# بازیابی هوشمند اطلاعات

## تمرین سوم

مینا فریدی
810100430

## بخش یک

سوال ۱)

ارتباط paradigmatic به این معناست که هر کلمه ای با چه کلمات دیگری قابلیت جایگزینی دارد. برای یافتن این ارتباط باید بردار contex هر کلمه را به دست بیاوریم و سپس برای اندازه گیری شباهت از روش کسینوسی استفاده می کنیم. بردار contex را میتوان به دو روش به دست آورد. روش EOWC و BM25.

در روش EOWC تعداد تکرار هر کلمه را بر تعداد کل کلمات داکیومنت تقسیم می کنیم.

در روش BM25 از فرمول زیر استفاده می کنیم:

$$BM25(w_i, d1) = \frac{(k+1)c(w_i, d1)}{c(w_i, d1) + k(1 - b + b * \frac{|d1|}{avdl})}$$

$$x_i = \frac{BM25(w_i, d1)}{\sum_{j=1}^{N} BM25(w_j, d1)} \qquad b \in [0,1] \quad k \in [0, +\infty)$$

فایل کد این نمایش برداری در فایل contex.py آورده شده است.

ابتدا مراحل پیش پردازش را روی کلمات انجام می دهیم. به ترتیب اول کلمات را توکن می کنیم سپس کلمات stop words را از آن ها حذف می کنیم. بعد نوبت به stem کردن می رسد.

پس از این مرحله ماتریس ترم ها در داکیومنت ها را می سازیم. برای ساختن این ماتریس از تابع countvectorizer استفاده می کنیم.

سوال ۲)

در ارتباط syntagmatic دنبال این هستیم که پیدا کنیم که چه کلماتی زیاد با هم تکرار می شوند. برای پیدا کردن این ارتباط برای دو کلمه ایران و تیچر از فرمول MI که فرمولی متقارن است استفاده می کنیم. از فرمول از مقادیر آنتروپی و آنتروپی نسبی به دست می آید.

برای محاسبه این رابطه باید ماتریس term document را بدست بیاوریم. این ماتریس که هر سطر آن مربوط به یک فایل است و هر ستون یک کلمه را نشان می دهد دارای درایه هایی است که نشان دهنده

تعداد تکرار هر کلمه در هر فایل است. سپس برای محاسبه میزان MI ستون مربوط به کلمه ایران و تیچر را در یک حلقه با تمامی ستون های کلمات دیگر به عنوان ورودی به تابع sklearn.metrics.mutual_info_score می دهیم که تابعی از کتابخانه ی sklearn می باشد.

در نهایت ۱۰ مقدار بیشتر MI را ذخیره کرده و ترم های مربوطه را پیدا می کنیم.

نتیجه حاصل به این شکل بوده است:

کلمه ایران:

offici 0.00729

tehran 0.00736

missil 0.01188

iran 0.02949

Iraq 0.00957

gulf 0.01105

iranian 0.00978

said 0.00747

oil 0.00863

Hormuz 0.00746

کلمه teacher:

bankwork 0.00175

strap 0.00175

hardli 0.00175

sao 0.00175

paulo 0.00175

unrest 0.00175

front 0.00175

fleischer 0.00175

strike 0.00175

educ 0.00175

مشاهده می شود که کلمات پیدا شده مرتبط هستند و نتیجه می گیریم تابع MI کارایی مناسبی دارد.

سوال ۳) در ارتباط paradigmatic دنبال این هستیم که کلماتی که با هم می توانند جایگزین شود را پیدا کنیم. برای این کار باید پنجره های کنار کلمه را بگردیم و سپس بین هر پنجره مورد نظر با تمام کلمات دیگر شباهت پیدا کنیم. برای این کار هر کلمه را بصورت بردار contex نشان می دهیم و پنجره های آن را با کلمه های دیگر مقایسه می کنیم.

سوال ۴)

تفاوت این دو نوع ارتباط در این است که در syntagmatic کلمات نمی توانند جایگزین هم بشوند مثلا ایران نمی تواند جایگزین کلمه oil شود اما این نشان می دهد که کلمه ایران همراه کلمه oil زیاد تکرار می شود. ارتباط paradigmatic به این معناست که دنبال کلمه های جایگزین می گردیم. مثلا کلمه ایران می تواند با کلمه NIGERIA جایگزین شود و نقش آن را بگیرد.

## بخش دو

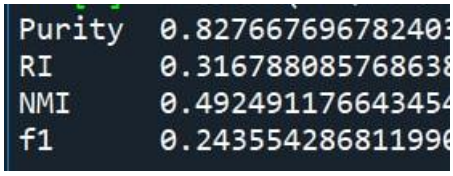
سوال ۱)

الگوریتم k means از مورد استفاده ترین الگوریتم‌های خوشه بندی است. در این الگوریتم k نشان دهنده تعداد خوشه ها است. این الگوریتم ابتدا k نقطه را به طور تصادفی انتخاب می کند که از آن ها به عنوان نقطه ثقل اولیه استفاده کند. در مرحله بعدی هر نقطه موجود در داده ها باید به نزدیک ترین نقطه ثقل موجود نسبت داده شود. سپس در دسته های جدید تشکیل یافته، مرکز ثقل هر دسته را (با میانگین گرفتن از مختصات نقطه ها) پیدا می کنیم و دوباره مراحل پیدا کردن نزدیک ترین مرکز ثقل و دسته بندی را انجام می دهیم تا به همگرایی برسیم.

