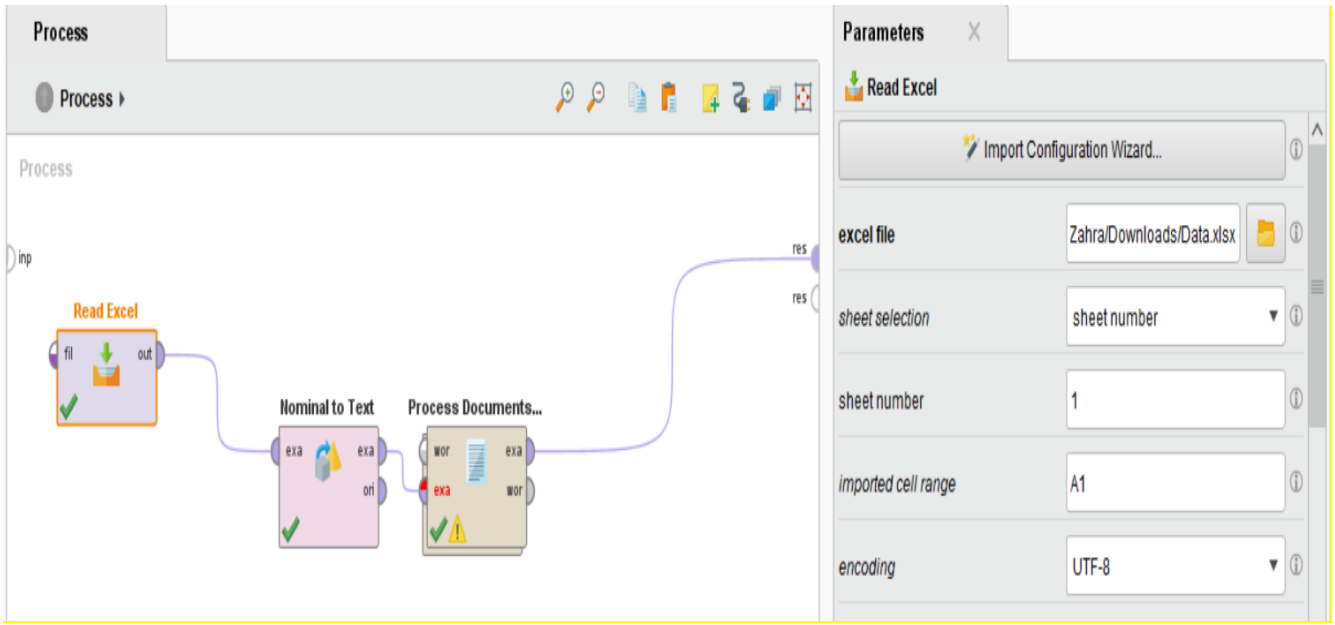
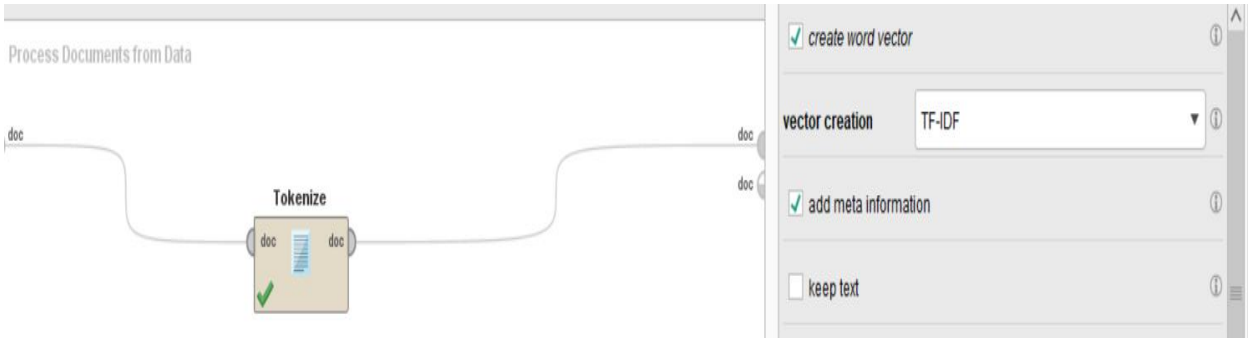
زهرا حیدری 986203146

پارسا زرگری 986203178

در اول میایم فایل رو به صورت read excel میگیریم. و برای اینک زبان فارسی رو بخواند encoding روی حالت UTF_8 میگذاریم.

