بازیابی هوشمند اطلاعات

تمرین چهارم

نام و نام خانوادگی : حسین سیفی

شماره دانشجویی : ۸۱۰۱۰۰۳۸۶

دی ۱۴۰۰

فهرست

بخش ۱ – Word Association

در این تمرین هدف استخراج روابط syntagmatic و paradigmatic بین کلمات با استفاده از Mutual Information است. روش کار به این صورت است که مقدار MI را برای هر زوج کلمه محاسبه می‌کنیم. مقادیر MI برای یک کلمه در مواجهه با کلمات دیگر نشان دهنده روابط Paradigmatic است و با ضرب همین بردار به عنوان بردار Context در بردار Context کلمات دیگر ، می‌توان روابط Syntagmatic را به دست آورد.

به عبارتی دیگر با به دست آوردن بردار MI هر کلمه در صورت مواجهه با کلمات دیگر نشان می‌دهیم که کدام کلمات بیشتر با همنشینی یکدیگر ظاهر شده‌اند تا به صورت تکی و با محاسبه شباهت کسینوسی همین مقادیر به عنوان بردار Context در بردار Context کلمات دیگر ، بررسی می‌کنیم که کدام کلمات با کلمات مشابهی همنشین هستند و در نتیجه قابل جانشینی با یکدیگر هستند.

سوال ۱

برای به دست آوردن بردار Context هر کدام از کلمات از فرمول زیر استفاده می‌کنیم.

$$I(X_{w1}; X_{w2}) = \sum_{u \in \{0,1\}} \sum_{v \in \{0,1\}} p(X_{w1} = u,\ X_{w2} = v)\, log_2 \frac{p(X_{w1} = u,\ X_{w2} = v)}{p(X_{w1} = u) p(X_{w2} = v)}$$

با استفاده از فرمول فوق و کمک گرفتن از ماتریس رویداد همزمان کلمات ، می‌توان Mutual Information را برای هر زوج کلمه در لغت‌نامه(Vocabulary) اسناد محاسبه کرد و در ادامه سوالات بخش ۱ مقدار MI یک کلمه به صورت زوج با سایر کلمات به عنوان بردار Context و برای یافتن روابط همنشینی و جانشینی کاربرد دارد.

سوال ۲

ده کلمه‌ای که بیشترین روابط Syntagmatic را با کلمات Teacher و Iran دارند ، به شرح زیر می‌باشند:

رتبه	*Teacher*	*Iran*
۱	fleischer	algier
۲	hardli	battlefield
۳	strap	battleship
۴	brasilia	civilian
۵	immun	colder
۶	laden	dissoci
۷	accord	efficaci
۸	adjust	enemi
۹	associ	es
۱۰	began	homeland

سوال ۳

ده کلمه‌ای که بیشترین روابط Paradigmatic را با کلمات Teacher و Iran دارند ، به شرح زیر می‌باشند:

رتبه	*Teacher*	*Iran*
۱	Laden	Iraq
۲	Brasilia	Iranian
۳	Grind	Missil
۴	Strap	Tehran
۵	Hardli	Hormuz
۶	Fleischer	Gulf
۷	Coalit	Attack
۸	Unrest	Naval
۹	Junior	Warship
۱۰	Bleak	Chines

سوال ۴

با بررسی جداول بالا متوجه می‌شویم که هر کدام از کلمات همنشین با کلمه Iran به نوعی با این کلمه مرتبط هستند و در اخبار متفاوت می‌توانند همزمان با کلمه Iran در متون متفاوت ظاهر شوند. همچنین با بررسی بردار Context کلمات ذکر شده که دارای مقدار بالایی برای رابطه Syntagmatic با کلمه Iran هستند مشخص می‌شود که این کلمات بردار Context تقریبا مشابه دارند و در نتیجه با کلمات یکسانی رابطه Paradigmatic دارند پس این کلمات می‌توانند جایگزین کلمه Iran باشند.

از طرفی دیگر به نظر می‌رسد کلمات دارای روابط Syntagmatic با کلمه Teacher ارتباط معنایی چندانی با آن ندارند در نتیجه با توجه به اینکه همین روابط Syntagmatic به عنوان بردار Context این کلمه استفاده می‌شوند ، روابط Paradigmatic نیز اعتبار پایینی به دست می‌آورند. با بررسی کلمات دارای ارتباط paradigmatic با کلمه Teacher متوجه می‌شویم که حدس درستی داشته‌ایم و به نظر نمی‌رسد که بتوانیم کلمه Teacher را با این کلمات جایگزین کنیم. دلیل این رویداد این است که کلمه Teacher احتمال بسیار پایینی برای ظهور در اخبار (برابر ۰.۰۰۰۲۷) دارد و تنها در یک سند ظاهر شده است.

