گزارش پروژه کامپیوتری صفرم درس آمار و احتمال

بابک حسینی محتشم 810101408 تاریخ: 1402/7/29

کتابخانه های استفاده شده: hazm,re,numpy,pandas

تعریف متغیرها:

categoriesدارای تمام موضوعات موجود در داده ها است و به هر موضوع یک عدد اختصاص میدهد.

cnt_categories_books به هر عدد هر موضوع در categories تعداد کتاب با آن موضوع را اختصاص میدهد.

Words به هر لغت منحصر به فرد عددی منحصر به فرد اختصاص میدهد.

words_processed به هر لغت منحصر به فرد، لغتی تغییر یافته به نحو خواسته شده در بخش های امتیازی اختصاص میدهد.

کلاس book دارای دو زیرکلاس book_train و book_test است.

متغیر config در این کلاس ها انجام یا ندادن قسمت های مختلف بخش های امتیازی را مشخص میکند.

در:config

do_additive_smoothing انجام دادن یا ندادن smoothing additive را مشخص میکند که در صورت False بودن یک بودن
no_additive_smoothing_probability برابریک در نظر گرفته میشود و در صورت صفر بودن no_additive_smoothing_probability برابر صفر در نظر گرفته میشود.

remove_stop_words جدا کردن یا نکردن لغات اضافی را مشخص میکند.

do_ lemmatization استفاده کردن یا نکردن از تابع lemmatizer هضم را مشخص میکند.

do_stemmation استفاده کردن یا نکردن از تابع stemmer هضم را مشخص میکند.

times_to_add_title تعداد دفعاتی که موضوع کتاب به توضیحاتش اضافه شود را مشخص میکند.

additive_smoothing_alpha پارامتر عددی به کار گرفته شده در فرمول additive smoothing.

ابتدا کتاب های آموزش ساخته میشوند و سپس متن شان با توجه به config پردازش میشود سپس bag of words کلمات آموزش ساخته میشود. بعد از آن کتاب های فایل آزمایش ساخته و پردازش میشوند و در نهایت طبق داده های به دست آمده احتمال های مختلف را به دست آورده و موضوع محتمل تر برای هر کتاب آزمایش را مشخص میکنیم و دقت پیشبینی ها را به دست می آوریم.

additive_smoothing	stop_words	lemmatization	stemmation	add_title=10 alpha=10	add_title=0 alpha=10	add_title=10 alpha=1	add_title=0 alpha=1
T	T	T	T	81.33	79.33	82.0	81.78
T	T	T	F	81.33	80.44	82.44	80.89
T	T	F	T	80.89	80.44	81.78	81.11
T	T	F	F	82.67	81.78	81.78	80.0
T	F	T	T	82.67	76.67	80.44	78.44
T	F	T	F	82.67	79.11	80.89	79.56
T	F	F	T	82.67	76.22	80.67	78.22
T	F	F	F	82.67	77.56	80.89	78.44
F	T	T	T	3.33	3.33	3.33	3.33
F	T	T	F	2.22	2.89	2.22	2.89
F	T	F	T	3.11	3.56	3.11	3.56
F	T	F	F	2.00	1.78	2.00	1.78
F	F	T	T	3.78	3.11	3.78	3.78
F	F	T	F	2.44	3.11	2.45	3.11
F	F	F	T	2.44	3.78	3.56	3.78
F	F	F	F	2.22	1.78	2.22	1.78

در جدول فوق درصد دقت حالات مختلف گزارش شده و در حالت False بودن do_additive_smoothing مقدار no_additive_smoothing_probability برابر یک یعنی احتمال لغات جدید یک در نظر گرفته شده ولی برای حالتی که احتمال لغات جدید را صفر در نظر بگیریم احتمال تمام حالاتی که lemmatization و stemmation انجام میشوند حاصل 16.44% و برای تمام حالاتی که این دو انجام نشوند دقت 16.22% گرفتم. که توجیه ریاضی این اعداد این است که چون در اکثر کتاب ها لغات جدید یافت میشود احتمال در اکثر کتاب ها برابر 0 میشود و کتاب به صورت تصادفی درست یا نا درست پیشبینی میشود که چون شش موضوع داریم احتمال انتخاب موضوع درست میشود 1/6.

پاسخ پرسش ها:

۱- اگر موقع حساب کردن احتمال ها با کلمه ای مواجه شویم که در فایل آموزش نبوده و یا برای محاسبه احتمال کلماتی که در فایل آموزشی برای بعضی از موضوع ها نیامده باشند : اگر احتمال آن را صفر در نظر بگیریم با توجه به رابطه ریاضی کل احتمال برابر صفر میشود که مورد پسند نیست زیرا ممکن است که آن موضوعی که صفر میشود در واقع موضوع آن کتاب باشد .اگر آن کلمه را در نظر نگیریم مانند آن است که احتمالش را یک بگیریم که از لحاظ ریاضیاتی توجیه ندارد و همچنین طبق نتایج جدول باال دقت بسیار پایین می آید .روش بهتر استفاده از smoothing additive است که احتمال آن را بسیار کم میشود و دقت مناسبی میدهد ولی با توجه به مشخص نبودن پارامتر آلفای فرمول شاید بتوان حتی دقت بهتری نیز کسب کرد.

۲- چون تعداد کلمات زیاد است و احتمال هر کلمه عددی بین صفر و یک است حاصل ضرب این اعداد عددی بسیار کوچک میشود و باعث وقوع underflowو صفر در نظر گرفته شدن احتمال میشود ولی اگر از دو. طرف رابطه لگاریتم بگیریم به حاصل جمع تعداد عبارت میرسیم که با مقایسه این حاصل جمع ها میتوان موضوع محتمل تر را یافت. باید توجه کرد که اگر از رابطه قبل استفاده شود و جلوی underflow گرفته شود مثال با ضرب هر احتمال در ۱۰۰۰ نتایج یکسانی با روش لگاریتم به دست می آید که از لحاظ ریاضی یکسان بودن این دو رابطه را توجیه میکند.