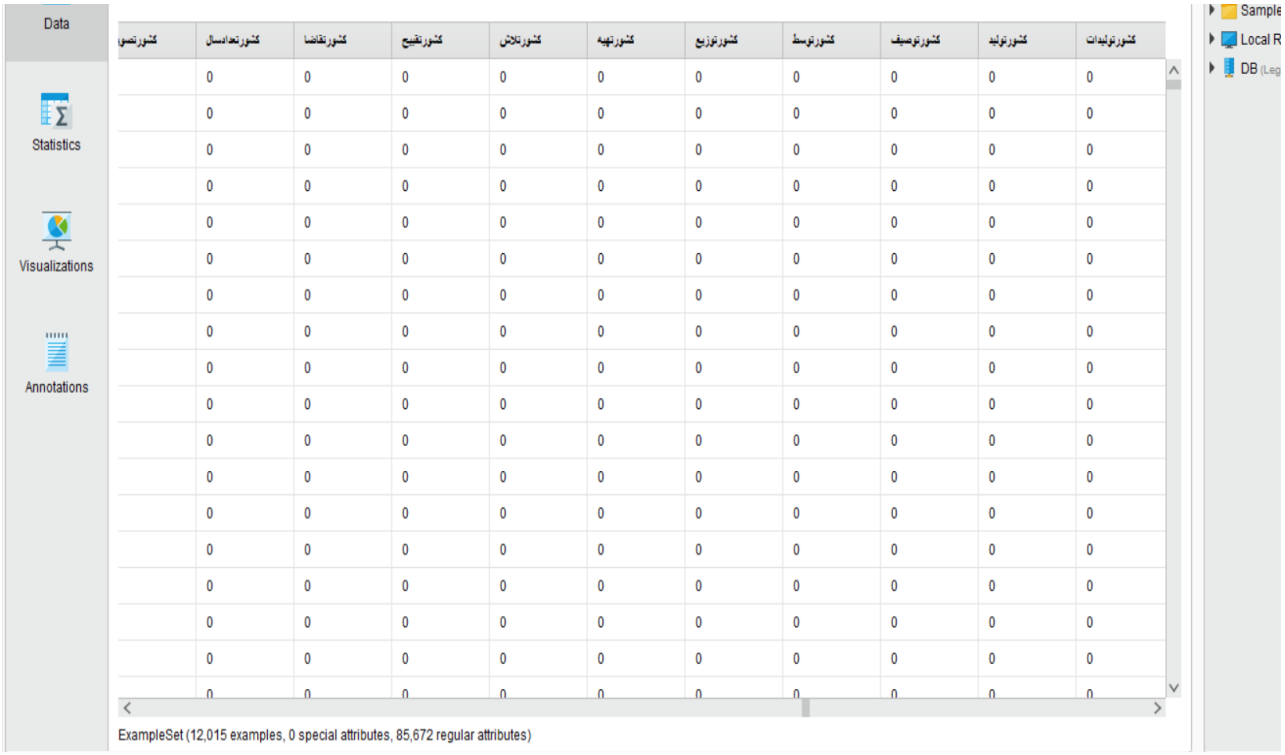


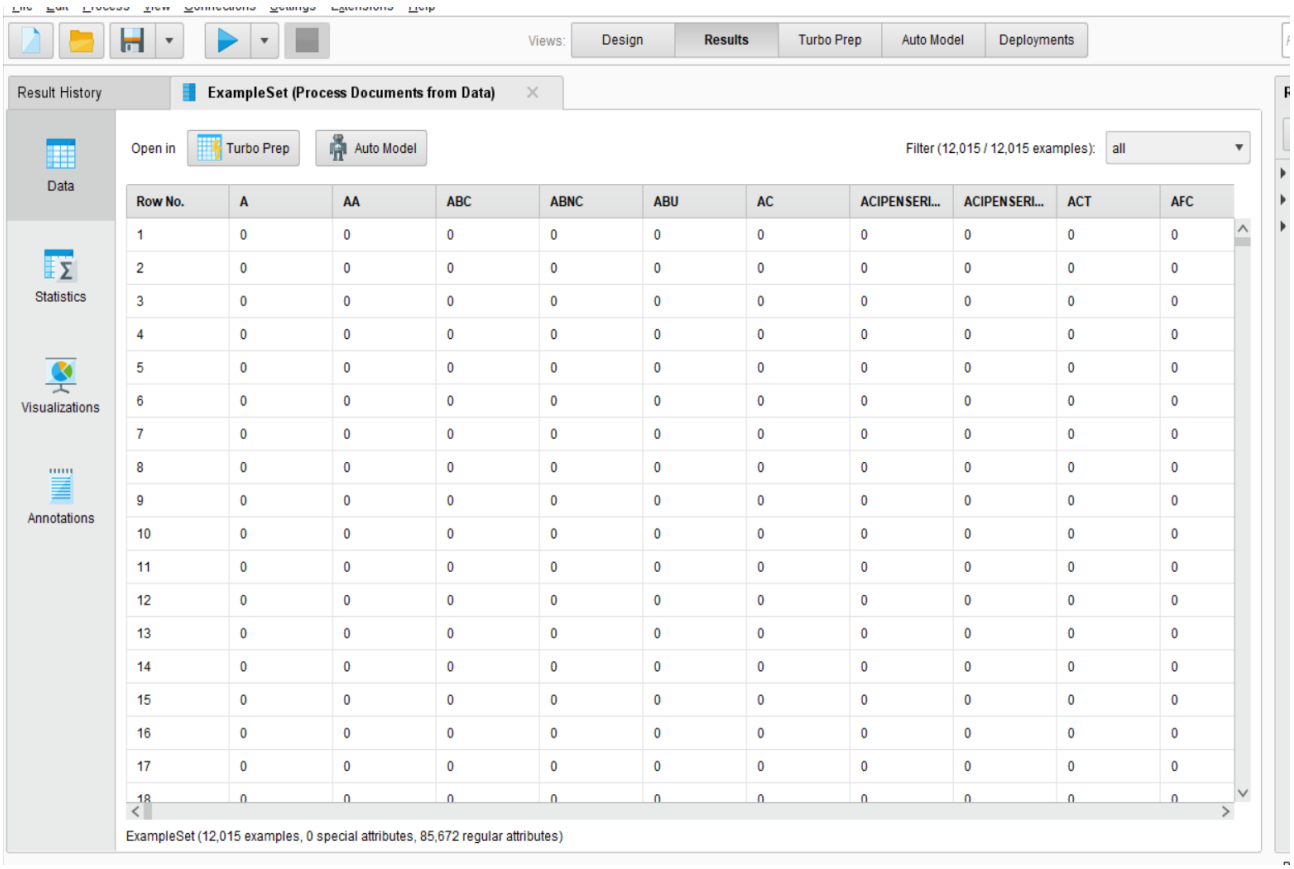
سپس برای اینک تمام رو یک متن ببنید از norminal to Text استفاده میکنیم. و بعد از آن از Process Doucmentis from data استفاده میکنیم که با استفاده از آن کلمات رو بتوانیم استخراج کنیم. در داخل این Process یک tokenize قرار میدهیم که به ما وزن های tf-idf هر کلمه رو نشان دهد.



