به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین ۵

آذرماه ۱۴۰۲

*فهرست

بخش ۱- سوالات عملی

شرح دادگان

برای سوال ۱ مجموعه داده بازخورد مشتریان برای خطوط هوایی بریتانیایی مورد استفاده قرار می‌گیرد. و شامل متن بازخوردها، وضعیت انتشار بازخورد، نویسنده هر بازخورد و ... است. در جدول زیر ستون‌های مهم این مجموعه داده معرفی شده‌اند:

نام ستون	توضیحات
OverallRating	امتیاز کلی داده شده توسط مشتری
ReviewHeader	عنوان بازخورد مشتری
VerifiedReview	تأیید یا عدم تایید بازخورد مشتری برای انتشار
ReviewBody	متن بازخورد مشتری
SeatType	نوع صندلی

در سوال ۲ باید از گراف وب گوگل استفاده کنید. این گراف در سال ۲۰۰۲ به عنوان بخشی از رقابت برنامه نویسی گوگل منتشر شد. گراف مورد استفاده شامل ۸۷۵۷۱۳ گره است که نشان دهنده صفحات وب هستند و ۵۱۰۵۰۳۹ یال را در بردارد که هر کدام یک هایپرلینک بین صفحات وب را نشان می‌دهد.

پیش‌نیازها

به منظور پاسخگویی به سوالات این بخش حتما از زبان برنامه‌نویسی **پایتون** استفاده نمایید. پیشنهاد می‌شود از کتابخانه‌های NLTK، Scikit-learn، Networkx و سایر کتابخانه‌های آماده برای خوشه‌بندی و محاسبه معیارهای ارزیابی خوشه‌بندی محاسبه کنید.

سوال ۱-Map/Reduce

در این سوال قصد داریم با نحوه پیاده‌سازی مدل‌های مبتنی بر Map/Reduce با استفاده از کتابخانه‌ی MRJob آشنا شویم. ابتدا مستندات MRJob[1] را مطالعه کنید، سپس با پیاده‌سازی مدل‌های مبتنی بر

[1] Fundamentals — mrjob v0.7.4 documentation

Map/Reduce مناسب، به هر یک از سوالات زیر پاسخ دهید. لازم به ذکراست که انجام پیش‌پردازش‌های مورد نیاز مانند حذف علائم نگارشی، حذف Stopwordها و ... را بر روی متون بازخوردها انجام دهید.

توجه:

- برای هر سوال یک فایل با فرمت نام گذاری job_#Qnumber.py ایجاد کنید(شماره سوال را به جای #Qnumber قرار دهید).
- هر فایل باید شامل توابع Map و Reduce مناسب باشد. توضیح این توابع باید به طور کامل در گزارش موجود باشد.
- به کمک دستورات معرفی شده در مستندات کتابخانه‌ی MRJob توابع خود را بر روی مجموعه داده معرفی شده اجرا کنید و نتایج به دست آمده را تحلیل کنید.

سوالات:

أ. تعداد بازخوردهای تایید شده و تایید نشده را به دست آورید.

ب. با استفاده از مجموعه داده معرفی شده و استفاده از اندیس هر بازخورد، یک لیست شاخص معکوس[1] برای کلمات موجود در بازخوردها ایجاد کنید به شکلی که هر کلمه در بین اندیس‌ها یکتا باشد و اندیس بازخوردهای مرتبط با یک لغت به شکل مجموعه‌ای از اندیس بازخوردها باشد.

ج. ۵ کلمه‌ای که در بیشترین باخوردها ظاهر شده‌اند را به همراه تعداد تکرار ذکر کنید.

د. چه کلماتی در بیشترین نظرات هر کدام از دسته بارخوردهای تایید شده و تایید نشده ظاهر شده‌اند؟

سوال ۲- PageRank/HITS

در این سوال به آشنایی با الگوریتم‌های رتبه‌بندی اسناد تنها بر اساس لینک‌های بین آن‌ها می‌پردازیم و الگوریتم‌های متفاوت را با یکدیگر مقایسه می‌کنیم. ابتدا با استفاده از کتابخانه‌های موجود و فایل یال‌های گراف مورد نظر، یک گراف جهت‌دار ایجاد کنید و سپس به پیاده‌سازی موارد خواسته شده بپردازید:

أ. بزرگترین مولفه همبند ضعیف این گراف را بیابید و تعداد گره‌ها و یال‌های باقیمانده را گزارش کنید. این زیرگراف را به عنوان گراف مرجع برای بخش‌های بعدی سوال مورد استفاده قرار دهید.

[1] Inverted Index List

ب. الگوریتم PageRank را بر روی این گراف اجرا کنید و توزیع امتیاز PageRank را با استفاده از یک هیستوگرام نشان دهید و نمودار به دست آمده را تحلیل کنید.

ج. الگوریتم HITS را بر روی این گراف اجرا کنید و توزیع امتیازهای Hub و Authority را با استفاده از هیستوگرام نشان دهید و نمودارهای به دست آمده را تحلیل کنید.

