بسم الله الرحمن الرحیم

عنوان پروژه:پردازش دیتاست اخبار

نام درس:بازیابی اطلاعات

استاد محترم:استاد شایق بروجنی

اعضای گروه : حانیه دهقان – نگین آخرتی – یگانه نعمتی

زمستان ۱۴۰۱

چکیده:

هدف اصلی ما استفاده از tf idf بر روی دیتاست ها ی متنی ست که با تکرار کلمات در یک سند مرتبط است.

برای این منظور دیتاست خبرها را انتخاب نمودیم و مراحل زیر را انجام دادیم

۱-وارد کردن یا ایمپورت کتابخانه های مورد لزوم

۲-بارگذاری یا لود کردن دیتاست

۳-مقدار دهی و انتقال اطلاعات از دیتاست با استفاده از TfidfVectorizer

۴-برگرداندن اطلاعات به دیتافریم و انتقال ان به اکسل

۵-انتقال از دیتافریم به جدول(pandastable) و چاپ ان در محیط Tkinter

مقدمه:

ما انسانها به موضوعات خاصی علاقه بیشتری نشان میدهیم.بنا براین یکی با شنیدن فوتبال گوشهایش تیز شود و دیگری با شنیدن مسایل علمی.در واقع ما اینجا با دو خبر ورزشی و علمی سر و کار داریم پس طبیعی ست برای پژوهشگری که میخواهد بر روی خبرها کار کند انها را دسته بندی نماید تا بتواند از روی هر خبری پیشگویی لازم را در مورد نوع ان انجام دهد.بنا براین او به الگوریتمهایی برای ساخت مدلش نیاز دارد تا این دسته بندی ها را انجام دهد. ضمن انکه او نیاز دارد تا مدلش را از حالت دستی خارج نماید و مکانیزه کند.زبان پایتون برای چنین مواردی بسیار قدرتمند است .توابع کتابخانه ای اماده که در واقع تبدیل الگوریتمهای دانشمندان داده پردازی به زبان برنامه است به پژوهشگر این امکان را میدهد تا پردازشش را با سرعت بیشتری انجام دهد.از انجا که دسته بندی ۰و۱ یا باینری ساده ترین روش ممکن است طبیعی ست پردازش دیتاست اخبار مشکلاتی چند نیز داشته باشد که از تنوع اخبار و تعداد کلاسها ناشی می شود .حتی این طبقه بندی ها میتواند برای خبرهای طولانی(شامل دیتاست ما) نادرست نیز باشد و ماشین را به گمراهی بیندازد.مثلا در بازی بین بارسلونا و مادرید که با درخشش مسی همراه بود و با نتیجه ۲-۰ بسود بارسا تمام شد تماشاچیان با یکدیگر درگیر شدند و برای استقلال کاتالانیا شعار دادند.این خبر ورزشی ست یا سیاسی و در کدام طبقه قرار میگیرد. اینجا بحث دیگری پیش می اید و ان اینکه انتخاب دیتاست چگونه باید باشد. ایا خبرهای طولانی میتواند از دقت مدل ما بکاهد؟ یکی از اصولی که تقریبا تمامی دانشمندان داده پردازی به ان معتقدند کیفیت دیتاست انتخابی برای پردازش داده هاست .ما به همه مواردی که میتواند از کیفیت داده ها کم کند کاری نداریم اما برای پروژه ما دو مورد بطرز چشمگیری خود را نمایان می سازد. یکی طولانی بودن سند (یا خبر) می باشد که می تواند ایجاد چند پهلویی و اشتباه در انتخاب دسته بندی نماید و دیگری کلمات نامربوط(stopwords) بوده که باید به ان پرداخته شود.ماشین معمولا بر اساس وزن کلمات میتواند پیشگویی اش را انجام دهد او با خواندن کلماتی (مثل فوتبال –دروازه بان – فوروارد – اسپک – کول انداز و غیره)که درسندهایی که در کلاس ورزشی قرار دارند حدس می زند خبرهای ورزشی باید شامل این کلمات باشد و نسبت به انها حساس شده و خبر جدید را پیشگویی میکند.اما معمولا خبر کلمات دیگری نیز دارد که با انها ادغام میشود تا تبدیل به جمله یا جملاتی گردد.مثلا اگر بگوییم یزدانی با بارانداز پیروز شد.کلمات با , شد و حتی پیروزی (stopwords) چه نقشی در

