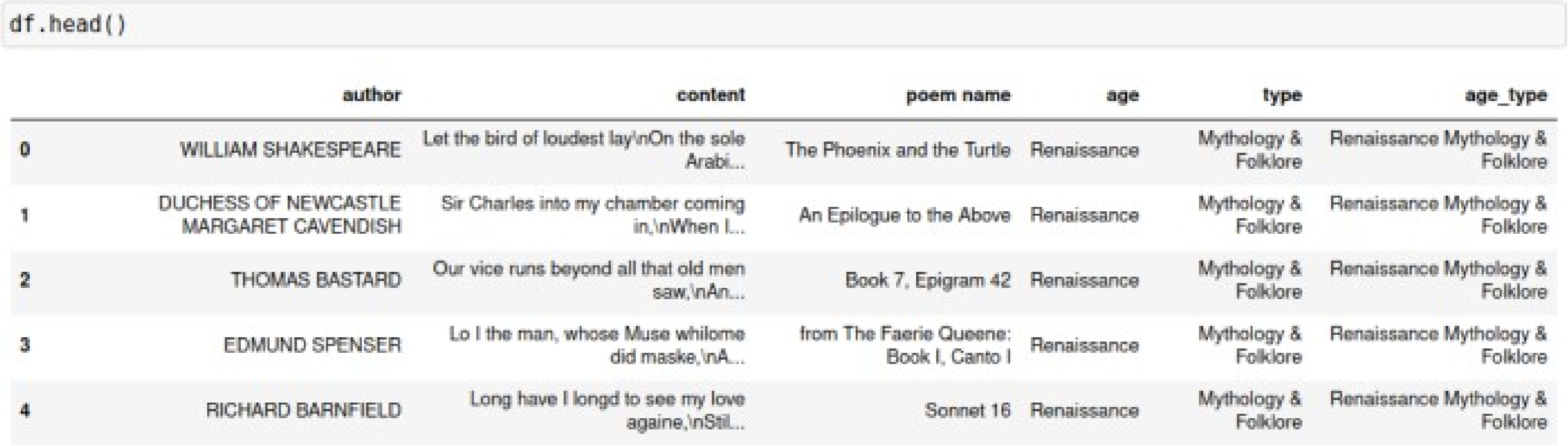
“Modern and renaissance poetry”

Analiza podataka i pretprocesiranje

- 573 instance
- 5 atributa

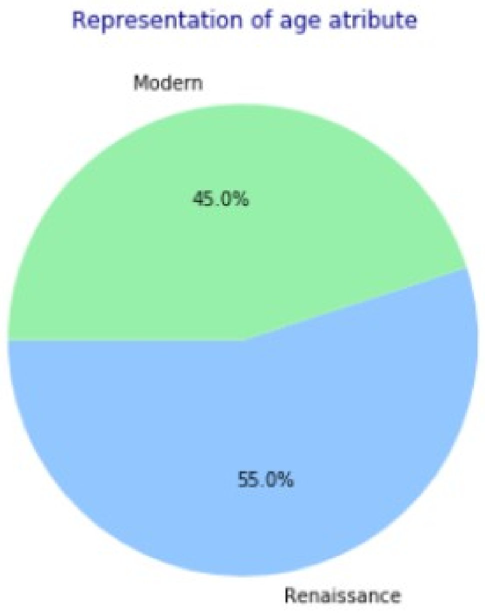


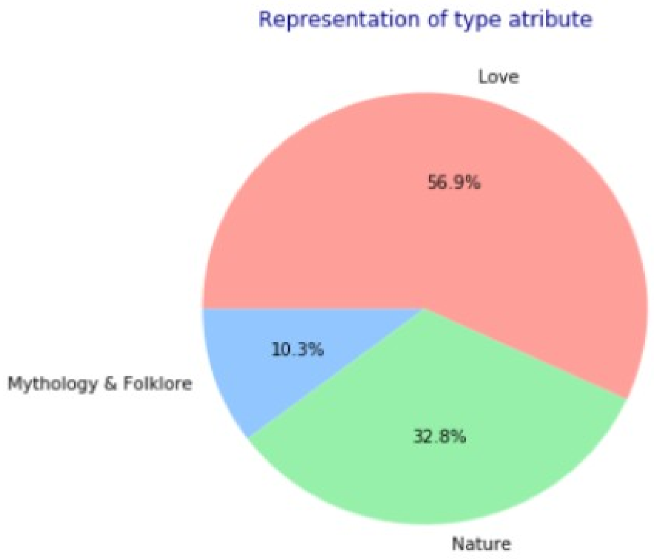
```
df.head()
```

	author	content	poem name	age	type	age_type
0	WILLIAM SHAKESPEARE	Let the bird of loudest lay\nOn the sole Arabi...	The Phoenix and the Turtle	Renaissance	Mythology & Folklore	Renaissance Mythology & Folklore
1	DUCHESS OF NEWCASTLE MARGARET CAVENDISH	Sir Charles into my chamber coming in,\nWhen I...	An Epilogue to the Above	Renaissance	Mythology & Folklore	Renaissance Mythology & Folklore
2	THOMAS BASTARD	Our vice runs beyond all that old men saw,\nAn...	Book 7, Epigram 42	Renaissance	Mythology & Folklore	Renaissance Mythology & Folklore
3	EDMUND SPENSER	Lo I the man, whose Muse whilome did maske,\nA...	from The Faerie Queene: Book I, Canto I	Renaissance	Mythology & Folklore	Renaissance Mythology & Folklore
4	RICHARD BARNFIELD	Long have I longd to see my love againe,\nStil...	Sonnet 16	Renaissance	Mythology & Folklore	Renaissance Mythology & Folklore