پیشنهاد: برای افزایش تفسیرپذیری و بهبود نمایش نمودار از لگاریتم تعداد ظهور هر امتیاز استفاده کنید.

د. ۱۰۰۰ گره با بیشترین امتیاز PageRank، Hub و Authority را به دست آورید و گره‌های مشترک بین هر دو زوج از دسته گره‌های زیر را به دست آورید و نتایج را تحلیل کنید.

i. ۱۰۰۰ گره برتر معیار PageRank

ii. ۱۰۰۰ گره برتر معیار Hub

iii. ۱۰۰۰ گره برتر معیار Authority

iv. گره‌هایی بدون یال ورودی

v. گره‌هایی بدون یال خروجی

ه. الگوریتم PageRank را با مقادیر مختلف برای هایپرپارامتر α (حداقل ۵ مقدار مختلف) بر روی گراف اجرا کنید و در هر مرحله شباهت ۱۰۰۰ عدد از بهترین گره‌های این الگوریتم با ۱۰۰۰ گره برتر Hub و Authority را بررسی کنید.

بخش ۲- سوالات تئوری

سوال ۱-Clustering

در این سوال به بررسی و مقایسه الگوریتم‌های خوشه‌بندی داده‌ها می‌پردازیم. ابتدا دادگان زیر را در نظر بگیرید(هر یک از حروف A، B، C و D نمایانگر یک ترم هستند):

Document	Text
1	A B A
2	B C A B C
3	B A D D
4	A D

أ. فرض کنید یک مدل Generative مبتنی بر روش EM، خوشه مربوط به هر کدام از داده‌ها را به شکل زیر تشخیص داده است. این مدل چه مقداری را برای احتمالات هر خوشه و همچنین احتمال ظهور هر کدام از لغات در هر خوشه($P(w|C)$) به دست آورده است؟(برای حل مشکل احتمالات صفر از Laplacian Smoothing استفاده کنید)

Document	Cluster
1	$Cluster_1$
2	$Cluster_2$
3	$Cluster_1$
4	$Cluster_1$

ب. برای هر یک از اسناد فوق، یک بردار به روش TF تشکیل دهید به صورتی که این بردار به شکل زیر باشد($Count_X$نشان دهنده تعداد تکرار کلمه X است):

$$D=(Count_A, Count_B, Count_C, Count_D)$$

حال، داده‌های موجود را به روش K-means و با استفاده از مراکز خوشه اولیه زیر و معیار فاصله منهتن خوشه‌بندی کنید.

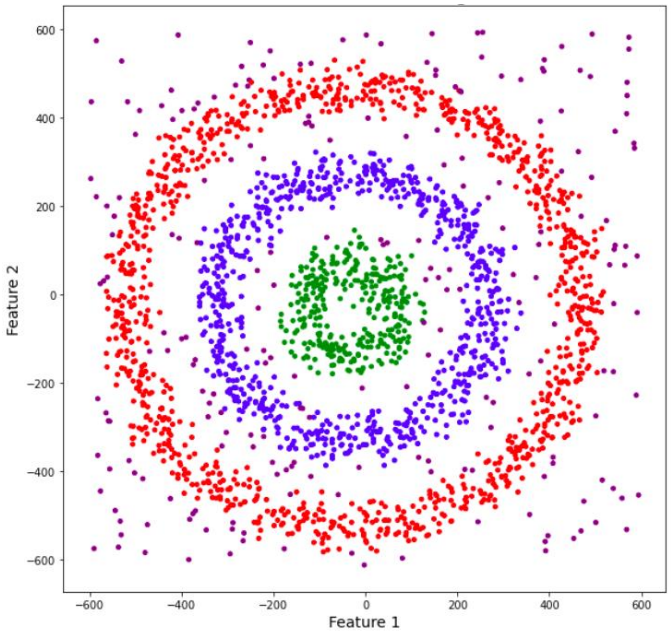
$$Centroid_1 = (1, 0, 0, 1)$$
$$Centroid_2 = (0, 2, 1, 0)$$

ج. در هر کدام از مدل‌های فوق نمونه زیر در کدام دسته قرار می‌دهند؟

$C\ A\ A$

د. به طور کلی، هر کدام از معایب و مزایای روش‌های فوق نسبت به روش دیگر را معرفی کنید.

ه. در شرایطی که داده‌ها به شکل زیر باشند، کدام یک از روش‌های فوق عملکرد بهتری از خود نشان می‌دهند؟ توضیح دهید.



ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA5_StudentID تحویل داده شود.

- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تأخیر تحویل تمرین تا یک هفته به ازای هر روز ۱۵ درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:hosein7seifi@gmail.com

مهلت تحویل بدون جریمه: ۹ دی ۱۴۰۲

مهلت تحویل با تاخیر، با جریمه ۱۵ درصد: ۱۶ دی ۱۴۰۲