Amirkabir University of Technology
(Tehran Polytechnic)

Department of
Computer Engineering

Natural Language Processing

Homework 1

Najmeh Mohammadbagheri

99131009

گزارش تمرین

بخش اول

همانطور که می‌دانیم برای پردازش بر روی داده‌ها ابتدا لازم است یک پیش‌پردازش داشته باشیم. در حوزه‌ی پردازش متن اصولا لازم است که برای پیش‌پردازش، گام‌های زیر انجام شود:

نرمالسازی، به معنای یکسان کردن حروف با شکل‌های متفاوت؛ یعنی شکل‌های عربی، شکل‌های مختلف فارسی یک حرف، با یک نماد استاندارد متناظر جایگیزین شوند. همچنین اعداد فارسی و انگلیسی به یک شکل تبدیل شوند یا اگر برایمان ارزشی ندارند از محتوا حذف شوند. همچنین اگر داده‌ها از وب گرفته شده‌اند لازم است علائمی چون tab ها یا متاتگ‌ها حذف شوند. همچنین لازم است نیم‌فاصله‌ها به یک فرمت تبدیل شوند.

استخراج توکن، بدین منظور که واحدهای مستقل جمله از هم جدا شوند. در اکثر موارد استخراج توکن در سطح کلمه انجام می‌شود.

حذف کلمات پرتکرار[1]، در این مرحله کلماتی که می‌دانیم ارزشی برای بررسی کردن ندارند را حذف می‌کنیم تا مدل بهتر کار کند. کلماتی مانند: از، در، به، است، که و غیره.

ریشه‌یابی کلمات، بدین منظور که اشکال مختلف یک کلمه به ریشه‌ی خود نگاشت شوند انجام می‌شود. این کار برای مواردی چون بازیابی اطلاعات بسیار اساسی و مهم است. به طور مثال در این گام کلمه‌ی درخت‌ها و درخت دو کلمه‌ی مجزا در دیکشنری ذخیره نمی‌شوند، یا افعال رفتم، رفتند، خواهند رفت همگی به ریشه‌ی رفت تبدیل شده و تنها رفت در دیکشنری ذخیره می‌شود.

یک ابزار مناسب برای انجام این پیش‌پردازش‌ها در زبان پایتون، کتابخانه‌ی پارسی‌ور است.

در این تمرین گام نرمالسازی و استخراج توکن انجام شده‌است. اما به دلیل اینکه جنس داده‌ها شعر بوده است و هر شاعر یک سبک برای استفاده از کلمات دارد، حذف کلمات پرتکرار و ریشه‌یابی انجام نشده است. زیرا بنده بر این باور بوده‌ام که میزان استفاده از کلماتی چون ز، در، کجا و از این قبیل موارد کاملا به سبک آن شاعر وابسته است و در نهایت می‌تواند به دسته‌بندی بهتر اشعار کمک کند. همچنین ریشه‌یابی نیز انجام نشده است زیرا برخی از شعرا اغلب به ضمایر خاصی شعر می‌سرایند و یکسان دانستن تمام افعال بر دسته‌بندی تاثیر منفی می‌گذارد. (البته بهتر بود که در هر دو حالت این تمرین را انجام می‌دادم و با دلیل و مدرک نتیجه‌گیری می‌کردم.)

در این تمرین توکن‌های هر شاعر در ساختار دیکشنری ذخیره شده‌است و در ادامه از همین دیکشنری‌ها استفاده شده است. دیکشنری‌ها در دوحالت بایگرم و یونیگرم ساخته شده‌اند.

نکته: در هنگام ساخت دیکشنری بایگرم، بایگرم‌های هر مصرع به صورت جداگانه ذخیره شده‌است. زیرا بنظرم قافیه‌ی بیت ارتباطی با کلمه‌ی شروع مصرع بعد ندارد.

[1] Stop words

بخش دوم

هموارسازی absolute discounting برای مدل یونی گرم:

$$P(w_i) = \frac{\max(\#(w_i) - \delta, 0)}{N} + \alpha P_{bg}, \qquad P_{bg} = \frac{1}{V}, \alpha = \frac{\delta}{N} B$$

که N برابر با تعداد کل توکن‌های داده‌های آموزشی است و V برابر با تعداد کلمات متمایز دیکشنری است. همچنین B نیز برابر با تعداد کلمات متمایز دیکشنری است.