پیاده سازی

برای پیاده سازی این تمرین از زبان پایتون و کتابخانه‌های متعددی استفاده شده است که به صورت خلاصه توضیح داده می‌شود. در این تمرین از کتابخانه nltk برای تبدیل اخبار به بردارهایی از کلمات و تکرار هرکدام استفاده شده است و از کتابخانه‌های دیگری مانند pandas ، numpy ، scikit-learn ، scipy ، tqdm و enchant کمک گرفتیم. در ابتدای کار فایل csv اخبار به کمک کتابخانه pandas داخل برنامه بارگزاری می‌شود و پیش‌پردازشی شامل حذف علائم نگارشی ، حذف کلمات بی‌معنی ، حذف اعداد و ریشه‌یابی روی اخبار به کمک کتابخانه‌های مرتبط انجام می‌گیرد. در ادامه تعداد تکرار کلمات در سندهای مختلف محاسبه می‌شود ، لغت‌نامه‌ای شامل تمام کلمات اخبار ایجاد می‌شود ، تعداد سندهای شامل هر کلمه مشخص می‌شود ، کلمات پرتکرار و عمومی که تکراری بیشتر از ۴۰۰ بار دارند حذف می‌شوند و ماتریسی شامل تعداد ظهور همزمان هر جفت کلمه ساخته می‌شود. سپس مقدار

mutual Information برای هر جفت کلمه با استفاده از فرمول ذکر شده در سوال ۱ محاسبه می‌شود. این مقادیر نشان دهنده ارتباط همنشینی می‌باشند و با محاسبه فاصله کسینوسی بین بردار کلمات همنشین(Context) کلمات مختلف می‌توان روابط جانشینی را برای هر کلمه به دست آورد.

بخش ۲ – Clustering

در بخش ۲ این تمرین به Train کردن مدلهای متفاوت خوشه بندی با استفاده از Dataset اخبار و ارزیابی بر اساس Labelهای موجود برای هر کدام از این اخبار می‌پردازیم.

سوال ۱

مقدار چهار معیار ارزیابی خواسته شده برای روش kmeans به شرح زیر می‌باشد:

معیار ارزیابی	*مقدار*
Purity	۰.۷۹۷
F1	۰.۳۹۰
NMI	۰.۴۸۴
RI	۰.۶۹۷

برای محاسبه معیارهای F1 ، NMI و RI توابع آماده کتابخانه scikit-learn استفاده شد و برای محاسبه Purity با اندکی جستجو و کمک گرفتن از ماتریس contingency نتیجه مطلوب به دست آمد. همچنین تعداد False Positive و False Negative به تفکیک کلاس در جدول زیر آورده شده است:

کلاس	*False Negative*	*False Positive*
Acq	۹۷	۵۳۸
Crude	۲۷	۱۸
Earn	۴۱۳	۵۴
Grain	۴۱	۰
Interest	۱۹۱	۰
Money-fx	۲۲۲	۰
Ship	۱۰۸	۰
Trade	۱۴	۵۰۳

مشاهده می‌شود که بیشترین تعداد False Negative مربوط به کلاس Earn و بیشترین تعداد False Positive مربوط به کلاس Acq می‌باشد. متاسفانه کلاس‌های کوچکتر به دلیل اینکه در هیچکدام از کلاس‌های مقصد بیشترین تعداد را نسبت به دیگر کلاس‌ها نداشتند ، هیچ کلاس اخصیص یافته‌ای ندارند و در نتیجه تمامی تعداد آنها به عنوان False Negative معرفی می‌شود.

سوال ۲

مقدار معیارهای ارزیابی ذکر شده برای سه روش سلسله مراتبی Complete link ، Average link و Single link به شرح زیر می‌باشد:

معیار ارزیابی	*Complete link*	*Average link*	*Single link*
Purity	۰.۶۱۴	۰.۵۵۳	۰.۵۱۷
F1	۰.۱۶۱	۰.۱۵۰	۰.۰۸۶
NMI	۰.۱۴۷	۰.۳۱۷	۰.۰۰۲
RI	۰.۶۲۴	۰.۶۳۳	۰.۳۵۸

به نوعی مهمترین معیار ارزیابی را می‌توان Normalized Mutual Information معرفی کرد و با توجه به جدول فوق بهترین مقدار برای این معیار در روش Average link و سپس با اختلاف کمی در روش Complete link به دست می‌آید اما معیاری مانند Purity اطلاعات کاملی به ما نمی‌دهد و فقط یکسان بودن اعضای خوشه‌های پیشبینی شده را به ما نشان می‌دهد.

سوال ۳

در مقایسه‌ی این چهار روش می‌توان بگوییم که بین این روش‌ها بهترین مقدار Rand Index متعلق به روش kmeans می‌باشد و این معیار نشان می‌دهد که Kmeans بیشترین مقدار داده‌هایی را دارد که در دسته‌ی درست قرار گرفته‌اند. سایر معیارهای ارزیابی NMI ، Purity و F1 نیز بیشترین مقدار خود را در روش Kmeans دارند. این نتایج به این معنی است که روش kmeans بیشترین مقدار یکسان بودن خوشه‌ها و بیشترین مقدار تعادل بین Precision و Recall را دارد. پس در مجموع بهترین روش بین روش‌های بررسی شده Kmeans می‌باشد و Average link با اختلاف بسیار زیادی پشت آن ایستاده است.

پیاده سازی

در پیاده‌سازی این بخش نیز همانند بخش ۱ ابتدا به پیش‌پردازش روی مجموعه داده می‌پردازیم و کلمات و کاراکترهایی که کمترین کمک را به ما در خوشه بندی اخبار می‌کنند یا ما را گمراه می‌کنند ، حذف می‌کنیم. ادامه کار از جمله به دست آوردن احتمال هر کلمه ، ساختن مدل‌ها و انجام تنظیمات آنها ، محاسبه معیارهای ارزیابی و به دست آوردن برچسب‌های پیشبینی شده توسط توابع پیشفرض کتابخانه scikit-learn انجام می‌گیرد.