سپس آن را اجرا میگیریم که خروجی آن شامل همه کلمات فارسی و انگلیسی داخل متن هست که هر سطر بیانگر یک متن میباشد که 12015 متن داریم. و 85672 کلمه متفاوت

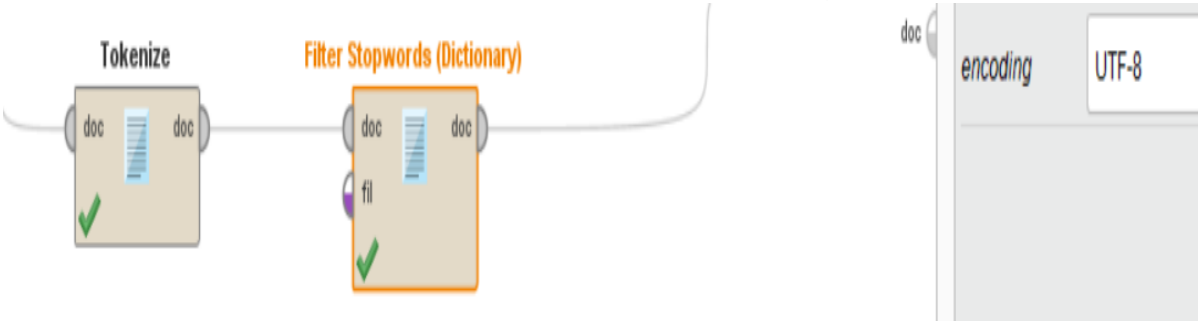
در داخل همه این متن ها موجود است که شامل کلمات تکراری و stop word ها و.. میباشد که در ادامه میخواهیم کلمات با اهمیت بیشتر را فقط وزن دهیم.



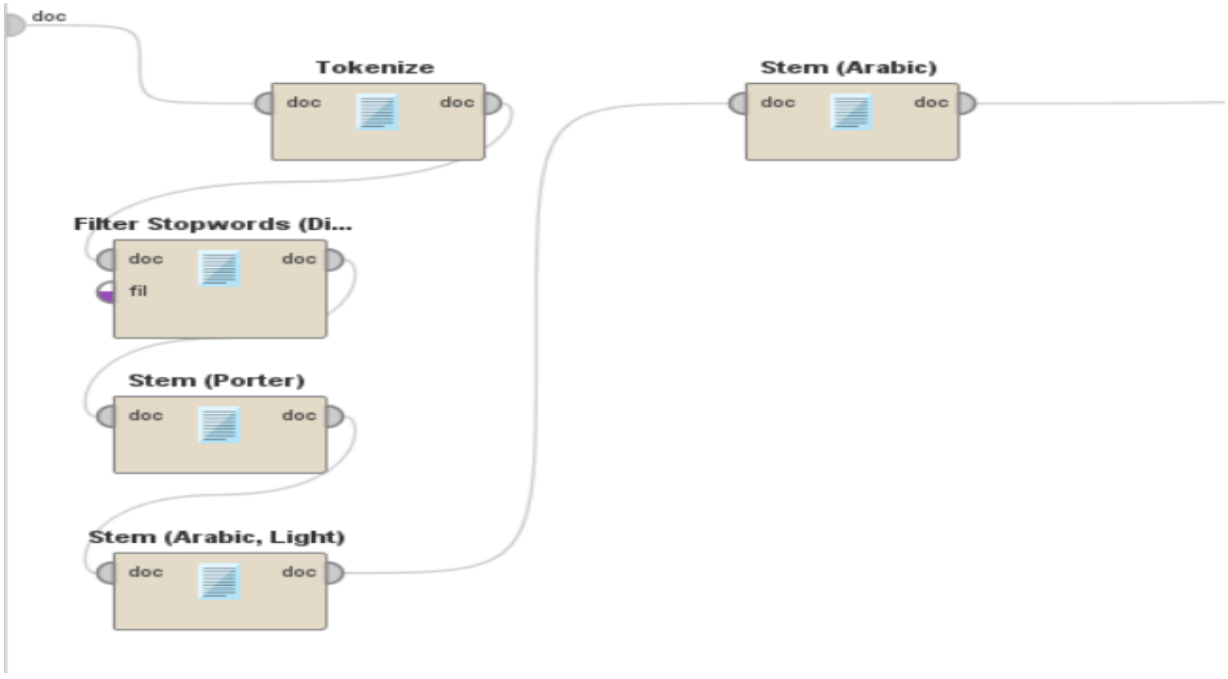


حال میایم یک تکس از stop word های فارسی از آدرس
https://github.com/kharazi/persian-stopwords/blob/master/persian

به صورت متن وارد میکنیم که آنها را حذف کند.



که حتما در اینجا encoding =UTF_8 قرار میدهیم. خروجی آن به شکل زیر است.