ورزشی بودن خبر دارند.بنا بر این سند خبر باید بشود یزدانی بارانداز.در واقع این سند برای پردازش مناسب است و نه خبر قبلی. اینها میتوانند پیش گویی ما را به چالش بکشند و از دقت مدل بکاهند.پس با انها باید مثل افت در یک زمین زراعی برخورد کرد . برای حذف این کلمات اضافی در متون انگلیسی و برخی کشورهای دیگر میتوان از توابعی استفاده نمود و این مشکل را تا حدود زیادی حل نمود اما برای زبان فارسی هنوز کار جدی (تا انجا که من میدانم) انجام نشده است.برای حل این مورد ما دو راه حل داریم یا باید لیستی از کلمات مزاحم در برنامه درست کنیم و یا در اکسل ردیف یا ستونی را به این کلمات اختصاص دهیم و انها را لود کرده و از دیتاست حذف نماییم که ما از هر دو روش میتوانیم در برنامه مان استفاده کنیم. اما باید اذعان نمود که این کاری یکی دو روزه نیست و ما نیز نخواستیم انها را کاملا حذف نماییم بلکه هدف ما ایجاد متدی بود که بتوانیم ان را به مرور انجام دهیم.متاسفانه در رابطه با زبان فارسی کار چشمگیری انجام نشده است در صورتی که حتی کشورهای عربی نیز در این زمینه اقداماتی جدی تری از ما انجام داده اند.

انچه که در بالا گفته شد در واقع مشکلات موجود بود اما این به هیچ وجه از ارزش پروژه ای که انجامش می دهیم نمیکاهد.این کار برای اغاز راهی ست که میتواند در پروسه زمان کامل تر شود . در این پروژه در ارتباط با دیتاست ما با دو ستون text و label_id سر و کار داریم که به ترتیب بیانگر خبرها و کلاسها (بر اساس اعداد از ۰ تا ۷) می باشد.ما اطلاعات اخبار را از فایل اکسل واکشی کرده و با تی اف آی دی اف و توابع کتابخانه ای چون svm,knnو dt و mnb محاسباتی را بر روی انها انجام دادیم و نتایج حاصله را نیز نشان دادیم .اما برای مطالعه بر روی خروجی ها فقط به نماش بر روی صفحه کامپیوتر بسنده نکردیم بلکه انها را در محیط اکسل نیز ذخیره نمودیم تا پژوهشگر بتواند انها را مطالعه نموده و کیفیت کار را با تغییراتی که لازم میداند انجام دهد بالاتر ببرد

جزییات برنامه نویسی

-خطوط ۱ تا ۴۴ ایمپورت کتابخانه های مورد لزوم برای نوشتن برنامه

-۴۶ تا ۸۶ ایجاد منو های لازم.این کار بیشتر به این دلیل انجام میشود تا شخص بتواند اطلاعات لازم را جداگانه بگیرد و بر روی انها مطالعه نماید وگرنه همه اطلاعات در یک جا می تواند ایجاد خستگی و گمراهی نماید.منوها شامل نشان دادن جداول بردار ویژگی ها و ویژگی ها در جدول پاندا و محیط اکسل می باشد.جدول ویژگی ها در D:\my\news1.xlsx و جدول بردار ویژگی ها در D:\my\news2.xlsx ذخیره شده اند.که همه انها در منوی tf idf قرار دارند.منوی spli dataset دیتاست را به سه دسته اموزش ازمون و اعتبار سنجی جدا میکند و تعداد سطر و ستون هر کدام را گزارش مینماید.منوی calculation با دو زیر منو دقت را با روشهای svm,knnو dt و mnb و prediction, recall, f1 score , accuracy را محاسبه میکند.عملکرد منوی confusion matrix از نامش پیداست و در اخرین منو نیز کادری ظاهر میشود که شما میتوانید پیشگویی خبر را مشاهده نمایید.مثلا اگر بنویسیم تیم فوتبال استقلال دیروز برنده شد پیغامی به این شکل ظاهر میشود.news is sport