Analiza podataka i pretprocesiranje

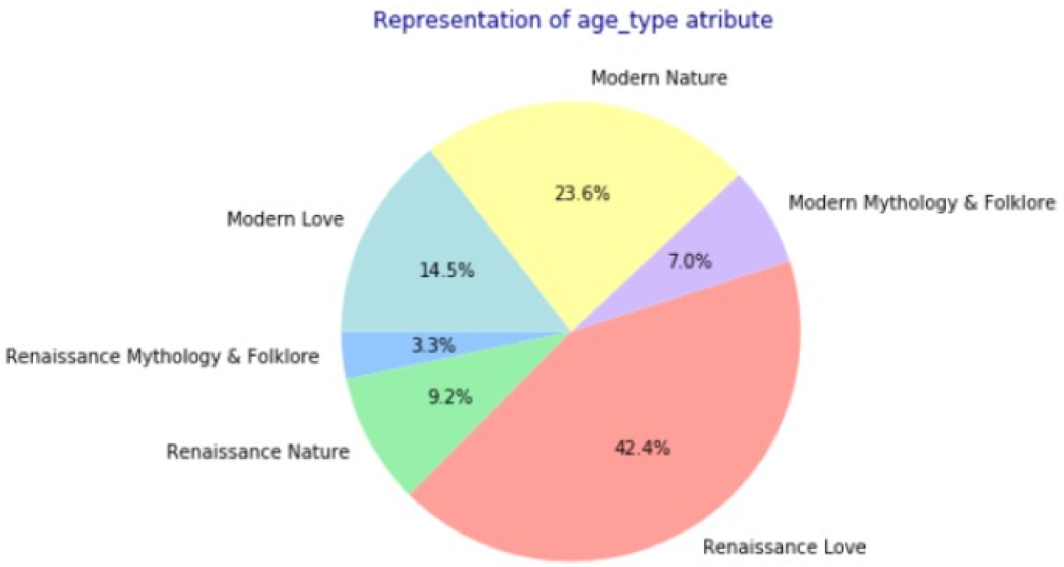
- Zastupljenost klasa





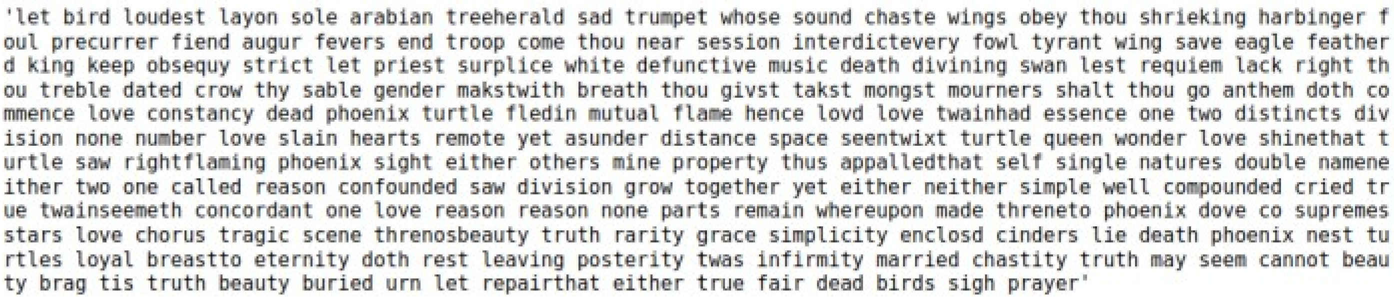
Analiza podataka i pretprocesiranje

- Zastupljenost klasa



Analiza podataka i pretprocesiranje

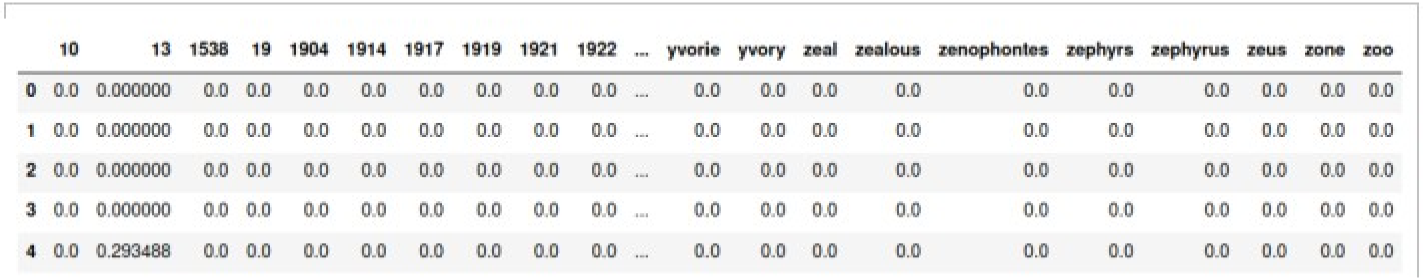
- Nedostajuce vrenosti, duplirani podaci
- 506 instanci
- Transformacija podataka:



```
'let bird loudest layon sole arabian treeherald sad trumpet whose sound chaste wings obey thou shrieking harbinger f
oul precurrer fiend augur fevers end troop come thou near session interdictevery fowl tyrant wing save eagle feather
d king keep obsequy strict let priest surplice white defunctive music death divining swan lest requiem lack right th
ou treble dated crow thy sable gender makstwith breath thou givst takst mongst mourners shalt thou go anthem doth co
mmence love constancy dead phoenix turtle fledin mutual flame hence lovd love twainhad essence one two distincts div
ision none number love slain hearts remote yet asunder distance space seentwixt turtle queen wonder love shinethat t
urtle saw rightflaming phoenix sight either others mine property thus appalledthat self single natures double namene
ither two one called reason confounded saw division grow together yet either neither simple well compounded cried tr
ue twainseemeth concordant one love reason reason none parts remain whereupon made threneto phoenix dove co supremes
stars love chorus tragic scene threnosbeauty truth rarity grace simplicity enclosd cinders lie death phoenix nest tu
rtles loyal breastto eternity doth rest leaving posterity twas infirmity married chastity truth may seem cannot beau
ty brag tis truth beauty buried urn let repairthat either true fair dead birds sigh prayer'
```

Analiza podataka i pretprocesiranje

- TF – IDF matrica :



	10	13	1538	19	1904	1914	1917	1919	1921	1922	...	yvorie	yvory	zeal	zealous	zenophontes	zephyrs	zephyrus	zeus	zone	zoo
0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.293488	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Klasifikacija - KNN

- Na neredukovanim podacima

  k = 23, kosinusna mera
- Na redukovanim podacima – prokletstvo dimenzionalnosti

  k = 23, euklidska mera

Klasifikacija – Naivni Bajes

- Multinominalni Naivni Bajes
- Primenjen na neredukovanom skupu
- Određivanje hiperparametara pohlepnim algoritmom

  alpha = 0.1

Klasifikacija - SVM

- Primena nad neredukovanim podacima
- Određivanje hiperparametara pohlepnim algoritmom