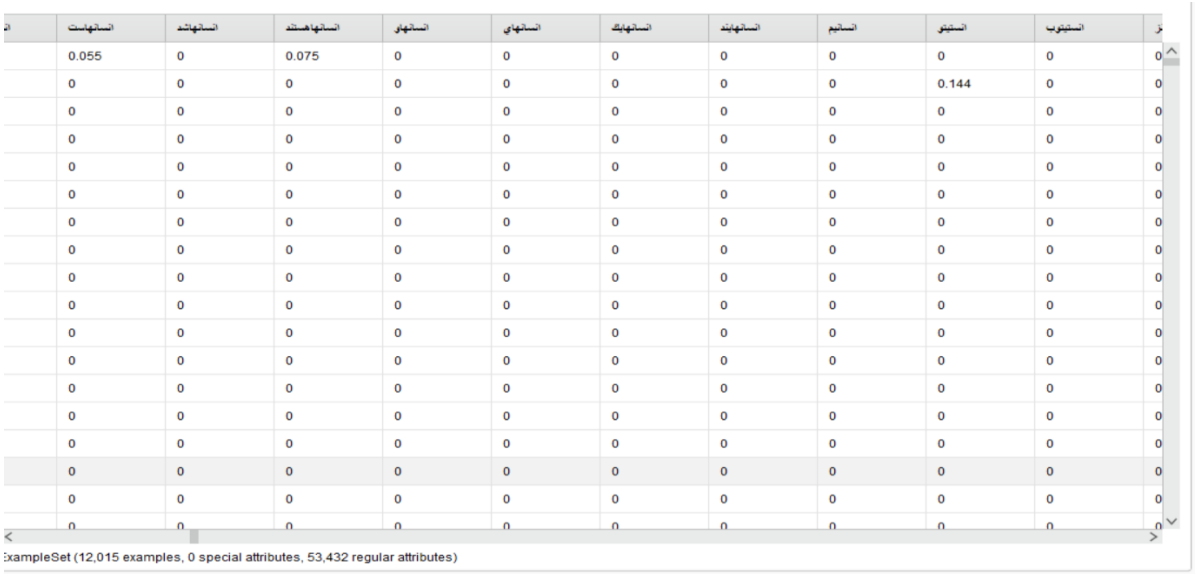


انزي	انس	انسا	انساب	انسان	انساندوستانه	انساندوستي	انسانشناسان	انسانشناسي	انسانند	انسانها
0	0	0	0	0.023	0	0	0	0	0	0.065
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0.219	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
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	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	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

ExampleSet (12,015 examples, 0 special attributes, 85,169 regular attributes)

که میبنیم تعداد کلمات کمتر و به 85169 کلمه تغییر میکند. ولی همچنان نتوانسته stop word هایی همانند (ها) رو در کلمات چسبیده به متن قبلی مانند تصویر بالا(انسانها) رو حذف کند و در ادامه stemming میکنیم و از آنجایی ک آماده در فارسی در این نرم افزار نداریم از stem کلمات عربی(که بیشتر زبان فارسی شامل همین کلماتی ک ریشه عربی دارند حاصل میشود) استفاده میکنیم و چون چندین کلمه انگلیسی هم داریم از stem انگلیسی هم استفاده میکنیم که در ادامه مشاهده میکنیم.

و خروجی آن:

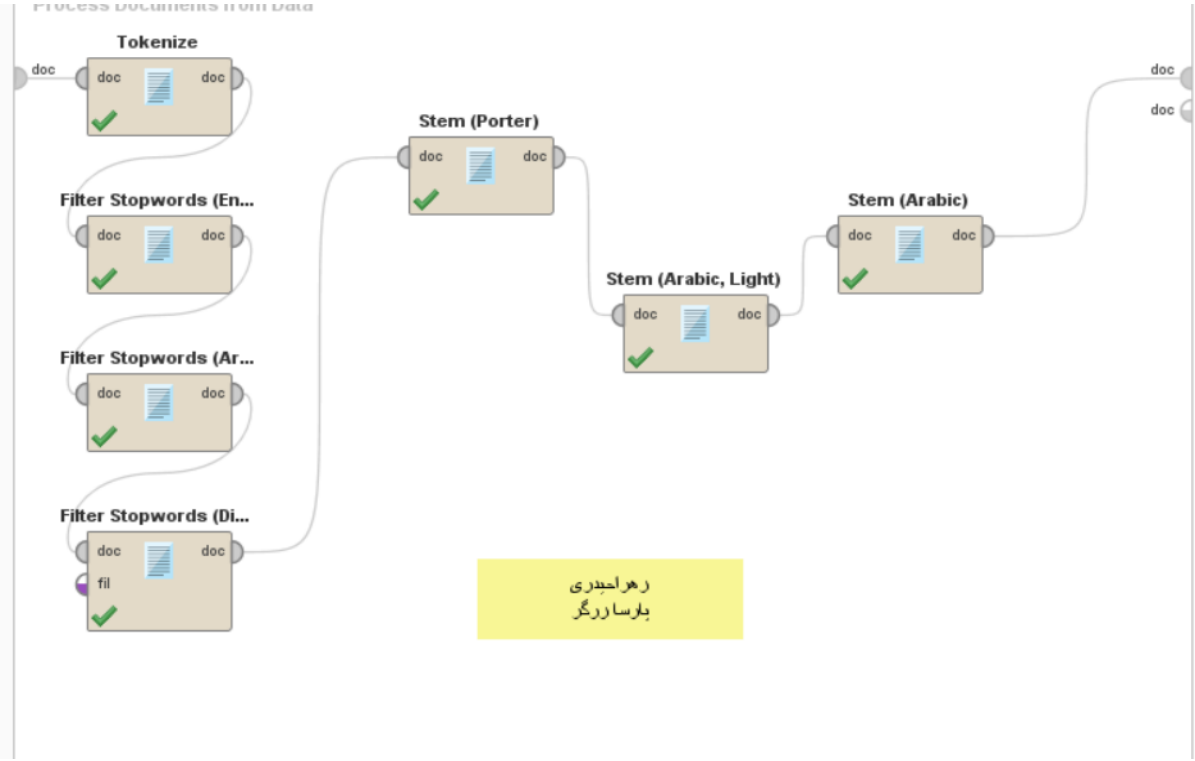


انسانهاست	انسانهاشد	انسانهاهستند	انسانهاو	انسانهاي	انسانهايك	انسانهايند	انسانيم	انستيتو	انستيتوب
0.055	0	0.075	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.144	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	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	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	0	0	0	0	0	0	0

xampleSet (12,015 examples, 0 special attributes, 53,432 regular attributes)

