** koji vrši automatsko određivanje sentimenta teksta proslijeđenog kao parametar. Sačuvati ključ dobiven nakon pretplate na servis, a koji je postavljen kao parametar za zahtjeve koji se nalaze kao primjer na stranici za testiranje servisa (ista stranica gdje je izvršena pretplata).

Preuzeti dataset koji se nalazi na sljedećem linku: [link](#), a koji sadrži komentare korisnika na filmove sa stranice IMDb.com. Oznaka 0 znači da je komentar negativan, a oznaka 1 da je komentar pozitivan.

Zadatak 2. (Sentimentalna analiza statičkog teksta)

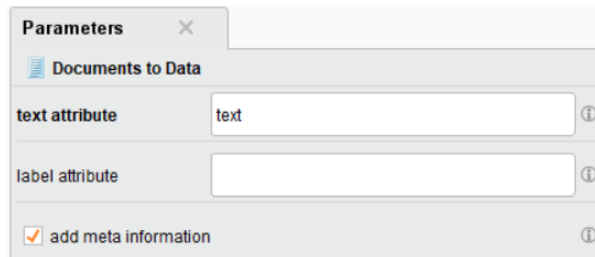
Kako bi se provjerio način na koji radi specificirani web API, prvo će se definisati nekoliko instanci statičkog teksta nad kojima će se izvršiti sentimentalna analiza.

Otvoriti RapidMiner okruženje i odabrati opciju za kreiranje praznog modela.

Korištenjem operatora **Create document** dodati prvi statički tekst u model. Dvostrukim klikom na komponentu otvoriti će se prozor. U prozor unijeti rečenicu „I love you.“

Dodati još jedan statički tekst na isti način. Pritom unijeti rečenicu „I hate you.“

Dodati operator **Documents to Data**. Kao ulaz dodati dva prethodno kreirana dokumenta. Selektovati komponentu, a zatim s desne strane u prozoru Parameters specificirati vrijednost za parametar text attribute i postaviti ga na **text**, kao što je prikazano na slici ispod.

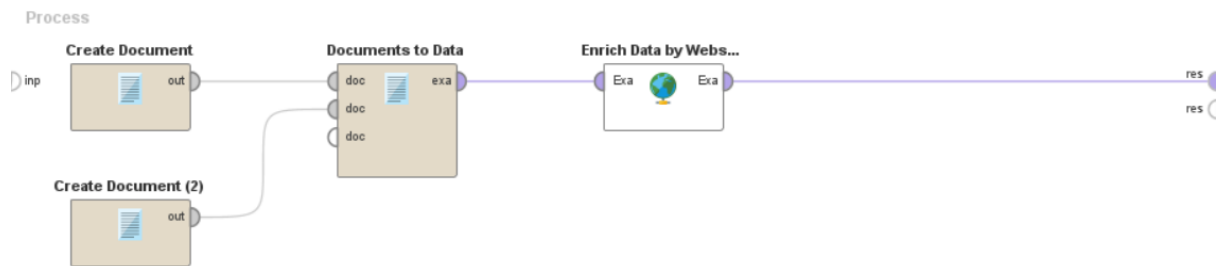


Dodati operator **Enrich Data by Webservice**. Specificirati parametre na sljedeći način:

query type	<i>string matching</i>		
string matching queries	attribute name	query expression	
	<i>pos</i>	<i>„pos“:</i>	<i>,</i>
	<i>neg</i>	<i>„neg“:</i>	<i>,</i>
	<i>neu</i>	<i>„neu“:</i>	<i>,</i>
attribute type	<i>Nominal</i>		
request method	<i>POST</i>		
service method			
body	<pre>{ "language": "english", "text": "<%text%>" }</pre>		
url	<i>https://text-analysis12.p.rapidapi.com/sentiment-analysis/api/v1.1</i>		
separator			
delay	<i>0</i>		
request properties	property	value	
	<i>content-type</i>	<i>application/json</i>	
	<i>x-rapidapi-key</i>	<i>ključ koji se nalazi na stranici gdje se izvršila pretplata na servis</i>	



	<i>x-rapidapi-host</i>	<i>text-analysis12.p.rapidapi.com</i>
	<i>useQueryString</i>	<i>true</i>
encoding	<i>SYSTEM</i>	
user agent		

Kao ulaz za komponentu *Enrich Data by Webservice* povezati izlaz iz komponente *Documents to Data*, a izlaz iz komponente postaviti na res sa desne strane modela. Krajnji model bi trebao izgledati kao na slici ispod.