  C = 1, linear

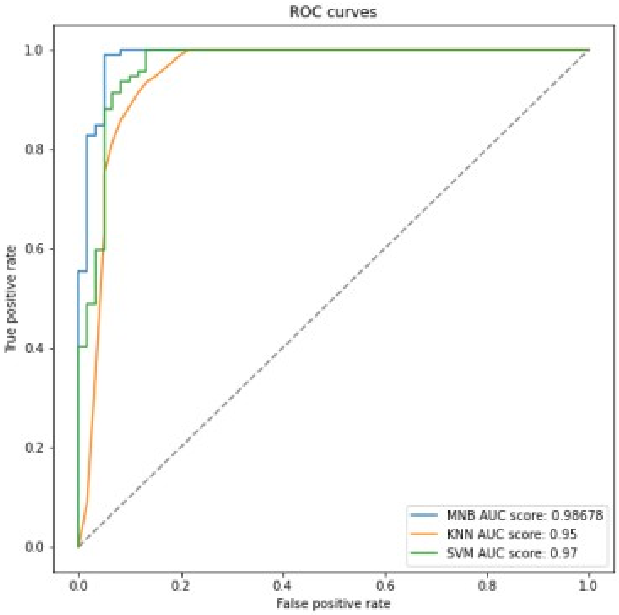
  Preprilagođavanje rešeno C = 0.5
- Primena na redukovanim podacima u cilju vizuelizacije

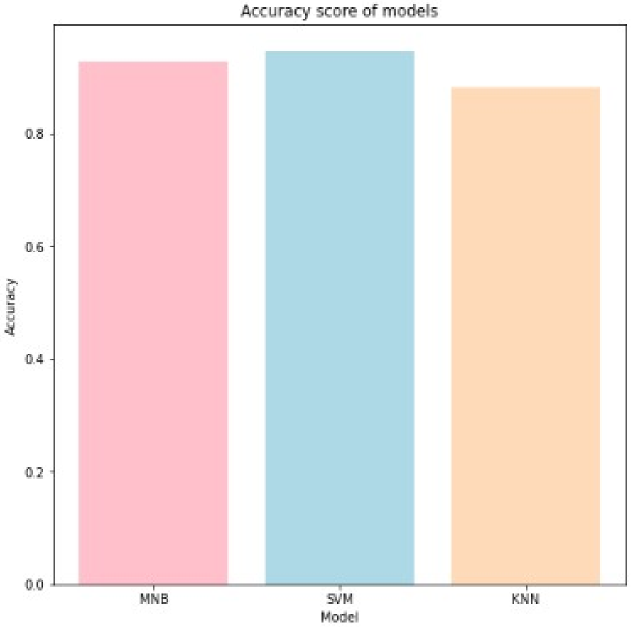
  Promena optimalnih hiperparametara

  C = 1, rbf jezgro

Klasifikacija - poređenje

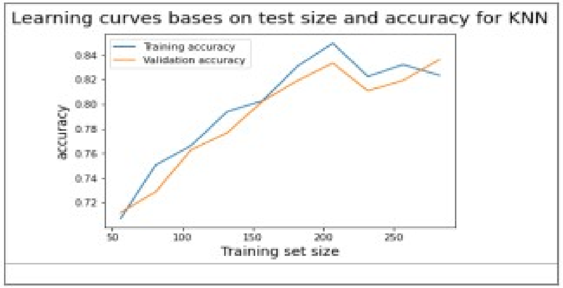
- Roc krive i tačnost:

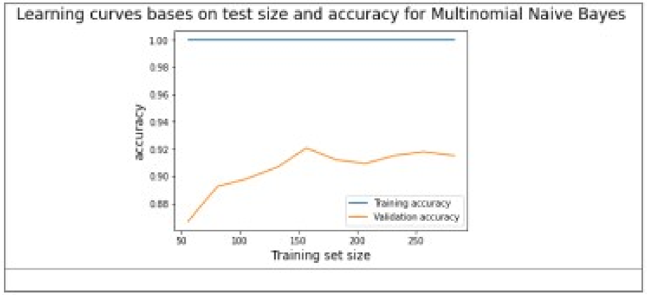


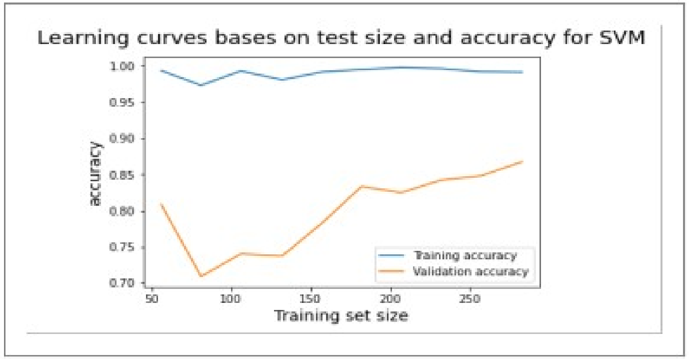


Klasifikacija - poređenje

- Krive učenja:

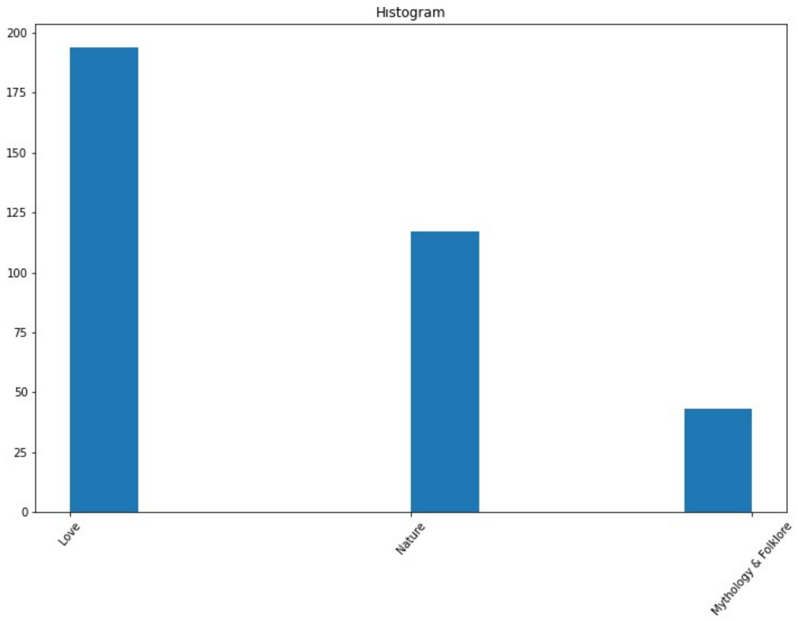

Learning curves bases on test size and accuracy for KNN

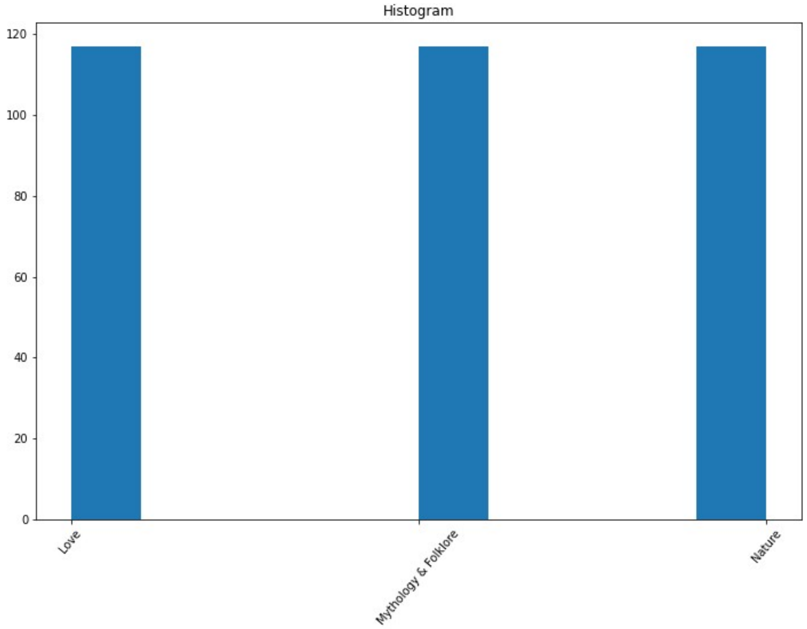

Learning curves bases on test size and accuracy for Multinomial Naive Bayes


Learning curves bases on test size and accuracy for SVM

Klasifikacija – nebalansiranost klasa

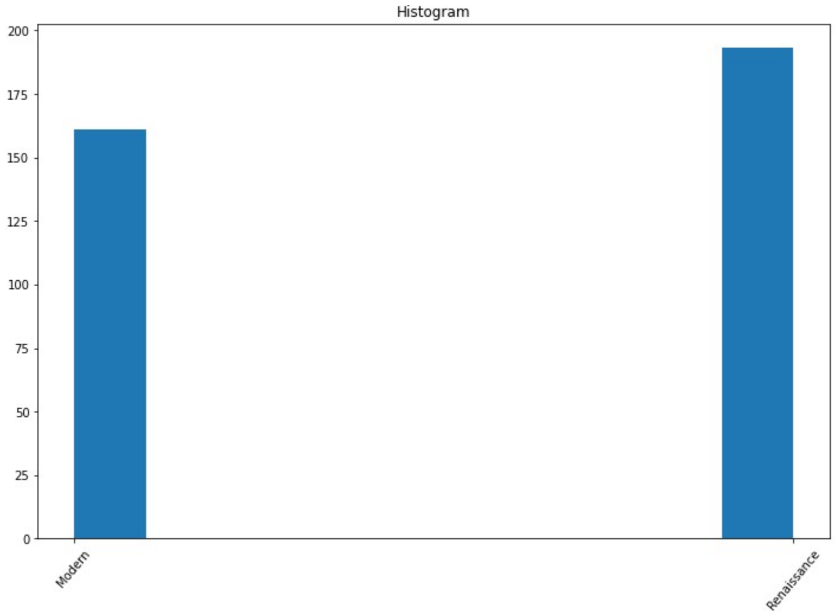
- Atribut type