برای مدل بایگرم این هموارسازی به شکل فرمول زیر استفاده شده‌است:

$$P(w_i | w_{i-1}) = \frac{\max(\#(w_i, w_{i-1}) - \delta, 0)}{\#(w_{i-1})} + \alpha P_{bg}, \qquad P_{bg} = P(w_i), \alpha = \frac{\delta}{\#(w_{i-1})} B$$

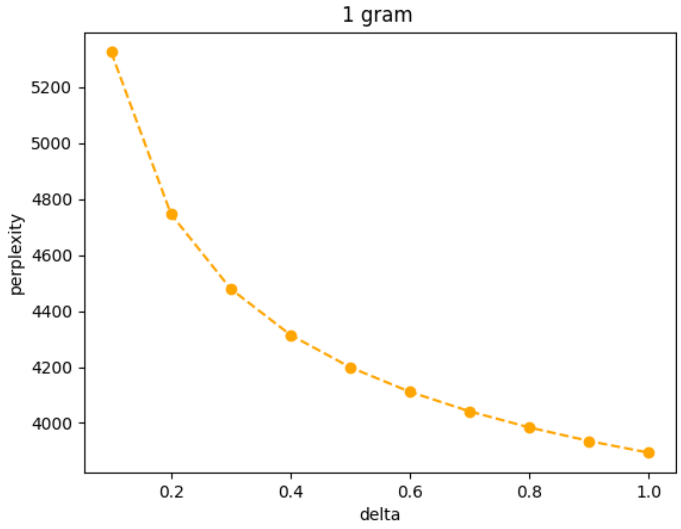
در این روش چندین چالش وجود داشت. چالش اول: صفر بودن $\#(w_{i-1})$ است. در این حالت بجای استفاده از احتمال بایگرم به طور کامل از احتمال یونیگرم کلمه‌ی iام استفاده شده‌است. چالش دوم: مقدار B بود که با بررسی‌های انجام‌شده در حالتی که B برابر با تعداد بایگرم‌های متمایز دیکشنری در نظر گرفته می‌شد، عملکرد بهتری بدست می‌آمد.

تابع perplexity نیز شبیه آنچه که در کلاس تدریس شده بود پیاده‌سازی شد. فرمول کلی در زیر آمده است:

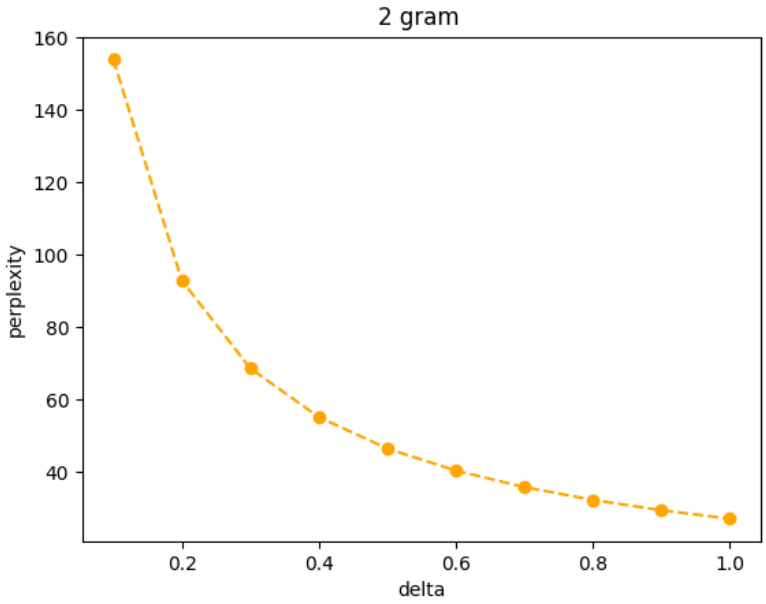
$$Perplexity(S) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1, w_2, \ldots, w_{i-1})}}$$

برای بدست آوردن perplexity داده‌های تست یا ارزیابی، این میزان برای هر بیت به صورت جداگانه محاسبه شد و در نهایت میانگین‌گیری صورت گرفت.

قسمت اول: در نمودار شکل ۱ perplexity برای دلتاهای ۰.۱ تا ۱ روی داده‌های اعتبارسنجی در حالت یونیگرم و در نمودار شکل ۲ در حالت بایگرم قابل مشاهده است.



شکل ۱



شکل ۲

همانطور که مشاهده می‌شود در هردو حالت بهترین مقدار دلتا یک است.

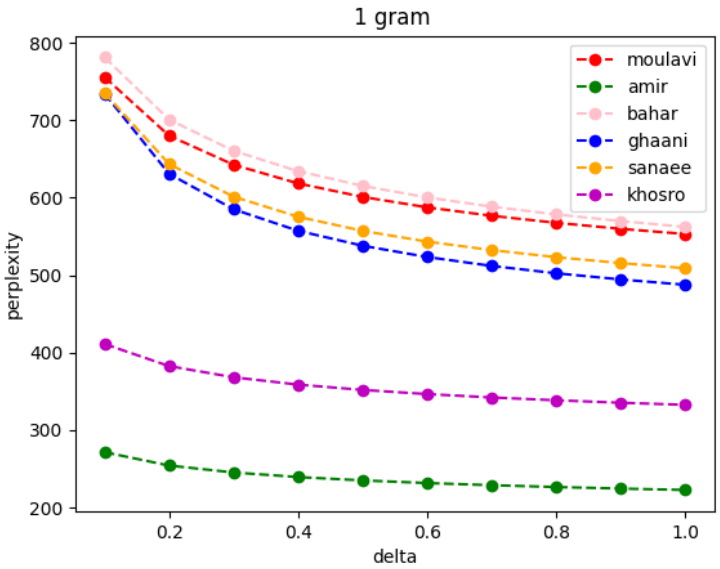
تحلیل: میزان پرپلکسیتی در حالت بایگرم به شدت کاهش یافته است. این امر کاملا قابل انتظار بود زیرا در حالت بایگرم احتمال‌ها با در نظر گرفتن کلمه‌ی قبلی در نظر گرفته می‌شوند و این باعث می‌شود که سرگشتگی در زبان کاهش یابد.