Pokrenuti proces klikom na tipku **Play**. Cijeli proces trebao bi biti završen za nekoliko sekundi, nakon čega se otvara **Results** tab.

Prikaz tabele sa rezultatima:

Open in  Turbo Prep  Auto Model

Row No.	text	pos	neg	neu
1	I love you.	0.808	0.0	0.192
2	I hate you.	0.0	0.787	0.213

Šta se može zaključiti iz rezultujućih vrijednosti kolona *pos*, *neu* i *neg*? Da li vrijednosti ovih parametara odgovaraju očekivanjima? Obrazložiti odgovor.

- Vrijednosti odgovaraju očekivanim rezultatima jer blize 0 odgovara negativnoj sentiment analizi, dok blize 1 odgovara pozitivnoj sentiment analizi.

Zadatak 2. (Sentimentalna analiza dataseta)

Sada je potrebno izvršiti sentimentálnu analizu nad prethodno preuzetim datasetom. U tu svrhu, prvo je potrebno definisati operator **Read CSV**. Selektovati lokaciju na kojoj se nalazi preuzeti dataset, a zatim slijediti korake bez promjene parametara.

Prikaz tabele sa učitanim podacima:

Format your columns.

Date format: ☐ Replace errors with missing values ⓘ

text	label
polynomial	integer
1 A very, very, very slow-moving, aimless movie about a distressed, drifting y...	0
2 Not sure who was more lost - the flat characters or the audience, nearly half...	0
3 Attempting artiness with black & white and clever camera angles, the movie ...	0
4 Very little music or anything to speak of.	0
5 The best scene in the movie was when Gerardo is trying to find a song that ...	1
6 The rest of the movie lacks art, charm, meaning... if it is about emptiness, it...	0
7 Wasted two hours.	0
8 Saw the movie today and thought it was a good effort, good messages for ki...	1
9 A bit predictable.	0
10 Loved the casting of Jimmy Buffet as the science teacher.	1
11 And those baby owls were adorable.	1
12 The movie showed a lot of Florida as it is best, made it look very appealing.	1
13 The Songs Were The Best And The Muppets Were So Hilarious.	1
14 It Was So Cool.	1
15 This is a very right on case movie that delivers everything almost right in you...	1
16 It had some average acting from the main person, and it was a low budget a...	0
17 This review is long overdue, since I consider A Tale of Two Sisters to be the ...	1
18 I will put this gem up against any movie in terms of screenplay, cinematogra...	1
19 It is practically perfect in all of them - a true masterpiece in a sea of faux ...	1

no problems.

Postojeća komponenta **Enrich Data by Webservice** može se kopirati desnim klikom i odabirom opcije Copy, a zatim Paste, kako bi se izbjeglo ponovno specificiranje parametara. Nad ovom komponentom nije potrebno vršiti nikakve promjene – samo se kao ulaz u komponentu povezuje izlaz iz komponente Read CSV.

Sada je potrebno dodati operator **Generate Attributes**, koji će na osnovu vrijednosti pos, neu i neg kolona generisati labele za redove dataseta. Kao parametar ove komponente potrebno je postaviti funkciju koja generiše labelu na osnovu postojećih vrijednosti kao na slici ispod. Kao ulaz za ovu komponentu potrebno je povezati izlaz iz komponente Enrich Data by Webservice.