-۱۱۵تا۱۲۲ تابعی برای خواندن یک خبر و پیش گویی کلاس مربوط به ان می باشد.یعنی یک تکست باکس باز میشود و ما میتوانیم خبر را وارد نماییم .سپس این تابع تابعی دیگر را (از خطوط ۸۸ تا ۱۱۰) را صدا میزند که با MultinomialNBداده

های ازمایشی را با () fit به عنوان ارگومان میگیرد و سپس با x = mnb.predict(prd) شماره کلاس را بدست می اورد و پیغامی مناسب با شماره کلاس صادر میکند.

-خطوط ۱۴۰ تا ۱۵۹ با سه تابع مجزا دیتاست را به سه دسته اموزشی و ازمون و اعتبار سنجی تفسیم نمودیم که خود توابع واضح می باشد.این سه تابع هر کدام تابع دیگری را صدا می زنند(print_df) که وظیفه ان ایجاد یک صفحه جدید .ایجاد دیتا فریم و جدول پاندا و ریختن اطلاعات داخل جدول می باشد.

-خطوط ۱۶۶تا ۱۸۶ ابعاد ذاذه های اموزشی تست و اعتبار سنجی را مشخص میکند.برای این منظور داده ها ی تفکیک شده را در دیتافریم های مختلف میریزد و تعداد سطر و ستونشان را مشخص میکند.سپس دیکشنری از این داده ها ایجاد میکند و با تابع دیگری print_df ان را چاپ میکند که تابع print_df قبلا توضیح داده شد.

-خطوط 298تا 305 ویژگی ها را در جدول پاندا نشان میدهد که از دیکشنری feu که در خط ۳۷۵ تعریف شده استفاده می نماید.

-۳۰۶ تا ۳۰۸ یک دیتا فریم ویژگی ها را استخراج و ان را در فایل اکسل ذخیره میکند

-۳۳۶ تا ۳۴۱ بردار ویژگی ها را در یک پاندا تیبل می ریزد.برای اینکار تابع دیگری بنام return_fv را صدا می زند که کار ان ایجاد دیتافریم(از خط 324 تا 328) با سل های بردار ویژگی بر حسب ارایه و ستون ویژگی ها می باشد(خط 327).

-خطوط 309 تا 312 توضیحات پاراگراف قبلی را در اکسل می ریزد.

-خطوط ۲۱۶ تا ۲۳۲ ماتریس سردرگمی را ترسیم مینماید.توضیح خطوط ان نیز پیچیدگی خاصی ندارد .در واقع confusion_matrix با دو تابع هدف سرو کار دارد یکی انهایی که درستند یا مقدار واقعی اند و یکی انهایی که از حدس روی داده های ازمون به وجود می ایند(خط 255) بقیه خطوط پایین ان در واقع فرمت و رسم ماتریس را نشان میدهند.

-خطوط ۲۸۲ تا ۲۹۴ محاسبه accuracy_score را با استفاده از svc, knn, mnb, 'dt' بر میگرداند که چیز خاصی برای توضیح دادن وجود ندارد مگر حلقه for که در ان با استفاده از روشهای svc,... بر روی داده های اموزشی و تست پردازش انجام میدهد و دقت را محاسبه میکند.

-خطوط ۱۸۷ تا ۲۲۹ برای محاسبه چهارگانه precision و recall و f1-score و accuracy می باشند که از دو تابع p_f_r_a2 و p_f_r_a1 استفاده شده است . از p_f_r_a1 با استفاده از scv محاسبات چهار گانه انجام شده است و p_f_r_a2 محاسبات چهارگانه را برای هر کلاس انجام میدهد .

-در خطوط ۳۵۴ تا ۳۵۹ محیط گرافیکی tkinter تعریف شده و برای پنجره ایجاد شده در این محیط ابعادی تعریف گردید.

-در خط ۳۶۰ داده های ما از اکسل لود میشود.