در حالت بایگرم تعداد کل بایگرم‌های متمایز دیکشنری ۵۵۷۱۰۵ و در حالت یونیگرم تعداد کل یونیگرم‌های متمایز دیکشنری ۶۷۹۹۲ است.

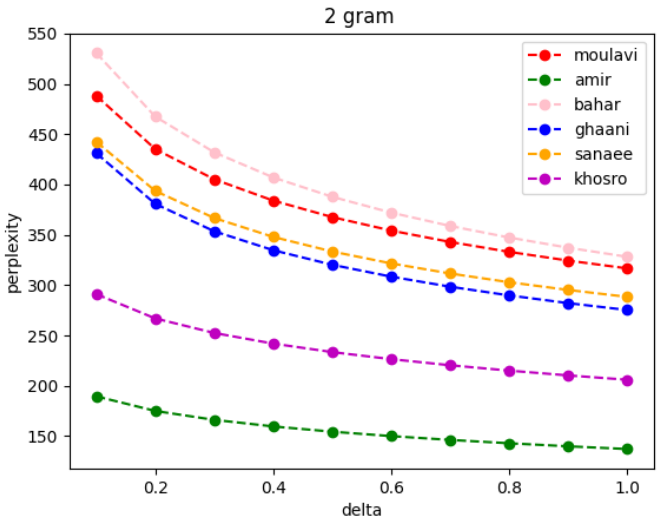
میزان پرپلکسیتی داده‌های آزمون در حالت یونیگرم با دلتای ۱ : ۳۹۲۵.۵۶

میزان پرپلکسیتی داده‌های آزمون در حالت بایگرم با دلتای ۱ : ۴۲.۴۲

قسمت دوم: در این قسمت تمام روابط به همان صورت که در قبل گفته‌شد استفاده شده‌است. با این تفاوت که هر شاعر به عنوان یک مجموعه داده‌ی مجزا درنظر گرفته شده و به طور مثال N,V برای هر مجموعه یعنی هر شاعر به صورت مستقل محاسبه و استفاده شده است. در نمودار شکل ۳ perplexity برای دلتاهای ۰.۱ تا ۱ روی داده‌های اعتبارسنجی در حالت یونیگرم و در نمودار شکل ۴ در حالت بایگرم قابل مشاهده است.



شکل ۳



شکل ٤

در این حالت‌ها نیز بهترین دلتا یک است. همانطور که مشاهده می‌شود در این قسمت میزان پرپلکسیتی کاهش یافته است زیرا تعداد کلمات متمایز هر شاعر کمتر از تعداد کل کلمات متمایز در قسمت قبل بوده است. با توجه به این نمودار متوجه می‌شویم اندازه‌ی مجموعه لغاتی که هر شاعر استفاده کرده است در چه رنجی است. به طور مثال بهار از تنوع بیشتری نسبت به امیر در استفاده از کلمات برخوردار است.

میزان پرپلکسیتی داده‌های آزمون در حالت یونیگرم با دلتای ۱ برای هر شاعر :

[۵۰۶.۸۶۵۱۶۰۵۴ مولوی] [۲۲۵.۱۷۳۳۲۲۲۲ امیر] [۵۸۷.۰۰۴۹۳۴۸۵ بهار] [۵۱۰.۰۹۹۸۸۲۳۸ قانی] [۴۷۵.۰۶۱۵۷۹۱۸ثنایی]
[۳۵۷.۰۱۳۳۶۲۱۸ خسرو]

میزان پرپلکسیتی داده‌های آزمون در حالت بایگرم با دلتای ۱ برای هر شاعر:

[۲۸۹.۲۳۹۶۳۲ مولوی] [۱۳۹.۶۴۳۵۰۳۵۷ امیر] [۳۳۳.۸۸۴۹۲۵۶۷ بهار] [۲۸۲.۹۴۹۰۳۷۴۶ قانی] [۲۷۰.۶۹۳۵۲۶۹۱ ثنایی]
[۲۱۵.۵۹۴۲۹۱۲۳ خسرو]