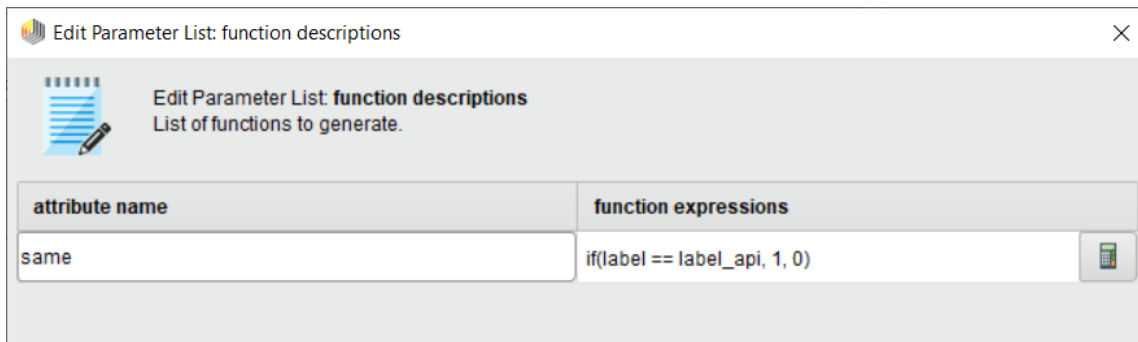
Edit Parameter List: function descriptions

Edit Parameter List: function descriptions
List of functions to generate.

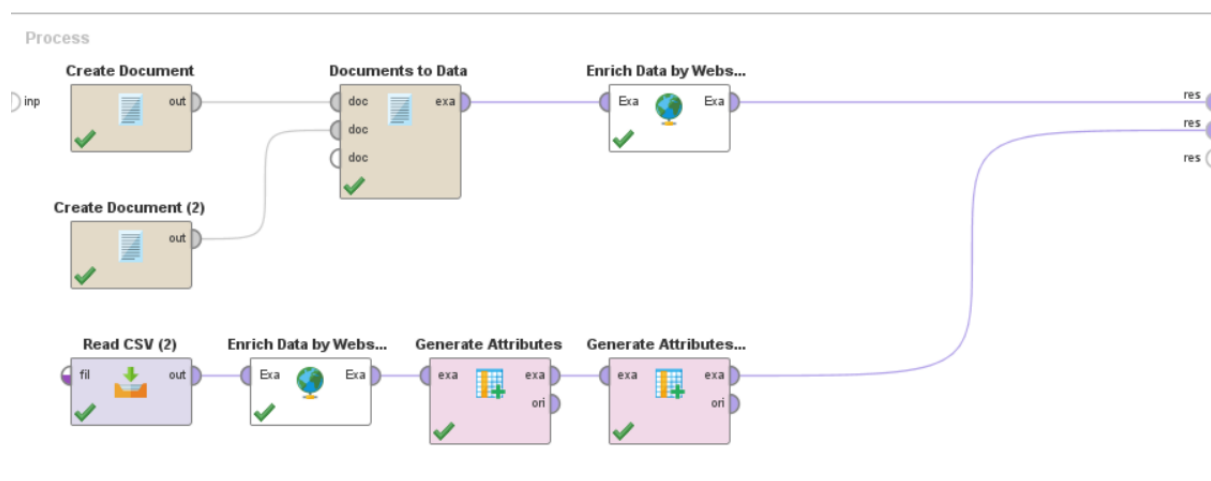
attribute name	function expressions
label_api	if(neg > (pos), 0, 1)

Izlaz iz ovih komponenti su sve postojeće kolone dataseta i nova kolona **label_api** sa labelama dobivenim koristeći web API servis. Još je neophodno izvršiti usporedbu da li su labele iz dataseta jednake labelama dobivenima koristeći web API servis, što se čini dodavanjem još jednog operatora **Generate Attributes**.

Ova komponenta dodati će novu kolonu koja će imati vrijednost 1 ukoliko su vrijednosti u kolonama label i label_api jednake, a 0 ukoliko je suprotno. Funkcija se generiše na sličan način kao u prethodnom koraku i prikazana je na slici ispod.



Kao ulaz u ovu komponentu potrebno je povezati izlaz iz prethodne Generate Attributes komponente, a izlaz iz komponente postaviti na res sa desne strane modela. Krajnji model bi trebao izgledati kao na slici ispod.



Pokrenuti proces klikom na tipku **Play**. Cijeli proces trebao bi biti završen za minutu, nakon čega se otvara **Results** tab, koji sadrži rezultate iz prethodnog zadatka i nove rezultate.