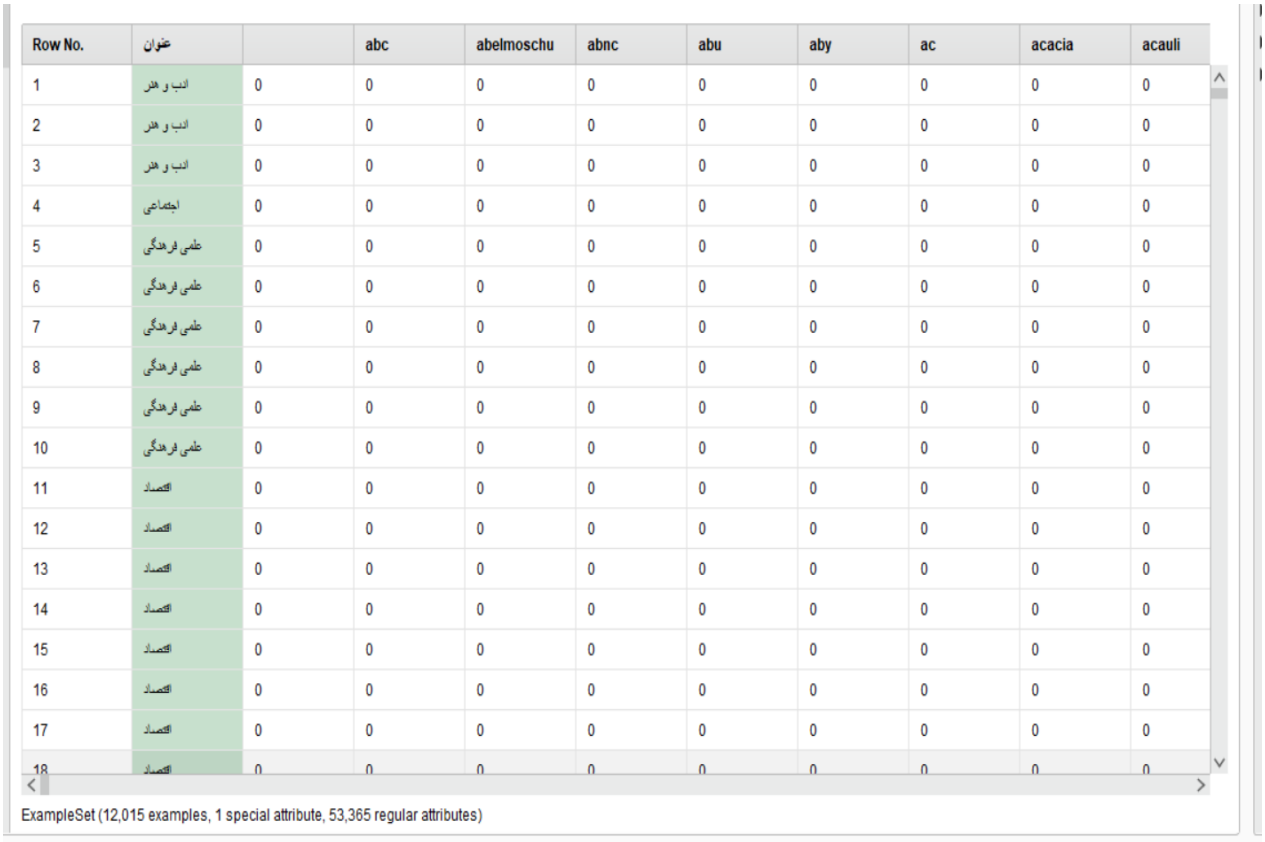
تعداد کلمات به 53432 تغییر یافته ولی همچنان مشکلات قبلی حل نشده است.

حالا یکم فایل stop word ها رو بیشتر قرار میدهیم و stem های عربی و انگلیسی رو به آن اضافه میکنیم.



زهراحیدری
پارسازرگر

خروجی آن به شکل زیراست.