بخش سوم

X_square,Information gain در این قسمت با همان رابطه‌ای که در اسلاید درس وجود دارد محاسبه شده است:

$$IG(w) = -\sum_{i=1}^{K} P(c_i)\log P(c_i) + P(w)\sum_{i=1}^{K} P(c_i|w)\log P(c_i|w) + P(\overline{w})\sum_{i=1}^{K} P(c_i|\overline{w})\log P(c_i|\overline{w})$$

$$\chi^2(w, c_i) = \frac{N \cdot (N_{iw}N_{\bar{i}\overline{w}} - N_{i\overline{w}}N_{\bar{i}w})^2}{(N_{iw} + N_{i\overline{w}}) \cdot (N_{\bar{i}w} + N_{\bar{i}\overline{w}}) \cdot (N_{iw} + N_{\bar{i}w}) \cdot (N_{i\overline{w}} + N_{\bar{i}\overline{w}})}$$

$$\chi^2(w) = \sum_{i=1}^{K} P(c_i) \cdot \chi^2(w, c_i)$$

خروجی IG به صورت زیر است :

	Poet	information gain	word
1	moulavi	0.03768	به
2	amir	0.0286101	تو
3	khosro	0.024643	خسرو
4	amir	0.0226031	است
5	amir	0.0192526	و
6	amir	0.0168301	او
7	khosro	0.0162689	که
8	khosro	0.0138132	من
9	amir	0.0124714	دولت
10	khosro	0.0113676	دل
11	amir	0.00891919	خود

12	amir	0.00861697	ملک	42	amir	0.00399308	اندر	72	sanaee	0.00296243	مرد
13	moulavi	0.00842973	آن	43	sanaee	0.0039702	عقل	73	moulavi	0.00292048	همه
14	amir	0.0082295	همی	44	khosro	0.00393973	ست	74	amir	0.00290437	نصرت
15	sanaee	0.00779017	سنایی	45	amir	0.00389561	عدل	75	amir	0.00284507	اقبال
16	khosro	0.00772011	چه	46	amir	0.0038145	همیشه	76	moulavi	0.00280464	وا
17	khosro	0.00761533	جان	47	khosro	0.00381032	خوش	77	bahar	0.00279147	چون
18	moulavi	0.0075783	گفت	48	khosro	0.00379502	دیده	78	khosro	0.00278262	می‌آید
19	amir	0.00695452	توست	49	ghaani	0.00378595	قاآنی	79	moulavi	0.00277878	هم
20	bahar	0.00678458	ترا	50	moulavi	0.00373012	هین	80	bahar	0.00276835	عالم
21	moulavi	0.00672965	چونک	51	khosro	0.00368911	کوی	81	amir	0.00273704	می‌کند
22	sanaee	0.00670002	دین	52	sanaee	0.00359545	شد	82	amir	0.00273313	جهان
23	amir	0.00663644	این	53	amir	0.00359318	کن	83	amir	0.00272661	خدمت
24	moulavi	0.00635822	کو	54	moulavi	0.00347231	تیغ	84	khosro	0.00272348	گریه
25	khosro	0.0055887	خون	55	bahar	0.00340262	در	85	amir	0.00272005	بر
26	amir	0.00554517	باد	56	amir	0.00337524	شاه	86	moulavi	0.00270767	سپهر
27	moulavi	0.00550321	زانک	57	sanaee	0.00335195	عشق	87	bahar	0.00269635	زن
28	moulavi	0.00549389	ز	58	khosro	0.0033226	غمزه	88	bahar	0.00269103	تا
29	amir	0.00522785	ما	59	khosro	0.00331713	دلم	89	amir	0.00267058	بزم
30	bahar	0.00522063	چو	60	amir	0.00330188	رای	90	amir	0.00266573	ظفر
31	ghaani	0.00507992	یی	61	amir	0.00328968	دارد	91	khosro	0.00265106	کنم
32	bahar	0.00501785	ایران	62	ghaani	0.00328581	باشد	92	moulavi	0.0026487	جمله
33	bahar	0.00493868	را	63	amir	0.00328004	رزم	93	khosro	0.00264721	مرا
34	amir	0.00473273	هست	64	amir	0.00323674	لیک	94	amir	0.00262523	زمین
35	moulavi	0.00443465	حق	65	khosro	0.00321989	ازان	95	amir	0.00258905	کاو
36	amir	0.00434883	تورا	66	amir	0.00320651	فتح	96	amir	0.00255934	شدست
37	khosro	0.00431487	غم	67	bahar	0.003199	هر	97	ghaani	0.0025286	مهر
38	moulavi	0.00417749	زلف	68	moulavi	0.00316843	گیتی	98	sanaee	0.00249889	همچو
39	khosro	0.00409474	جود	69	khosro	0.00311298	مر	99	khosro	0.00249061	طبع
40	amir	0.00405506	بخت	70	moulavi	0.0030964	رو	100	moulavi	0.00248597	خدا
41	amir	0.00401474	مدح	71	moulavi	0.00306653	کی	101	ghaani	0.00248048	چرخ