Prikaz tabele sa rezultatima za cijeli dataset:

Open in Turbo Prep Auto Model Filter (100 / 100 examples): all

Row No.	text	label	pos	neg	neu	label_api	same
1	A very, very...	0	0.0	0.219	0.781	0	1
2	Not sure wh...	0	0.0	0.222	0.778	0	1
3	Attempting ...	0	0.083	0.25	0.667	0	1
4	Very little m...	0	0.0	0.0	1.0	1	0
5	The best sc...	1	0.181	0.0	0.819	1	1
6	The rest of t...	0	0.126	0.16	0.714	0	1
7	Wasted two ...	0	0.0	0.615	0.385	0	1
8	Saw the mov...	1	0.326	0.0	0.674	1	1
9	A bit predict...	0	0.0	0.0	1.0	1	0
10	Loved the c...	1	0.302	0.0	0.698	1	1
11	And those b...	1	0.39	0.0	0.61	1	1
12	The movie s...	1	0.231	0.0	0.769	1	1
13	The Songs ...	1	0.444	0.0	0.556	1	1
14	It Was So Co...	1	0.464	0.0	0.536	1	1
15	This is a ver...	1	0.0	0.0	1.0	1	1
16	It had some ...	0	0.124	0.096	0.78	1	0
17	This review l...	1	0.189	0.0	0.811	1	1
18	I will put thi...	1	0.0	0.0	1.0	1	1
19	It is practica...	1	0.562	0.0	0.438	1	1
20	The structur...	1	0.138	0.0	0.862	1	1

ExampleSet (100 examples, 0 special attributes, 7 regular attributes)

Šta se na prvi pogled može zaključiti pogledom na kolonu *label* i usporedbom sa kolonama *pos*, *neu* i *neg*? Da li instance koje imaju labelu 0 imaju veliku negativnost i obrnuto? Objasniti par primjera iz *dataseta*.

- Može se vidjeti korelacija sa *neu* i *pos* vrijednostima odnosno sa label 1 vrijednostima.
- Nema ju instance veliku negativnu vrijednost za labelu 0, a ni za labelu 1 nemaju veliku pozitivnost.
- Primjer se može pokazati na primjeru 1 ispod.

Pronaći par primjera u *datasetu* za koju je vrijednost u koloni *same* jednaka 0.

Primjer 1

Tekst instance: Very little music or anything to speak of.

Vrijednost u *pos*, *neu* i *neg* kolonama: 0.0, 1.0, 0.0.

Vrijednost manuelne labele, vrijednost web API labele: 0, 1

Zbog čega je instanca drukčije labelirana koristeći web API servis? Koja labela je tačnija i zašto? Objasniti odgovor. Zato što je *neu* jednak 1 i servis je vratio 1 kao rezultat, dok je u *datasetu* 0. Tesko je reci koja je tačnija jer se može presuditi na obje strane.

Primjer 2

Tekst instance: It had some average acting from the main person, and it was a low budget as you clearly can see.

Vrijednost u *pos*, *neu* i *neg* kolonama: 0.124, 0.78, 0.096

Vrijednost manuelne labele, vrijednost web API labele: 0, 1.

Zbog čega je instanca drukčije labelirana koristeći web API servis? Koja labela je tačnija i zašto? Objasniti odgovor. - U *datasetu* je predstavljeno kao labela 0, dok je servis vratio 1. Labela se može postaviti 1 za ovaj tekst jer je najviše neutralna, tačnija je vrijednost sa servisa (naravno subjektivno).

Primjer 3

Dubinska Analiza Podataka

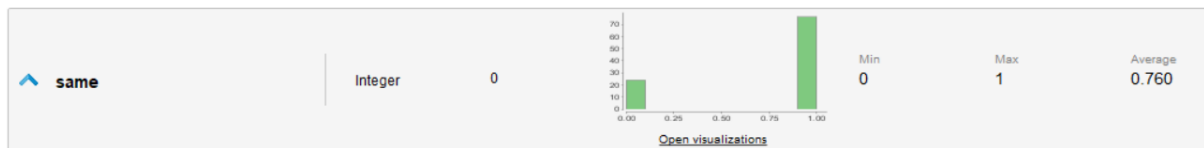
Tekst instance: This short film certainly pulls no punches.

Vrijednost u *pos*, *neu* i *neg* kolonama: 0.25, 0.521, 0.229.

Vrijednost manuelne labele, vrijednost web API labele: 0, 1.