-در واقع خطوط حیاتی برنامه ۳۷۰ تا ۳۸۰ می باشد که در آن دو تابع کلیدی CountVectorizer و TfidfVectorizer وجود دارند که در واقع اولی تعداد دفعاتی را که یک کلمه در یک سند ظاهر می شود شمارش میکند و دومی نه تنها این

شمارش را در نظر میگیرد بلکه میزان اهمیت ان را برای همه سندها در نظر میگیرد و منظور ما از سند یک خبر یا یک سل از ستون text میباشد و همه سندها همه ردیفهای شامل خبر یا ستون خبر می باشد.

-اما هنوز یک مورد پر اهمیت توضیح داده نشده وان هم حذف stopwords می باشد.در خط ۳۷۰ ما ستون خبر(در دیتاست ما text) را به تابع delet_stopword میفرستیم تا ان را پالایش نماییم.تابع مذکور در خط ۳۴۲ این مهم را انجام میدهد .ما برای تست چند کلمه زاید را در لیستی قرار دادیم تا انها از دیتاست حذف گردند. ضمن اینکه توضیح این نکته ضروری ست که اساس کلاسه کردن داده های متنی این است که جملات به کلمات مجزا تقسیم میشوند و همه بحثها بر روی این کلمات انجام میگیرد مثلا گفت: یک گلمه محسوب میشود و مجزا حساب نمیگردند مگر اینکه ما در اکسل با استفاده از find and replace بین انها فاصله بیندازیم تا : قابل حذف گردد.

لازم است به دو نکته حائز اهمیت جهت اجرا اشاره کنم:

۱-بدلیل حجم بالای fv ارسال ان به اکسل با مشکل مواجه می شود مگر انکه df1.to_excel(r'D:\my\news2.xlsx') را به df1.head(200).to_excel(r'D:\my\news2.xlsx')تغییر دهید(خط ۳۱۲).که ۲۰۰ ردیف را میریزد البته ۲۰۰ یک مثال است و شما میتوانید از ۵۰۰ یا بیشتر نیز استفاده نمایید.

۲-فایل اصلی dev1.xlsxاست.چون فایل ارسالی شما دو ردیف nan داشت که ان را حذف کردم.

ارزیابی نتایج:

پیش بینی: با ازمونهایی که انجام دادم نتایج برای خبرهای کوتاه دقت بالایی دارد.مثلا خبر:تیم فوتبال استقلال برنده بازی دیروز بود .بازخورد برنامه خبر ورزسی ست و یا خبر :جنگ سرد بین روسیه و امریکا از سر گرفته شد بازخورد ان سیاسی ست .یا اگر قسمتی از یک خبر را کپی past نماییم باز بازخورد صحیح است .در هرصورت چسبندگی کلمات حجم دیتاست و نبود تابعی برای حذف کلمات نامربوط و طولانی بودن خبرها از کیفیت دیتاست ما می کاهد .

precision و recall و f1-score و accuracy : از نتایج حاصله چنین برمی اید که کلاسهای ۰ و ۳و ۷ یعنی خبرهای اجتماعی سیاسی و پزشکی از دقت کمتری برخوردار است و این میتواند بخاطر نوع این خبرها و یا نبود کلمات وزین در انها که به میزان کافی تکرار نشده باشند برگردد. خبرهای ورزشی و اقتصادی از دقت بالای ۹۰٪ برخوردارند که به دلایل گفته شده برمیگردد.در هر صورت بنظر من با توجه به کیفیت دیتاست مقادیر بدست امده قابل قبول میباشند.

ماتریس سردرگمی:هر سطر ماتریس بیانگر یک کلاس است مثلا برای کلاس ۰ (اجتماعی) مطابق ماتریس ترسیمی ۱۵ مورد در مدل بدرستی پیش بینی شده است و یک مورد به اشتباه کلاس ۰ به عنوان کلاس یک طبقه بندی شده و هیچ موردی به عنوان کلاس دو طبقه بندی نشده و دو مورد به اشتباه کلاس ۳ طبقه بندی شده و ... و ۴ مورد (اخرین ستون ردیف یک یا کلاس ۰) به اشتباه به عنوان کلاس ۷ گرفته شده(معنی ساده ان این است :پیش بینی کننده ۴ بار خبر سیاسی را به اشتباه خبر پزشکی حدس زده است) و به همین ترتیب میتوان ادامه داد اخر.امیدوارم این پروژه مفید وقع شده باشد.