102	khosro	0.00246327	می‌رود	132	moulavi	0.0020951	آنچ	162	khosro	0.00188282	جانم
103	ghaani	0.00241314	بحر	133	bahar	0.00208884	ری	163	sanaee	0.00187795	دان
104	amir	0.00239577	شرف	134	amir	0.00208542	شهریار	164	moulavi	0.00185794	شعر
105	khosro	0.00239291	سرو	135	bahar	0.00208376	کشور	165	amir	0.0018481	پیروزی
106	khosro	0.00238223	فر	136	bahar	0.00207809	بهار	166	amir	0.00183819	حشمت
107	amir	0.00234731	ار	137	khosro	0.00207771	چشم	167	moulavi	0.00182637	کاین
108	moulavi	0.00234547	پس	138	moulavi	0.00205581	آنکه	168	amir	0.0018123	نی
109	moulavi	0.00232515	هرکه	139	moulavi	0.00205399	یا	169	amir	0.00180498	ره
110	moulavi	0.00231964	ای	140	ghaani	0.00205278	چهر	170	amir	0.001792	ماه
111	ghaani	0.00231474	از	141	ghaani	0.00205075	خصم	171	ghaani	0.00178693	بود
112	amir	0.00230918	روزگار	142	amir	0.00204348	کلک	172	amir	0.00178093	مکن
113	khosro	0.00229637	صبا	143	amir	0.00204302	ملت	173	ghaani	0.00177817	کند
114	bahar	0.00228765	وطن	144	khosro	0.00203008	جگر	174	ghaani	0.00175899	بسکه
115	amir	0.00227598	ا	145	bahar	0.00201915	خراسان	175	amir	0.00174124	تویی
116	amir	0.00226817	همت	146	moulavi	0.00200464	خر	176	amir	0.00173187	درد
117	khosro	0.00225689	سینه	147	moulavi	0.00200134	اگر	177	ghaani	0.00173092	رخش
118	amir	0.00224146	سعادت	148	sanaee	0.00198949	علم	178	ghaani	0.00171317	فارس
119	khosro	0.00222985	چند	149	amir	0.00195601	پی	179	khosro	0.00170216	خواهم
120	khosro	0.00222644	خسروا	150	khosro	0.00194599	خوبان	180	khosro	0.00169001	سو
121	khosro	0.00222167	گل	151	amir	0.0019458	گردون	181	sanaee	0.00168467	راه
122	khosro	0.00220588	باز	152	moulavi	0.00193884	روی	182	amir	0.00168393	تاکه
123	bahar	0.00219695	مردم	153	amir	0.00193216	فخر	183	moulavi	0.00167668	صد
124	sanaee	0.00219267	شرع	154	amir	0.00192906	شاهی	184	bahar	0.00166825	زبن
125	amir	0.00218401	سلطان	155	moulavi	0.00192731	داشتن	185	ghaani	0.00166794	هرچه
126	khosro	0.00218359	خواب	156	sanaee	0.00192499	شود	186	khosro	0.00166143	باری
127	khosro	0.00217026	لب	157	amir	0.0019245	ایزد	187	amir	0.00165913	لولو
128	khosro	0.00214818	وه	158	bahar	0.00192156	تست	188	khosro	0.00165697	دیوانه
129	bahar	0.00214771	گر	159	ghaani	0.00191762	پیکر	189	ghaani	0.00163916	سوی
130	moulavi	0.0021092	فلک	160	khosro	0.00190165	مست	190	ghaani	0.00162774	نه
131	moulavi	0.00210888	خدای	161	ghaani	0.00188781	ابر	191	sanaee	0.00162325	سرای

192 moulavi 0.00162035 رخ	195 amir 0.00160925 طلعت	198 amir 0.00158895 گویی
193 amir 0.0016102 برای	196 ghaani 0.00159532 پیش	199 khosro 0.00158108 لبت
194 khosro 0.00160981 عاشقان	197 moulavi 0.00159102 لا	200 amir 0.00158031 پا

خروجی X^2 به صورت زیر است:

0 moulavi 985.493 به	26 moulavi 158.812 کو	52 amir 100.966 فتح
1 khosro 929.681 خسرو	27 amir 151.465 تورا	53 khosro 98.2103 گریه
2 amir 655.626 تو	28 moulavi 151.328 حق	54 amir 96.4559 جود
3 amir 584.387 است	29 moulavi 150.923 ز	55 khosro 96.2501 خوش
4 khosro 475.385 که	30 moulavi 150.44 هین	56 sanaee 96.1822 شد
5 amir 447.934 و	31 amir 143.669 باد	57 amir 95.0645 اندر
6 khosro 431.164 من	32 khosro 142.281 غم	58 amir 94.0404 نصرت
7 amir 417.694 دولت	33 khosro 137.492 ست	59 moulavi 92.1162 رو
8 amir 409.477 او	34 ghaani 134.92 قاآنی	60 amir 91.7864 اقبال
9 khosro 357.087 دل	35 moulavi 131.466 چو	61 khosro 90.4318 می‌آید
10 sanaee 304.189 سنایی	36 amir 128.894 هست	62 amir 87.7406 رای
11 moulavi 275.545 آن	37 moulavi 127.721 را	63 amir 87.5084 رزم
12 moulavi 260.161 چونک	38 sanaee 127.377 عقل	64 sanaee 87.4269 عشق
13 moulavi 245.546 گفت	39 khosro 124.655 غمزه	65 moulavi 87.0194 کی
14 amir 245.049 توست	40 moulavi 123.717 ما	66 moulavi 86.6414 تیغ
15 amir 225.586 همی	41 sanaee 122.825 ترا	67 amir 86.4093 شاه
16 amir 225.306 ملک	42 amir 122.359 همیشه	68 khosro 86.0949 صبا
17 moulavi 223.621 زانک	43 khosro 121.755 دیده	69 amir 85.5328 خدمت
18 khosro 220.839 چه	44 khosro 116.3 ازان	70 khosro 85.2479 کنم
19 sanaee 202.304 دین	45 khosro 115.956 کوی	71 moulavi 84.1849 آنچ
20 khosro 196.052 جان	46 moulavi 114.672 وا	72 amir 84.0935 دارد
21 khosro 190.889 خون	47 khosro 112.17 دلم	73 sanaee 83.4651 مر
22 ghaani 182.769 یی	48 khosro 110.827 زلف	74 amir 83.4563 ظفر
23 moulavi 171.714 این	49 amir 107.089 بخت	75 sanaee 82.9792 شرع
24 amir 171.324 خود	50 amir 103.394 مدح	76 khosro 82.5625 مرا
25 bahar 165.269 ایران	51 amir 100.995 عدل	77 moulavi 81.6142 هم