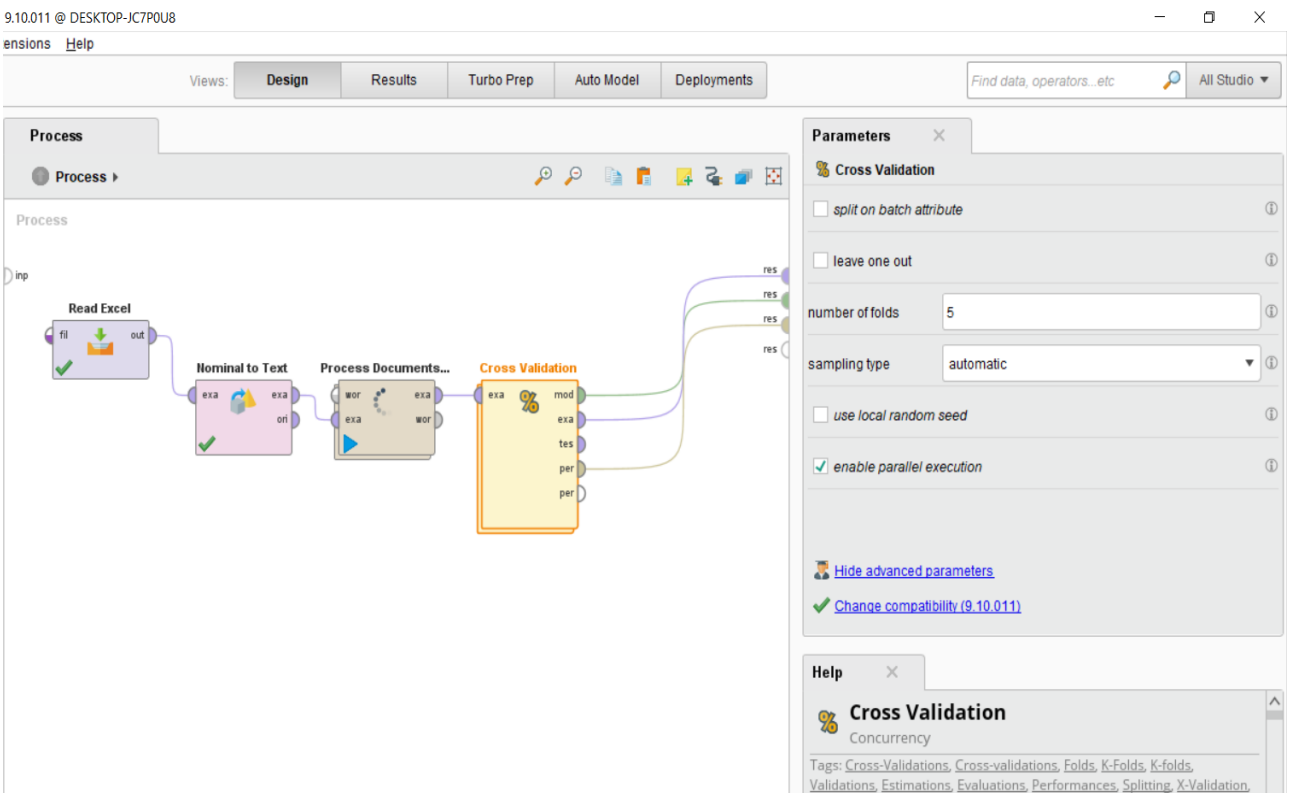


Row No.	عنوان		abc	abelmoschu	abnc	abu	aby	ac	acacia	acauli
1	ادب و هنر	0	0	0	0	0	0	0	0	0
2	ادب و هنر	0	0	0	0	0	0	0	0	0
3	ادب و هنر	0	0	0	0	0	0	0	0	0
4	اجتماعی	0	0	0	0	0	0	0	0	0
5	علمی فرهنگی	0	0	0	0	0	0	0	0	0
6	علمی فرهنگی	0	0	0	0	0	0	0	0	0
7	علمی فرهنگی	0	0	0	0	0	0	0	0	0
8	علمی فرهنگی	0	0	0	0	0	0	0	0	0
9	علمی فرهنگی	0	0	0	0	0	0	0	0	0
10	علمی فرهنگی	0	0	0	0	0	0	0	0	0
11	اقتصاد	0	0	0	0	0	0	0	0	0
12	اقتصاد	0	0	0	0	0	0	0	0	0
13	اقتصاد	0	0	0	0	0	0	0	0	0
14	اقتصاد	0	0	0	0	0	0	0	0	0
15	اقتصاد	0	0	0	0	0	0	0	0	0
16	اقتصاد	0	0	0	0	0	0	0	0	0
17	اقتصاد	0	0	0	0	0	0	0	0	0
18	اقتصاد	0	0	0	0	0	0	0	0	0

ExampleSet (12,015 examples, 1 special attribute, 53,365 regular attributes)

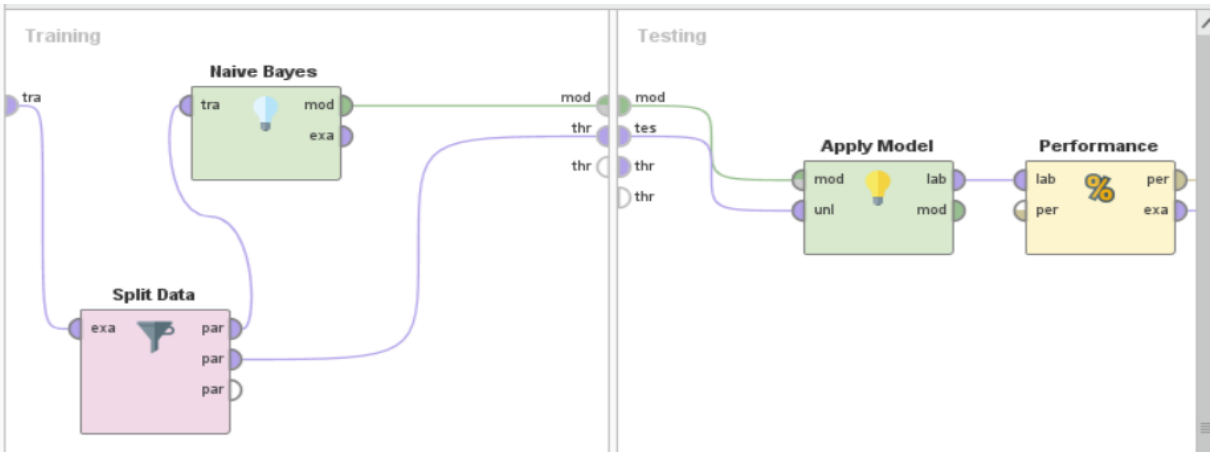
که کلمات به 53365 تا تغییر کرد که نسبت به تمام روش هایی که امتحان کردم بهینه ترین حالت ممکن بود.

حال با استفاده از Cross Validation می آییم آن رو روی 5 بار ست میکنیم و با استفاده از روش بیزین داخل آن را پیاده میکنیم.