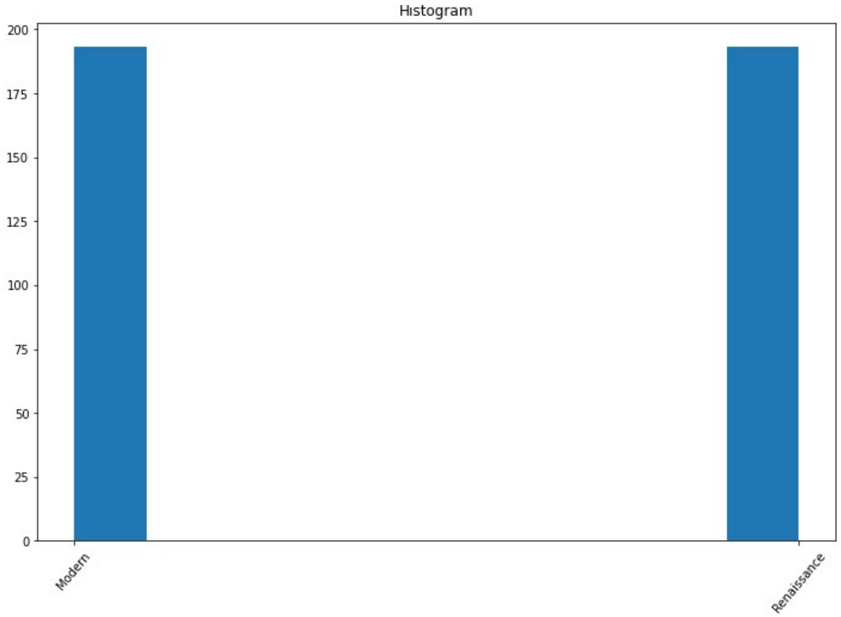




Klasifikacija – nebalansiranost klasa

- Atribut age

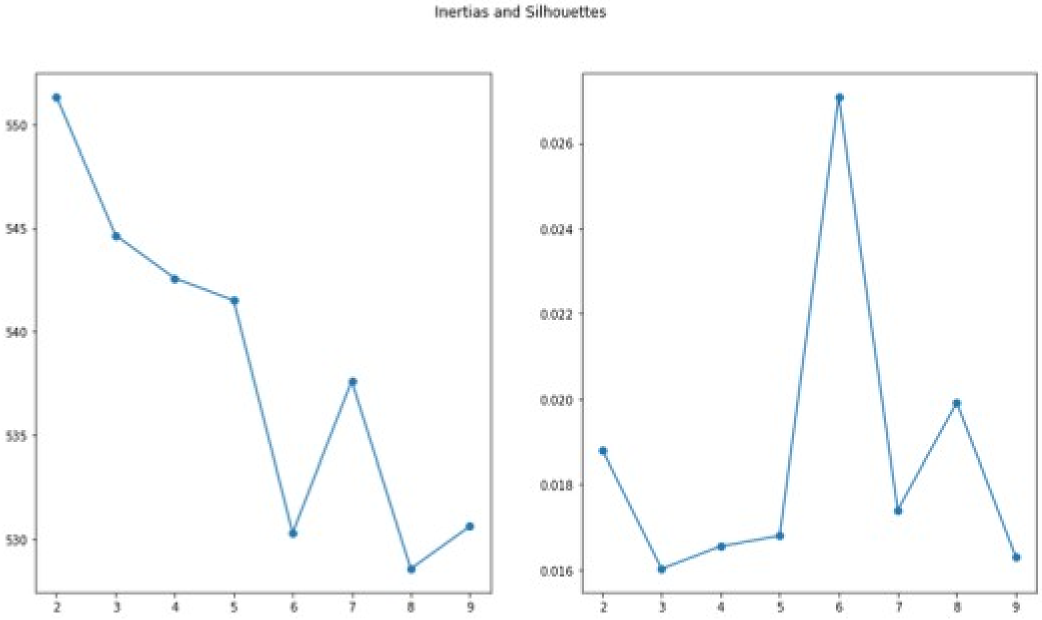




Klasterovanje - kmeans

- Na neredukovanim podacima:

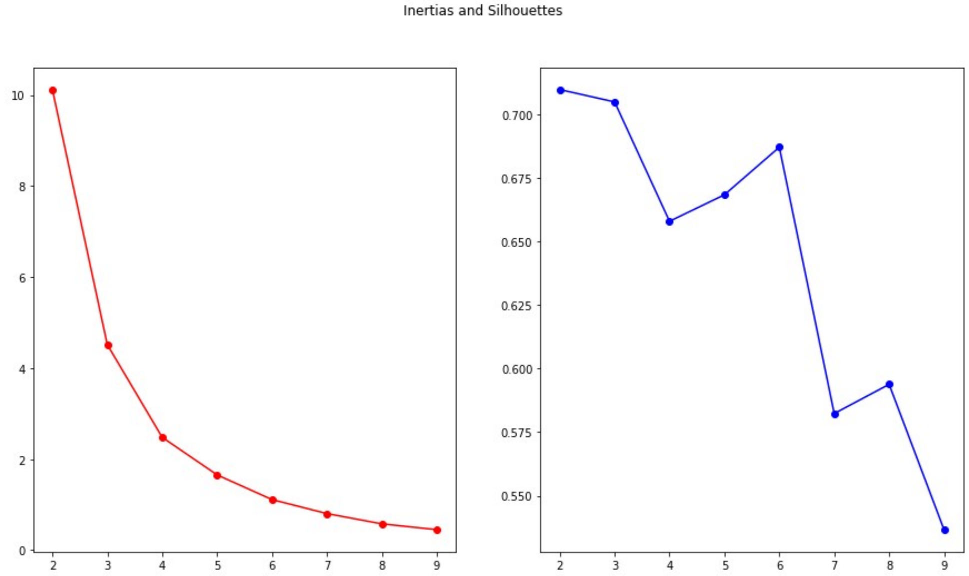
  Prokletstvo dimenzionalnosti

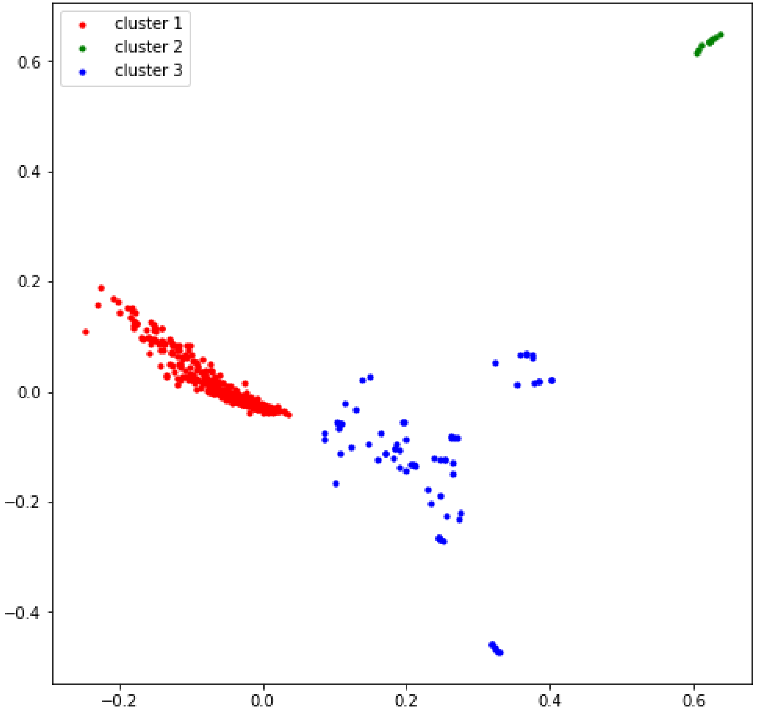


Klasterovanje - kmeans

- Na redukovanim podacima:

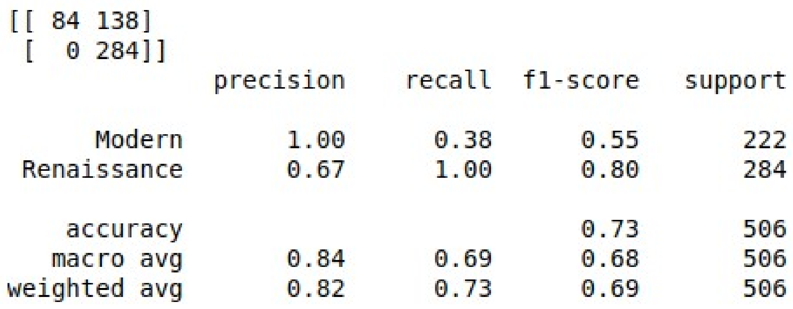
  k=3



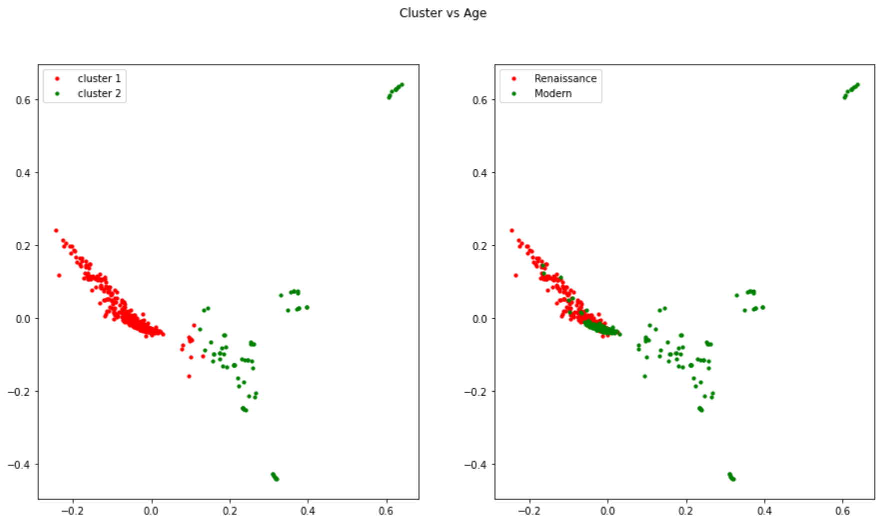


Klasterovanje - kmeans

- Spoljašnji kriterijumi:



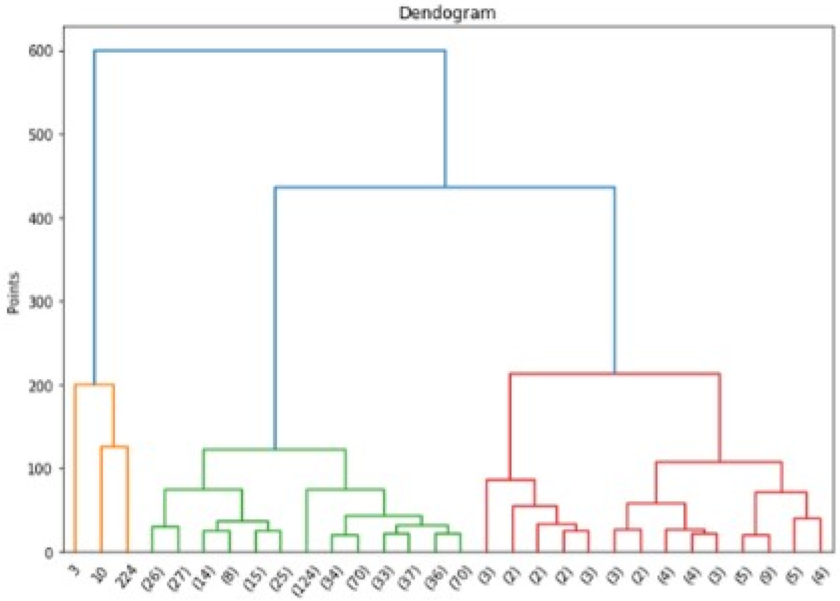
```
[[ 84 138]
 [  0 284]]
              precision    recall  f1-score   support

      Modern       1.00      0.38      0.55       222
 Renaissance       0.67      1.00      0.80       284

    accuracy                           0.73       506
   macro avg       0.84      0.69      0.68       506
weighted avg       0.82      0.73      0.69       506
```



- Klasterovanje kao dobar načina analize pre nadgledanog učenja

Klasterovanje - hijerarhijsko

- Sakupljajuće hijerarhijsko klasterovanje
- Dendogram, Vardov metod



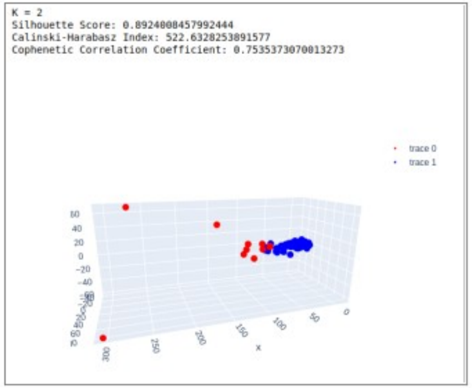
Klasterovanje - hijerarhijsko

- Unutrašnji kriterijumi

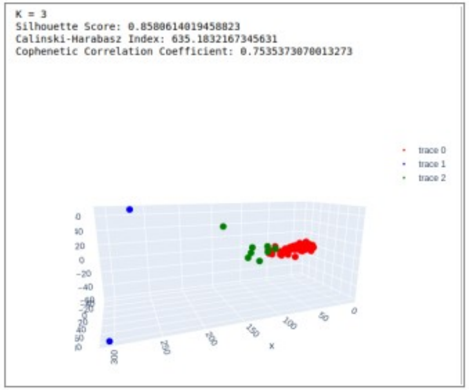
  1. Siluate skor

  2. Kriterijum odnosa varijnse

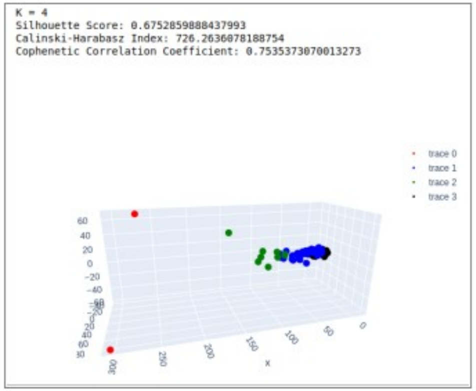
  3. Kofenetski koeficijent korelacije



```
K = 2
Silhouette Score: 0.8924008457992444
Calinski-Harabasz Index: 522.6328253891577
Cophenetic Correlation Coefficient: 0.7535373070013273
```