78	khosro 81.4849	خسروا	108	khosro 67.7733	لب
79	bahar 81.2022	وطن	109	khosro 67.6798	چند
80	moulavi 81.0399	جمله	110	amir 67.4463	جهان
81	khosro 80.4363	وه	111	ghaani 67.0647	مهر
82	amir 79.957	کن	112	bahar 66.8593	تا
83	bahar 78.0198	در	113	amir 66.4195	همت
84	sanaee 77.5978	مرد	114	moulavi 66.3114	می‌کند
85	khosro 77.5245	سینه	115	amir 65.8704	ا
86	bahar 77.28	هر	116	amir 65.2325	کاو
87	moulavi 76.9478	همه	117	amir 65.0681	پیروزی
88	amir 76.4481	سعادت	118	khosro 65.0586	جانم
89	amir 76.198	بزم	119	bahar 64.3993	زن
90	sanaee 76.1147	همچو	120	ghaani 64.0744	بسکه
91	amir 75.0715	شدست	121	ghaani 62.7105	چرخ
92	moulavi 73.1872	ای	122	bahar 62.6226	بهار
93	khosro 73.0733	جگر	123	amir 62.297	سلطان
94	moulavi 72.4931	لیک	124	moulavi 61.9009	پس
95	khosro 72.3834	خوبان	125	amir 61.7995	روزگار
96	sanaee 72.1991	عالم	126	amir 61.7372	بر
97	amir 72.1794	گیتی	127	amir 61.4581	شاهی
98	khosro 71.6296	گل	128	moulavi 61.4327	جو
99	khosro 71.502	سرو	129	moulavi 61.0029	یا
100	khosro 71.4008	می‌رود	130	ghaani 60.9064	چهر
101	moulavi 71.0434	خدا	131	khosro 60.7171	خواهم
102	bahar 70.3954	چون	132	bahar 60.1104	زبن
103	khosro 70.1467	خواب	133	bahar 59.8384	مردم
104	amir 69.1232	زمین	134	moulavi 59.7287	بلک
105	ghaani 68.5088	باشد	135	khosro 59.4253	چشم
106	amir 68.157	شرف	136	bahar 59.3471	ری
107	ghaani 67.9703	سپهر	137	sanaee 59.3394	علم

138	amir 58.5589	شهریار	167	moulavi 50.5071	صد
139	khosro 58.3309	باز	166	ghaani 50.9126	ابر
140	ghaani 57.4641	از	165	khosro 51.2282	نتوان
141	khosro 57.2616	لبت	164	sanaee 51.5573	سرای
142	amir 57.2598	طبع	163	moulavi 52.2313	خدای
143	ghaani 57.028	بحر	162	sanaee 52.6193	دان
144	ghaani 56.4825	رخش	161	moulavi 52.6639	فلک
145	khosro 56.4453	کویت	160	amir 52.684	حشمت
146	ghaani 56.0983	فارس	159	amir 52.7809	فخر
147	khosro 55.7677	سوخته	158	bahar 53.5213	خراسان
148	khosro 55.4554	دیوانه	157	ghaani 53.5748	خصم
149	moulavi 55.1289	اگر	156	moulavi 53.5968	آنکه
150	bahar 55.047	کشور	155	moulavi 53.8886	روی
151	amir 54.9973	کلک	154	amir 53.9267	فر
152	khosro 54.6488	مست	153	khosro 54.5215	باری