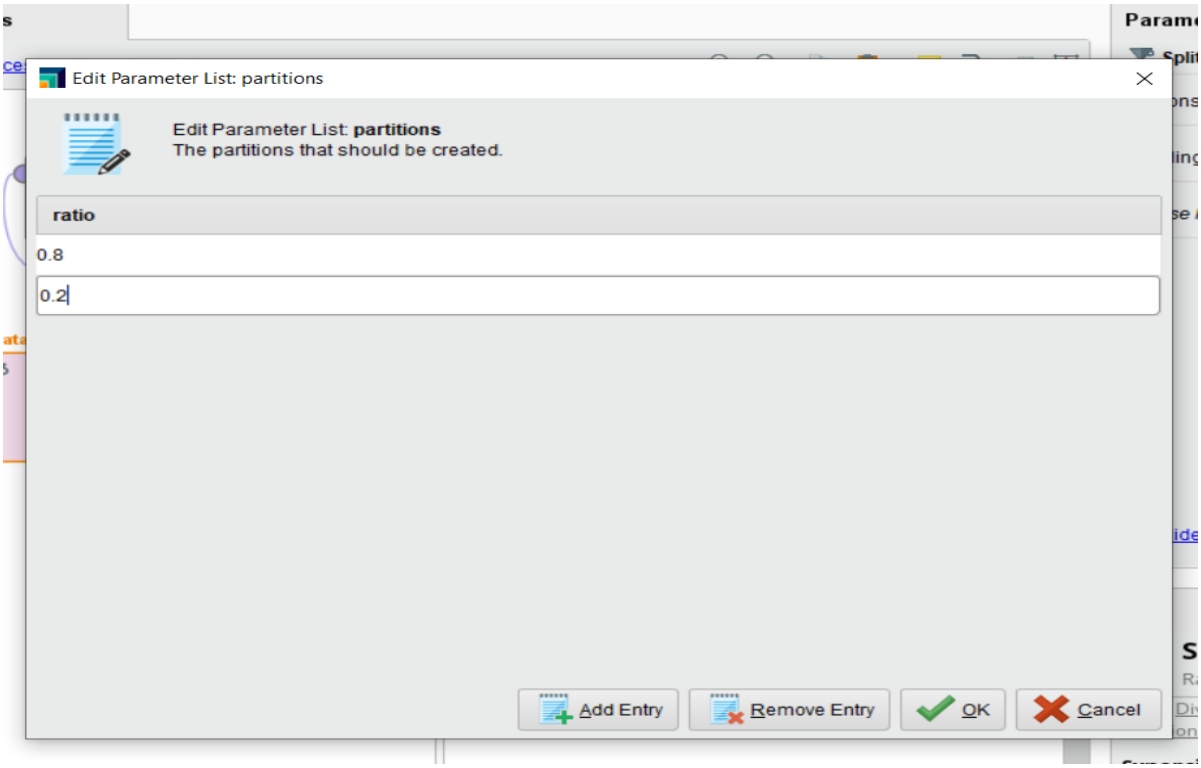


حال داخل آن دو بخش تردیند و تستینگ دارد که در هر قسمت موارد مورد نیاز آن را قرار میدهیم.

مثلا برای روش بیزین به صورت زیر است.



که در آن split data رو همانطور که گفته شده است قرار میدهیم که 0.8 به روش بیزین وصل میشد و 0.2 به اپلای مدل



و خروجی روش بیزین به صورت زیراست.

SimpleDistribution

```
Distribution model for label attribute عنوان


Class  ادب و هنر(0.043)
53365 distributions

Class  اجتماعی(0.028)
53365 distributions

Class  علمی فرهنگی(0.045)
53365 distributions

Class  اقتصاد(0.099)
53365 distributions

Class  گوناگون(0.171)
53365 distributions

Class گوناگون. حوادث(0.034)
53365 distributions

Class گوناگون. خارجی(0.127)
53365 distributions
```

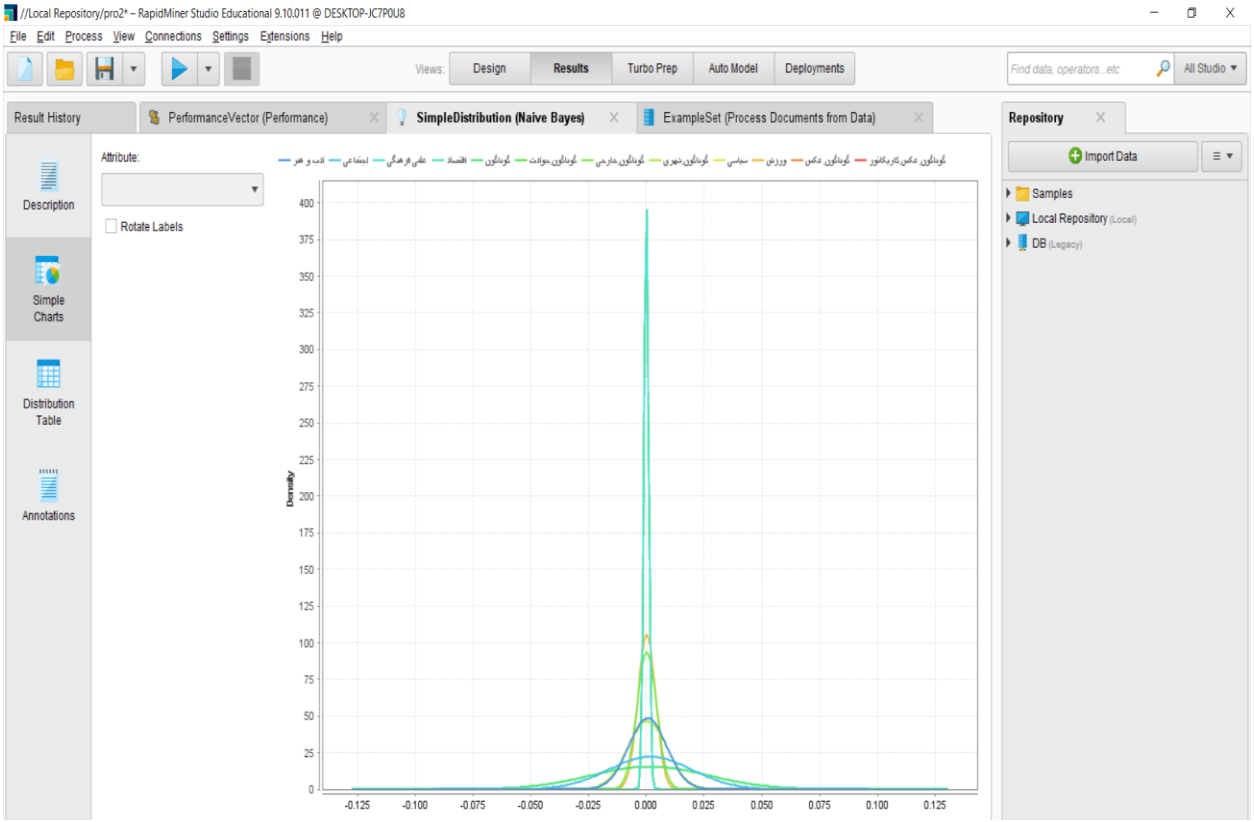
```
Class گوناگون. شهری(0.221)
53365 distributions

Class  سیاسی(0.110)
53365 distributions

Class  ورزش(0.109)
53365 distributions

Class گوناگون. عکس(0.009)
53365 distributions

Class گوناگون.عکس. کاریکاتور(0.005)
53365 distributions
```