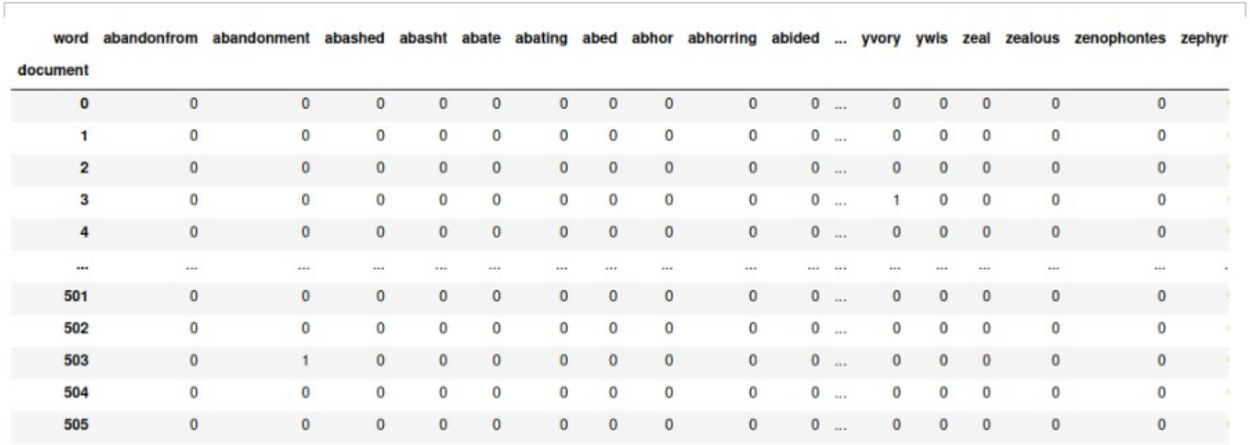
```
K = 3
Silhouette Score: 0.8580614019458823
Calinski-Harabasz Index: 635.1832167345631
Cophenetic Correlation Coefficient: 0.7535373070013273
```



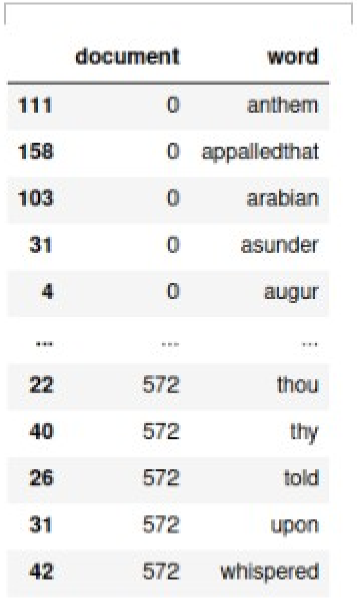
```
K = 4
Silhouette Score: 0.6752859888437993
Calinski-Harabasz Index: 726.2636078188754
Cophenetic Correlation Coefficient: 0.7535373070013273
```

Pravila pridruživanja

- Potrošačka korpa – pretprocesiranje



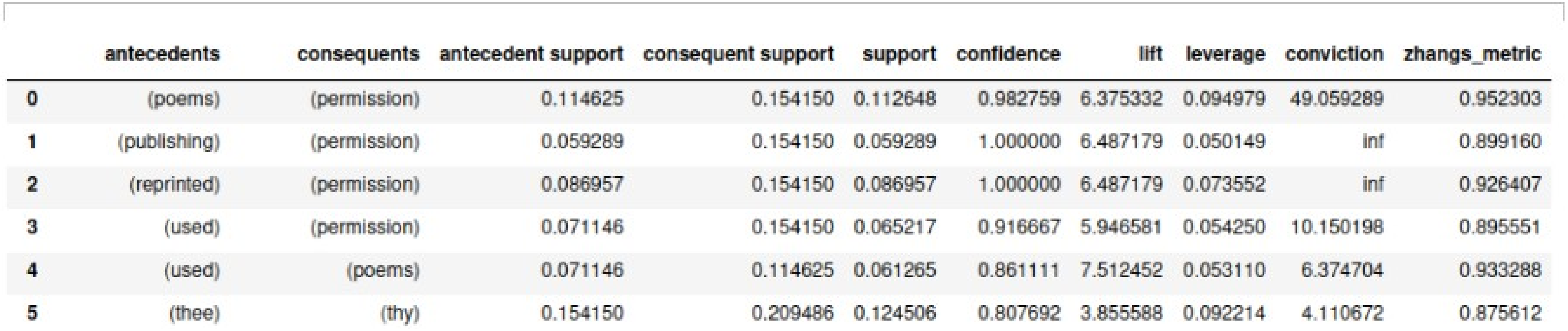
word	abandonfrom	abandonment	abashed	abasht	abate	abating	abed	abhor	abhorring	abided	...	yvory	ywis	zeal	zealous	zenophontes	zephyr
document																	
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
501	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
502	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
503	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
504	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
505	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	



	document	word
111	0	anthem
158	0	appalledthat
103	0	arabian
31	0	asunder
4	0	augur
...	...	...
22	572	thou
40	572	thy
26	572	told
31	572	upon
42	572	whispered

Pravila pridruživanja

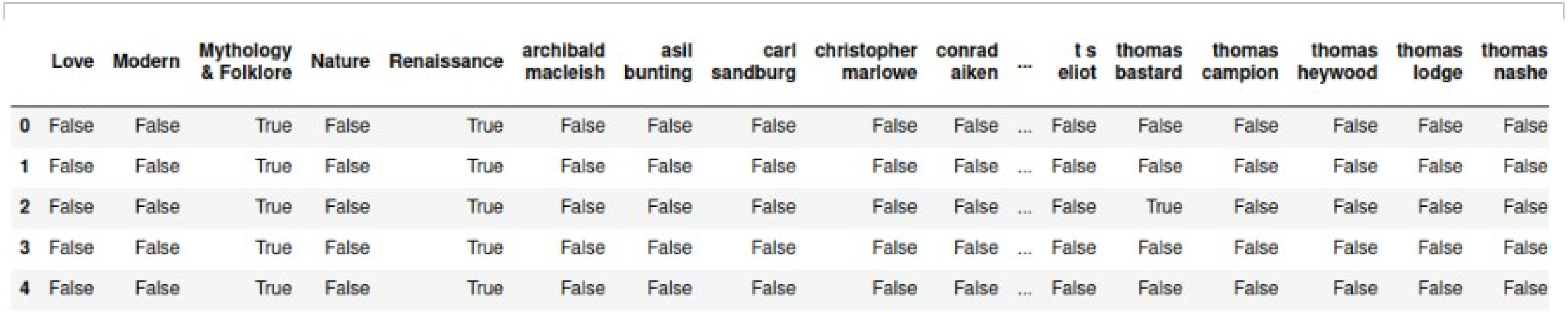
- Potrošačka korpa – pravila



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(poems)	(permission)	0.114625	0.154150	0.112648	0.982759	6.375332	0.094979	49.059289	0.952303
1	(publishing)	(permission)	0.059289	0.154150	0.059289	1.000000	6.487179	0.050149	inf	0.899160
2	(reprinted)	(permission)	0.086957	0.154150	0.086957	1.000000	6.487179	0.073552	inf	0.926407
3	(used)	(permission)	0.071146	0.154150	0.065217	0.916667	5.946581	0.054250	10.150198	0.895551
4	(used)	(poems)	0.071146	0.114625	0.061265	0.861111	7.512452	0.053110	6.374704	0.933288
5	(thee)	(thy)	0.154150	0.209486	0.124506	0.807692	3.855588	0.092214	4.110672	0.875612