168 amir 50.4138 ایزد	179 sanaee 49.399 شود	190 moulavi 47.2555 پیغامبر
169 moulavi 50.27 خر	180 moulavi 48.9541 بود	191 moulavi 47.0659 آنک
170 amir 50.1863 گردون	181 khosro 48.8482 درت	192 amir 46.6101 لولو
171 moulavi 50.1533 بانگ	182 amir 48.4465 ار	193 amir 46.0548 ماه
172 sanaee 50.0362 پی	183 moulavi 48.3363 لا	194 moulavi 45.2583 نی
173 sanaee 50.0309 داشتن	184 sanaee 48.316 راه	195 khosro 45.2492 کرشمه
174 khosro 49.7688 غمت	185 amir 48.2713 ملوک	196 amir 45.1576 تویی
175 khosro 49.7485 شب	186 sanaee 48.0678 ورا	197 khosro 45.0937 کشتن
176 amir 49.6632 جهانداری	187 khosro 47.8566 شبی	198 amir 44.7922 سنجر
177 amir 49.4439 ملت	188 khosro 47.854 هرکه	199 amir 44.6187 ایا
178 bahar 49.4094 گر	189 ghaani 47.8085 پیکر	

جنس اعداد بدست آمده در این دو روش متفاوت است. زیرا در X^2 از تعداد استفاده می‌شود و در IG از احتمال.

تحلیل:

همانطور که مشاهده می‌شود برخی کلمات چون گر، به، در و از این قبیل به عنوان کلمات مهم استخراج شده است. این به این دلیل است که کلمات پرتکرار حذف نشده اند و دلیل حذف نشدن نیز در بخش اول توضیح داده شده است. اگر کلمات پرتکرار حذف می‌شدند ۲۰۰ کلمه‌ی مهم، کلماتی بود که ممکن بود در داده‌های تست وجود نداشته باشد و از این لحاظ در قسمت بعدی تمرین دچار مشکل می‌شدیم.

مقایسه:

تا حد خیلی زیادی کلمات در هر دو روش یکسان هستند و تنها در رتبه‌شان تفاوت است. این امر در بخش بعد قسمت اول نیز مشهود است. زیرا معیارهای ارزیابی در این دو حالت نتایج بسیار شبیه به یکدیگری دارند. اگر کلمات در این دو روش متفاوت بودند خروجی این معیارها نیز متفاوت می‌شد.

بخش چهارم

در این قسمت برای محاسبه‌ی هر کدام از معیارهای ارزیابی از کتابخانه‌ی sklearn در پایتون استفاده شده است. برای دسته‌بندی بیز ساده، از رابطه‌ی زیر استفاده شده‌است:

$$\hat{c} = \text{argmax}_{c_i} P(d|c_i) \cdot P(c_i)$$

قسمت اول:

که در حالت یونیگرم احتمال هر بیت به صورت ضرب احتمال یونیگرم‌های داده‌ی تست در صورت وجود در ۲۰۰ ویژگی بدست آمده در قسمت قبل است.

خروجی در حالت IG:

f1_score , macro average is : 0.3233113653457745

f1_score, micro average is : 0.3756296296296296

precision, macro average is : 0.4429633459476863

precision, micro average is : 0.37562962962962965

recall, macro average is : 0.37900444992178417

recall, micro average is : 0.37562962962962965

خروجی در حالت x^2 :

f1_score , macro average is : 0.32097750946882814

f1_score, micro average is : 0.3727407407407407

precision, macro average is : 0.44405976780431106

precision, micro average is : 0.37274074074074076

recall, macro average is : 0.376638124400618

recall, micro average is : 0.37274074074074076

قسمت دوم:

در این حالت اگر کلمه‌ای در در دیکشنری وجود نداشت به روش هموارسازی absolute discounting احتمال آن کلمه محاسبه شده‌است.

خروجی این قسمت:

f1_score , macro average is : 0.4590946186362401

f1_score, micro average is : 0.4577037037037037

precision, macro average is : 0.5549190622703205

precision, micro average is : 0.4577037037037037

recall, macro average is : 0.4497288408460512

recall, micro average is : 0.4577037037037037

قسمت سوم:

در حالت دلتا ۰.۱ :

f1_score , macro average is : 0.6253361797522307

f1_score, micro average is : 0.6286666666666667

precision, macro average is : 0.6376574210073086

precision, micro average is : 0.6286666666666667

recall, macro average is : 0.6298981516669863

recall, micro average is : 0.6298981516669863

درحالت دلتا ۰.۴ :

f1_score , macro average is : 0.5961529118174586

f1_score, micro average is : 0.6037777777777777

precision, macro average is : 0.6229425383083792

precision, micro average is : 0.6037777777777777

recall, macro average is : 0.6046523710235708

recall, micro average is : 0.6046523710235708

در حالت دلتا ۰.۷:

f1_score , macro average is : 0.5714851389766665

f1_score, micro average is : 0.5817777777777777

precision, macro average is : 0.6105886773296654

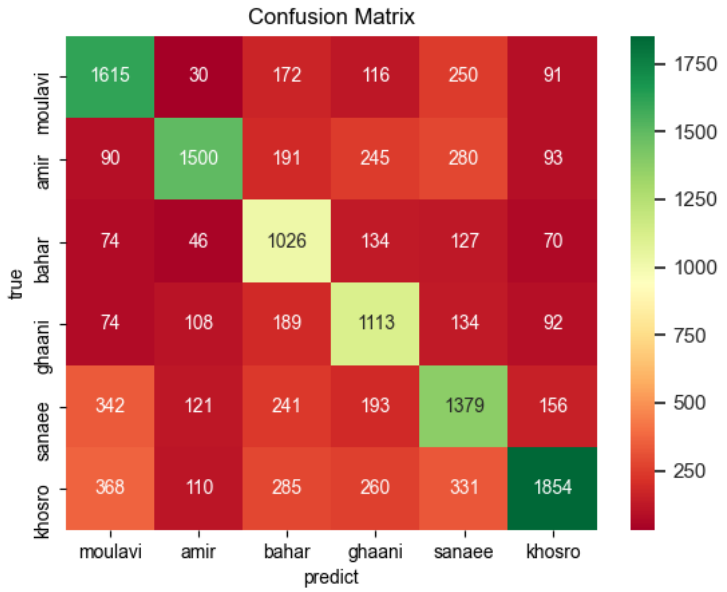
precision, micro average is : 0.5817777777777777

recall, macro average is : 0.5827475589983601

recall, micro average is : 0.5827475589983601

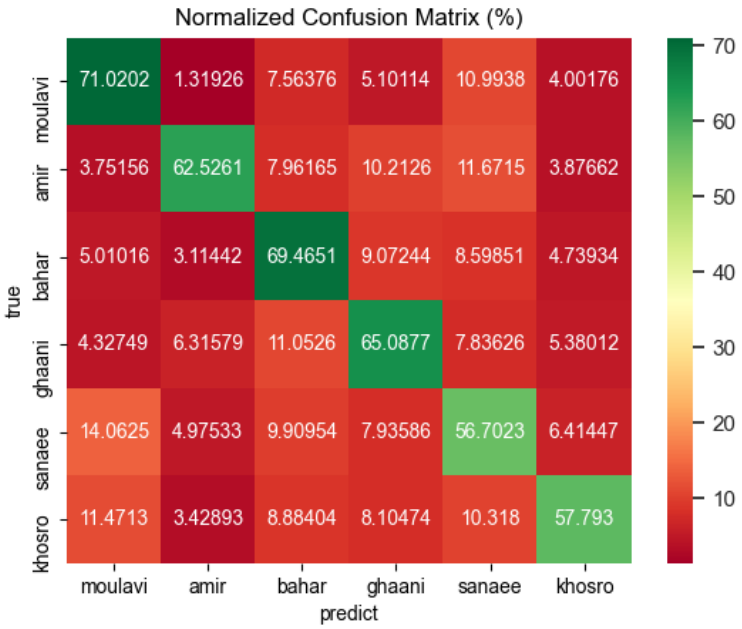
با توجه به نتایج بدست آمده مدل قسمت سوم با دلتای ۰.۱ بهترین مدل است. از این مدل در قسمت بعد استفاده می‌شود.

بخش پنجم



شکل ۵

در شکل ۵ ماتریس درهم ریختگی قابل مشاهده است. برای تصمیم‌گیری در مورد میزان شباهت شعرا لازم بود این تعدادها نرمالایز شوند. چون تعداد اشعار شعرا یکسان نبود. به همین دلیل اعداد هر سطر تقسیم بر تعداد ابیات همان شاعر در مجموعه‌ی آزمون شده‌اند و در شکل ۶ قابل مشاهده است.



شکل ۶

براساس این نمودار دو شاعر ثنایی و مولوی بیشترین شباهت را بهم دارند. زیرا ۱۴٪ از ابیات این دوشاعر به صورت اشتباه دسته‌بندی شده است. کمترین شباهت‌ها مربوط به دو شاعر مولوی و امیر است زیرا تنها ۱٪ از ابیات این دو شاعر اشتباه دسته‌بندی شده است.(دو شاعری که شباهت بیشتری بهم دارند مدل در هنگام دسته‌بندی بیشتر دچار خطا می‌شود. البته بهتر بود از جمع دو درایه‌ی متناظر برای هر دو شاعر استفاده می‌کردم ولی به صورت کلی این حالت فعلی هم اشتباه نیست.)

از نظر بنده دو زوج سخت: مولوی و ثنایی دو زوج آسان: امیر و مولوی دو زوج متوسط: امیر و قآنی (با درصد خطای ۶.۳٪)

خروجی‌ها بر اساس هر زوج :

f1_score , macro average is : 0.8043260776163546 #hard

f1_score , macro average is : 0.9208916381036756 #easy

f1_score , macro average is : 0.8606638428800599 #intermediate

تحلیل: در دو شاعر سخت همچنان دقت کمتر از دو زوج دیگر است. در دو شاعر آسان ابیات به خوبی تفکیک شده اند و در حالت متوسط عمکرد بین دو حالت قبلی است و این خروجی‌ها کاملا در جهت انتظارمان هستند.