برای ارزیابی دسته بندی ها می توان از روش های مختلفی استفاده کرد. در این تمرین از روش های purity، RI، F1 و NMI استفاده می کنیم.

در این سوال ابتدا پیش پردازش روی داده ها را انجام می دهیم. سپس با تابع TfidVectorizer، مقدار tfid آن ها را به دست می آوریم.

برای محاسبه دسته ها از طریق روش k means از تابع آماده kMeans که در scikit موجود است استفاده می‌کنیم و برای اندازه گیری معیارهای ارزیابی گفته شده از توابع purity_score که در کد پیاده شده و adjusted_rand_score، normalized_mutual_info_score، f1_score موجود در کتابخانه sklearn استفاده می کنیم. برای این که خروجی های بدست آمده و لیبل های قبلی را تطبیق دهیم بررسی می کنیم که کدام دو لیبل با هم بیشتر یکسان تکرار شده اند.




مشاهده می شود که مقادیر purity و NMI و RI بیشتر هستند.

برای پیدا کردن لیبلی که false negetive و false positive را افزایش می دهد در هر بار اجرا یکی از لیبل ها را حذف می کنیم و نتیجه می گیریم که لیبل های earn و interest مشکل زا هستند.

سوال ۲)

یکی از روش های معروف خوشه بندی روش خوشه بندی سلسه مراتبی است که در کل دو نوع هست: agglomerative وpartitioning . در روشِ agglomerative یا پایین به بالا هر نقطه را یک خوشه در نظر می گیریم و در هر مرحله خوشه ها را با هم ادغام می کنیم تا خوشه های بزرگتری تولید شوند. در آخر همه خوشه ها با هم یک خوشه می شوند در هنگام ادغام خوشه ها از روش های مختلف اندازه گیری شباهت می توان استفاده کرد. سه روش single link، average link و complete link. در روش single link کوتاه ترین فاصله بین جفت نقاطی که در دو خوشه قرار دارند را پیدا می کنیم. در complete link دورترین نقطه های دو خوشه را در نظر می گیریم و در روش average link میانگین فواصل بین نقاط موجود در خوشه ها را به دست می آوریم.

در این سوال ابتدا پیش پردازش روی داده ها را انجام می دهیم. سپس با تابع TfidVectorizer، مقدار tfid آن ها را به دست می آوریم.

برای یافتن دسته ها از طریق روش خوشه بندی سلسله مراتبی از تابع آماده AgglomerativeClusteringکه در scikit موجود است استفاده می‌کنیم و نوع روش یافتن شباهت را با استفاده از پارامتر linkage مشخص می کنیم و میتوانیم به آن سه مقدار complete، single و average را بدهیم. برای اندازه گیری معیارهای ارزیابی گفته شده از توابع purity_score که در کد پیاده شده و adjusted_rand_score ، normalized_mutual_info_score، f1_score موجود در کتابخانه sklearn استفاده می کنیم. برای این که خروجی های بدست آمده و لیبل های قبلی را تطبیق دهیم بررسی می کنیم که کدام دو لیبل با هم بیشتر یکسان تکرار شده اند.

در روش single link نتایج به صورت زیر ظاهر می شوند:

Purity = 0.51

RI = 0.001

F1 = 0.05

NMI = 0.005

در روش complete link نتایج به صورت زیر ظاهر می شوند:

Purity = 0.64

RI = 0.17

F1 = 0.12

NMI = 0.255

در روش average link نتایج به صورت زیر ظاهر می شوند:

Purity = 0.521

RI = 0.005

F1 = 0.067

NMI = 0.014

سوال ۳)

با تحلیل نتایج ارزیابی خوشه بندی های مختلف به نتایج زیر می رسیم:

بیشترین مقدار purity در روش k means است که برابر با ۰/۸۲ می باشد.

بیشترین مقدار RI در روش k means به دست می آید که برابر با ۰/۳۱ است.

دو معیار دیگر نیز در روش k means بیشترین هستند.

در بین روش های مختلف در الگوریتم خوشه بندی سلسله مراتبی، بهترین نتایج معیار ها مربوط به complete link هستند. نتیجه می گیریم در کل روش k means بهتر است. با توجه به این که حجم داده ها زیاد است روش سلسه مراتبی کارایی مناسبی ندارد. اما در خود روش سلسله مراتبی هم روش complete link برای اندازه گیری شباهت خوشه ها بهتر است.