Pravila pridruživanja

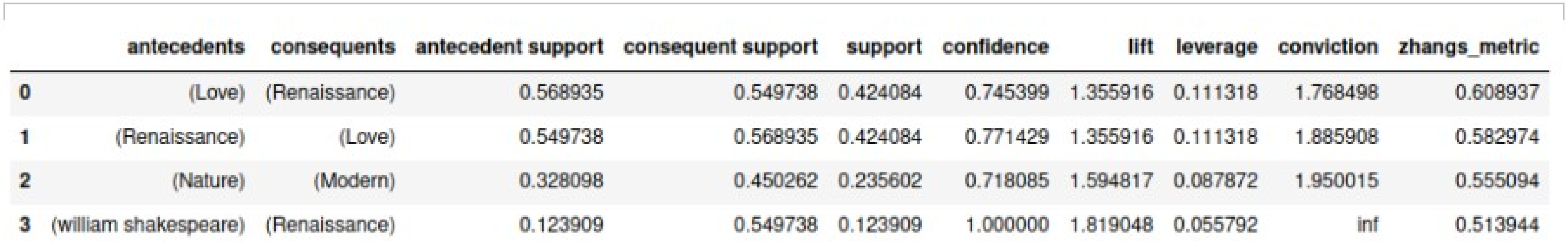
- Kategorički atributi – pretprocesiranje



	Love	Modern	Mythology & Folklore	Nature	Renaissance	archibald macleish	asil bunting	carl sandburg	christopher marlowe	conrad aiken	...	t s eliot	thomas bastard	thomas campion	thomas heywood	thomas lodge	thomas nashe
0	False	False	True	False	True	False	False	False	False	False	...	False	False	False	False	False	False
1	False	False	True	False	True	False	False	False	False	False	...	False	False	False	False	False	False
2	False	False	True	False	True	False	False	False	False	False	...	False	True	False	False	False	False
3	False	False	True	False	True	False	False	False	False	False	...	False	False	False	False	False	False
4	False	False	True	False	True	False	False	False	False	False	...	False	False	False	False	False	False

Pravila pridruživanja

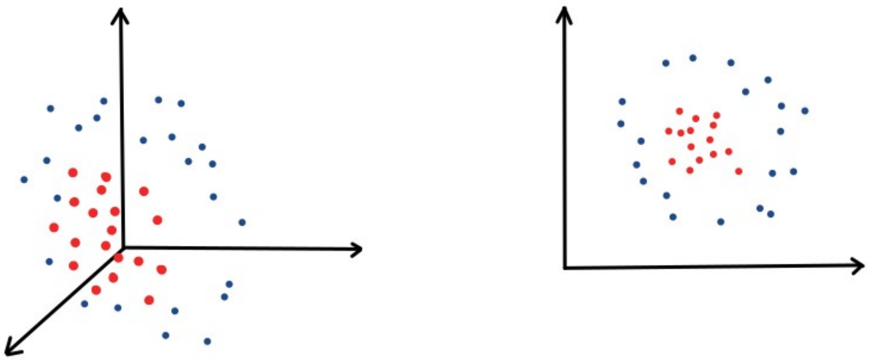
- Kategorički atributi – pravila



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(Love)	(Renaissance)	0.568935	0.549738	0.424084	0.745399	1.355916	0.111318	1.768498	0.608937
1	(Renaissance)	(Love)	0.549738	0.568935	0.424084	0.771429	1.355916	0.111318	1.885908	0.582974
2	(Nature)	(Modern)	0.328098	0.450262	0.235602	0.718085	1.594817	0.087872	1.950015	0.555094
3	(william shakespeare)	(Renaissance)	0.123909	0.549738	0.123909	1.000000	1.819048	0.055792	inf	0.513944

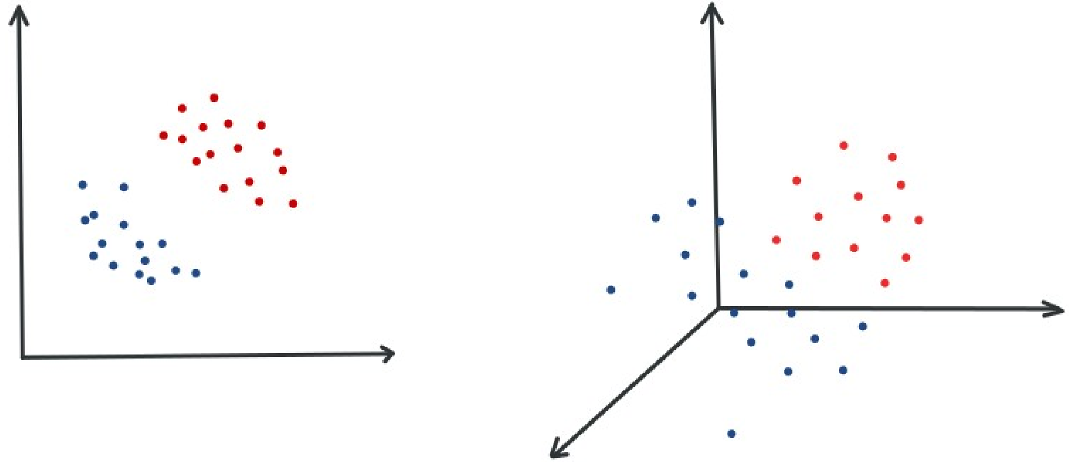
Zaključak

- Nije potrebno domensko predznanje
- Klasterovanje kao mocan alat za analizu i pretprocesiranja
- Problemi nebalansiranosti klasa
- Prokletstvo dimenzionalnosti



Zaključak

- Prokletstvo dimenzionalnosti



- Uvek treba imati na umu da su podaci, iako transormisani u numeričke, ustvari tekstualni!