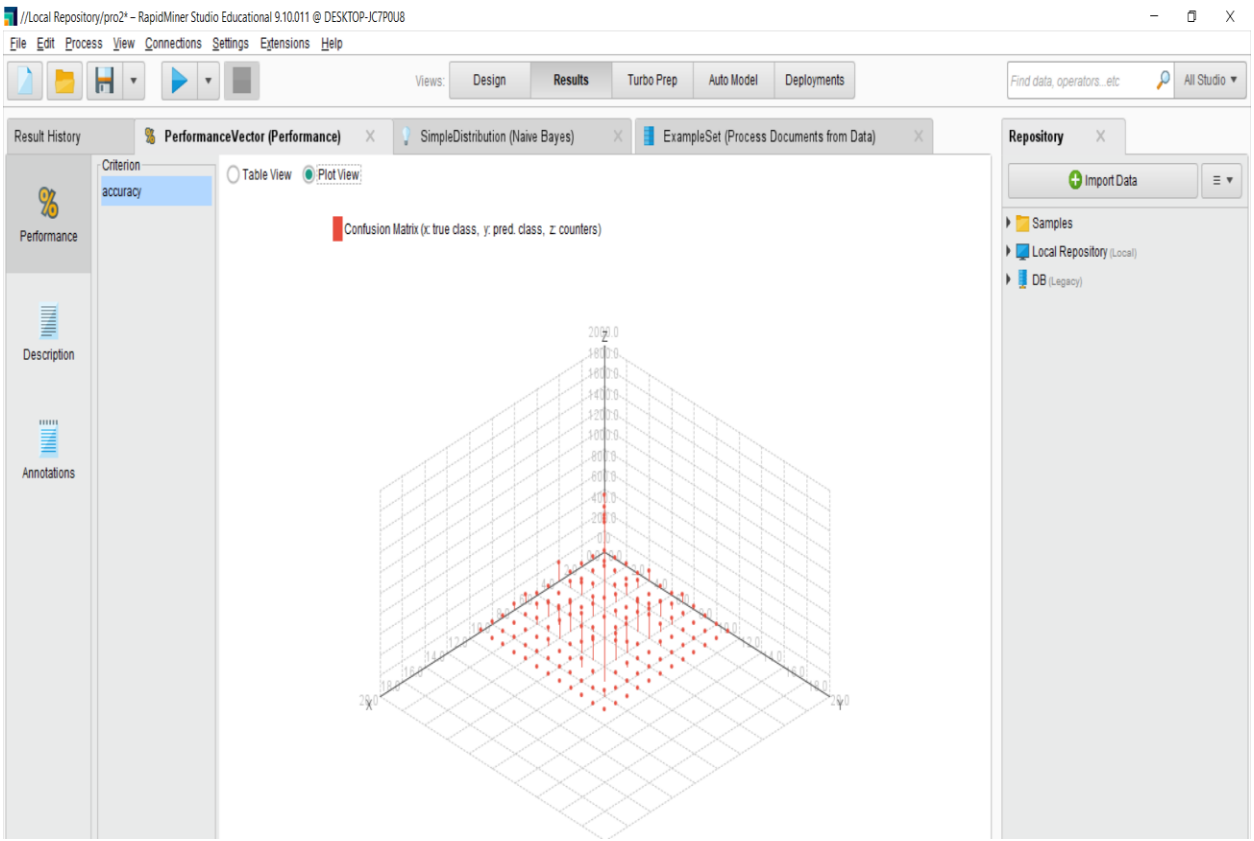


که نشان میدهد اگر یک داده جدید وارد شود به چ احتمال شامل موارد بالاست و هرکلاس چند درصد مجموعه کل رو تشکیل میدهند و جدول performance آن در ادامه نشان داده میشود.

accuracy: 58.07% +/- 1.14% (micro average: 58.07%)

	true ... ادب و	true اجتماعی	true ...علمی ف	true اقتصاد	true گوناگون	true ...گوناگو	true ...گوناگو	true ...گوناگو	true سیاسی	true ورزش	true ...گوناگو	true ...گوناگو
...اد	279	14	24	4	171	10	15	70	48	8	14	4
...اجا	8	144	25	12	59	7	7	48	17	4	1	0
...	15	17	187	38	119	10	7	126	51	7	5	12
اقتصـ	1	9	19	742	50	14	20	171	30	5	2	4
...گـ	94	66	88	75	994	117	94	270	165	50	11	7
...گـ	7	3	5	14	51	65	25	106	11	5	16	6
...گـ	12	6	7	38	46	33	1195	57	98	8	0	1
...گـ	68	53	138	192	301	135	48	1527	215	82	10	5
سیاـ	24	22	47	63	226	13	102	248	671	28	4	4
ورز	5	1	6	7	33	1	8	31	15	1116	4	5
...گـ	0	0	0	0	0	0	0	0	0	0	46	0
...گـ	0	0	0	0	0	0	0	0	0	0	0	11
...c	54.39%	42.99%	34.25%	62.62%	48.49%	16.05%	78.57%	57.54%	50.79%	85.00%	40.71%	18.64%

که نشان میدهد مدل در هر یک موارد چقدر را به درستی نشان داده است. و نمودار شکلی آن به این صورت است.



نکته:برای استفاده از این روش حتما باید برای داده ها لیبل انتخاب کنیم ک لیبل رو همان عنوان انتخاب کردیم.