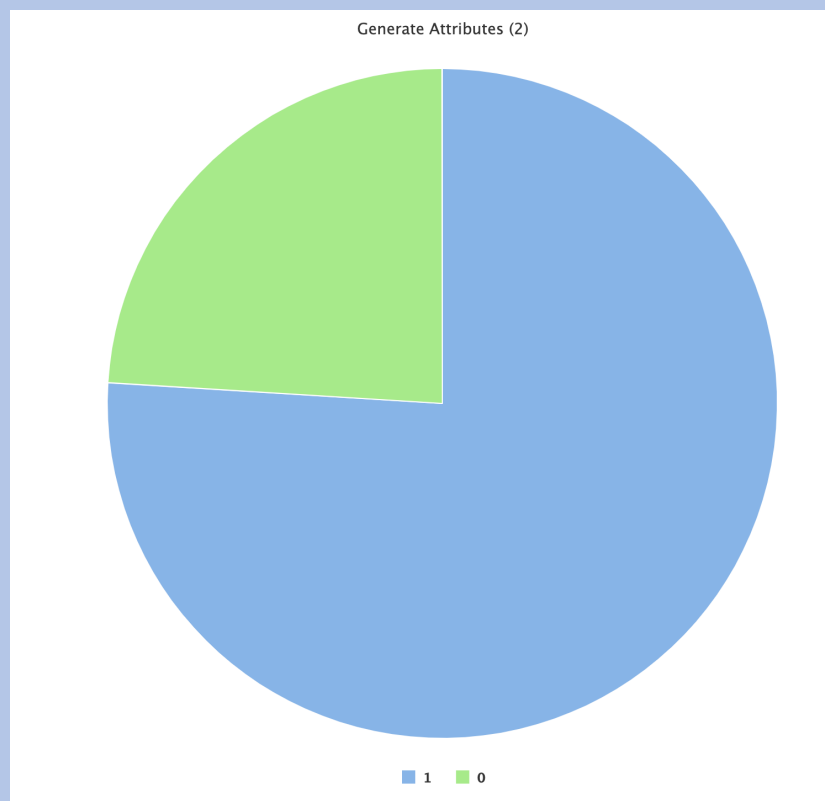
Zbog čega je instanca drukčije labelirana koristeći web API servis? Koja labela je tačnija i zašto? Objasniti odgovor. U datasetu je predstavljen text kao labela 0, servis je vratio 1. Text je najviše neutralan shodno time je dobijena vrijednost 1. Tačnija je pozitivnija labela.

Kliknuti **Statistics** tipku sa lijeve strane ekrana. Kliknuti na **same** labelu, a zatim odabrati opciju **Open Visualizations** kao na slici ispod.



Za **Plot type** odabrati **Pie/Donut**, za **X-Axis Column** i **Value Column** odabrati kolonu **Same**, a zatim odabrati opciju **Aggregate Data** kako bi se rezultati unificirali.

Prikaz grafika:



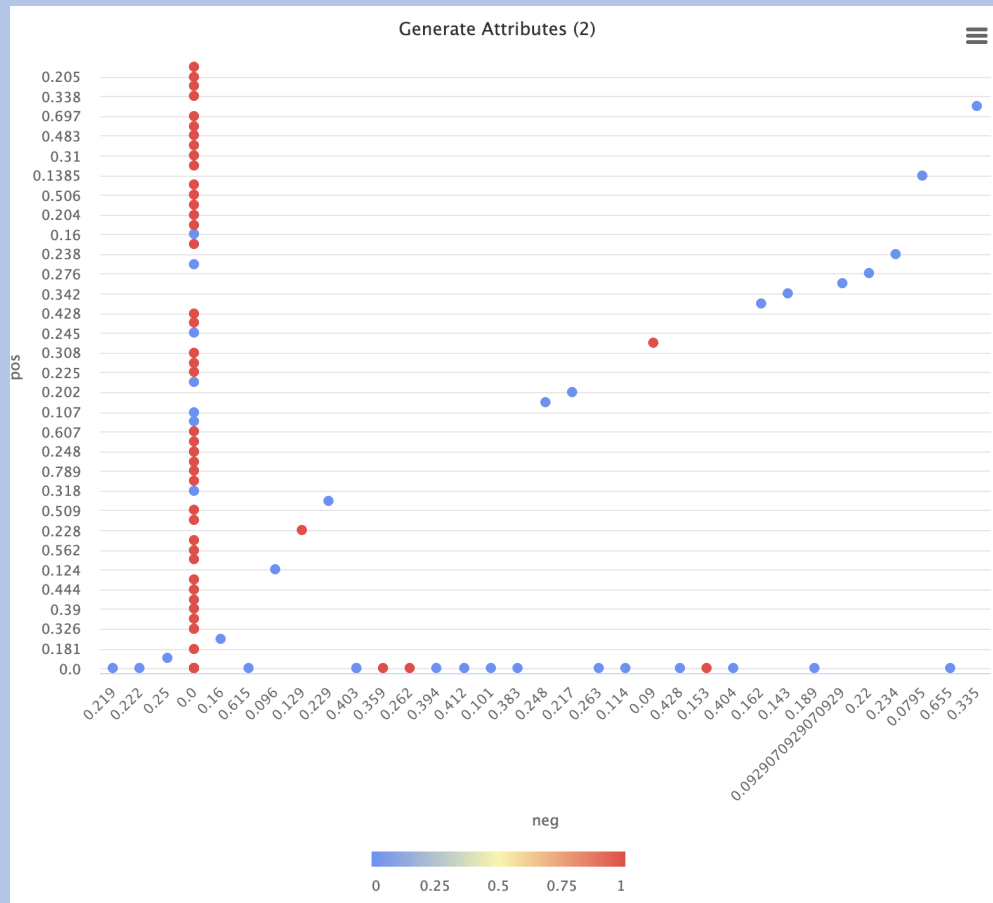
Kakav je odnos između labela koje su labelirane na isti način koristeći web API servis i labela koje nisu labelirane na isti način? Da li je procenat istovjetno labeliranih instanci visok ili ne? Obrazložiti svoj odgovor.

- 76% labela su iste u odnosu na web servis i labela iz dataseta.

- Procenat je relativno visok (76%).

Za **Plot type** odabrati **Scatter/Bubble**, za **X-Axis Column** odabrati kolonu **neg**, za **Value Column** odabrati kolonu **pos**, a za **Color** odabrati kolonu **label_api**.

Prikaz grafika:

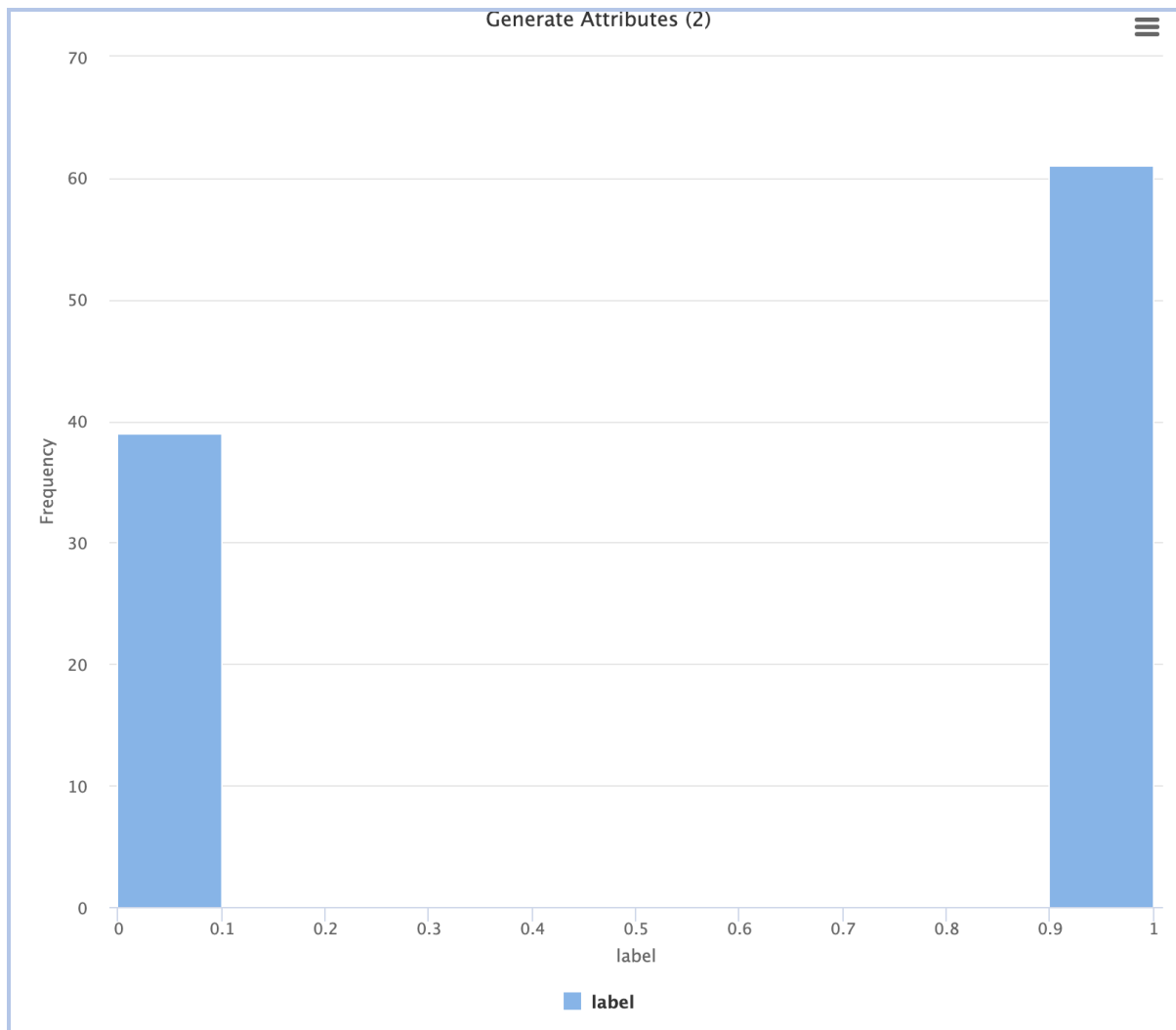


Kakav je odnos između X i Y vrijednosti sa bojom instance? Da li postoji univerzalna logika prema kojoj se labela dodjeljuje na osnovu vrijednosti u *pos* i *neg* kolonama? Kakve veze to ima sa formulom koja je iskorištena u modelu? Objasniti odgovor.

- Veci je broj predstavljeno crvenom bojom tj. vise je vrijednosti negativnih vrijednosti.
- Ne postoji univerzalna logika za dodjeljivanje labela prema pos i neg kolonama.
- Formula je iskoristena da ako je $neg > pos$, da bude vrijednost veka, shodno time ima i vise vrijednosti crvenom bojom.

Vratiti se na **Statistics** tab i sada odabrati vizualizaciju za **label** kolonu. Nije potrebno mijenjati početne postavke za grafik.

Prikaz grafika:



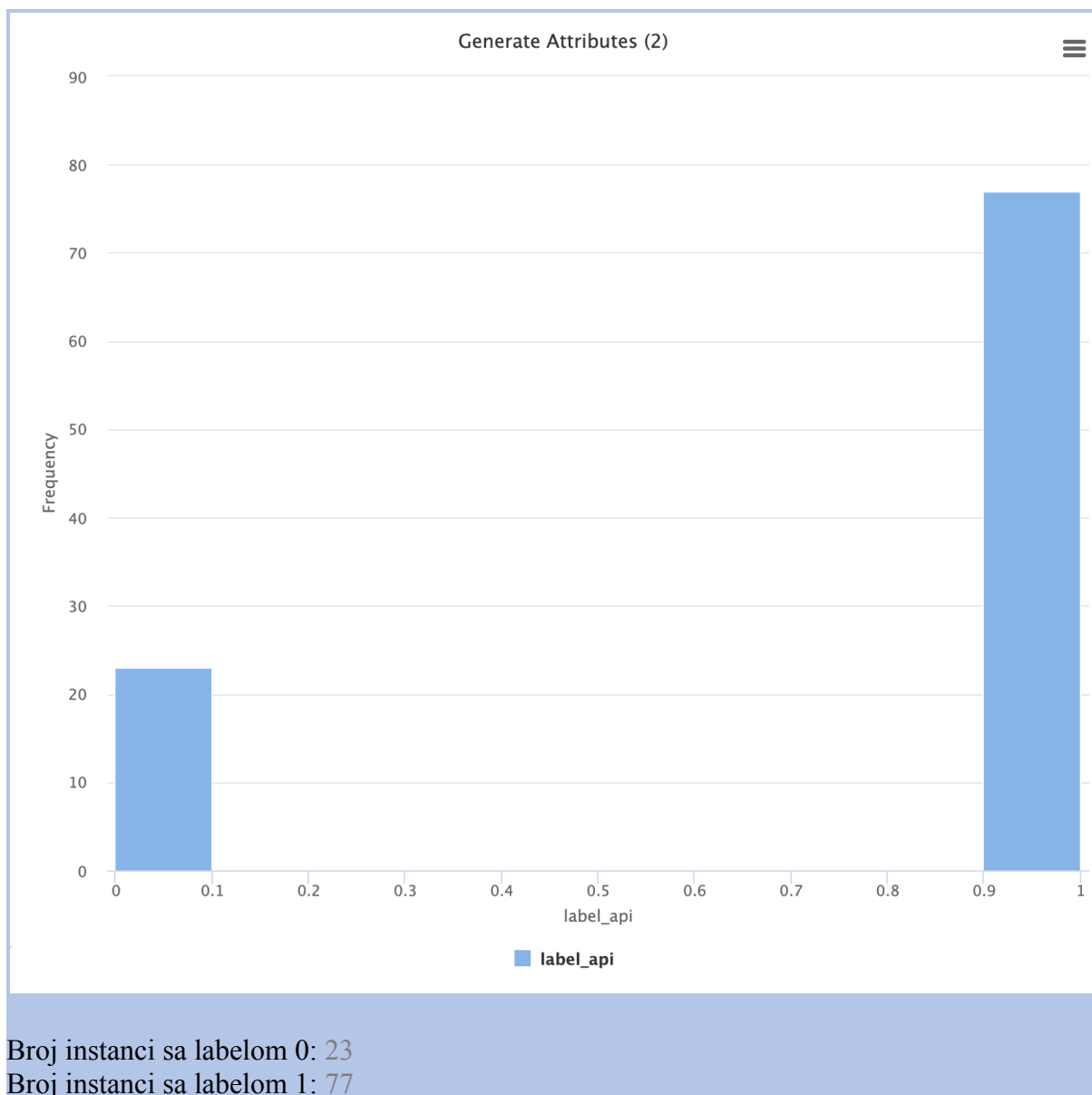
Broj instanci sa labelom 0: 39

Broj instanci sa labelom 1: 61

Vratiti se na **Statistics** tab i sada odabrati vizualizaciju za **label_api** kolonu. Nije potrebno mijenjati početne postavke za grafik.

Prikaz grafika:

Dubinska Analiza Podataka



Formirati konfuzijsku matricu koristeći vrijednosti u tabeli ispod:

label\label_api	0	1
0	19	20
1	4	57

Šta se može zaključiti iz konfuzijske matrice? Da li je veći broj FN ili FP i šta to indicira? Povezati odgovor sa analizom pojedinačnih instanci koje su bile pogrešno labelirane koristeći web API servis.

- Može se zaključiti da je većina labels ispravno odredjena.

- Veci je broj FP i to inicira da su postavljene vrijednosti labels drugacije od dobijenih od strane servisa, ovo je prouzrokovano to sto servis za neutralne vrijednosti daje labelu 1, dok u datasetu su postavljene kao 0, shodno time je veci broj FP.

Da li je korištenje web API servisa za automatsko labeliranje instanci dovoljno pouzdano na osnovu dobivenih rezultata? Posebno obratiti pažnju na veličinu *dataseta* i vrijeme procesiranja u *RapidMiner* okruženju. Detaljno objasniti svoj odgovor.

- Pouzdano je dovoljno, u smislu dobijenih rezultata sentimentalne analize, tj. web API servis je pouzdan za automatsko labeliranje instanci.

- Dataset je male velicine i za dataset te velicine vrijeme procesiranja traje dugo, tako da vrijeme izvršavanja predstavlja manu u ovom pristupu.