Department of
Computer Engineering

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

تمرین دوم پردازش داده‌های حجیم

استاد درس: دکتر چهرقانی

بهار ۱۴۰۱

بخش اول – سوالات تشریحی

Clustering

۱ در این مسئله، قصد داریم خوشه‌بندی کلمات را در سطح کاراکترهای هر کلمه، با استفاده از معیار فاصله ویرایش[۱] به دست بیاوریم. توجه داشته باشید که تنها عمل‌های حذف و درج مجاز هستند. فرض کنید خوشه‌ی داده شده شامل کلمات he, she, hen, then و when است.

الف) اگر مرکز خوشه[۲] را کلمه‌ای تعریف کنیم که مجموع فاصله‌ی آن با سایر کلمات خوشه کمترین باشد، این کلمه را برای مثال داده شده به دست آورید.

ب) بیشترین فاصله‌ای که مرکز خوشه با سایر کلمات دارد، چقدر است؟

ج) اگر معیار انسجام[۳] یک خوشه را به صورت بیشترین فاصله‌ای که دو عضو خوشه از یک‌دیگر دارند، تعریف کنیم، معیار انسجام را برای خوشه‌ی ذکر شده حساب کنید.

DGIM

۲ کدام یک از گزینه‌های زیر حالت درستی از بازنمایی جریان داده با توجه به قوانین الگوریتم DGIM است؟

1) 1 0 1 1 1 0 1 0 1 1 1 1 0 1 0 1
2) 1 0 1 1 1 0 0 0 0 1 1 0 0 0 1 0 1 1 1 0 0 1
3) 1 1 1 1 0 0 1 1 1 0 1 0 1
4) 1 0 1 1 0 0 0 1 0 1 1 1 0 1 1 0 0 1 0 1 1

۳ فرض کنید برای تقریب تعداد بیت‌های یک جریان داده از الگوریتم DGIM استفاده می‌کنیم. اندازه‌ی پنجره را ۱۰۰۰ در نظر بگیرید.

الف) بیشترین اندازه‌ی باکت ممکن در بازنمایی این جریان داده چقدر است؟

ب) فرض کنید تمامی هزار بیت آخری که آمده‌اند، ۱ هستند. حداقل اندازه‌ی ممکن برای بزرگ‌ترین باکت در بازنمایی با پنجره‌ی گفته شده چند است؟

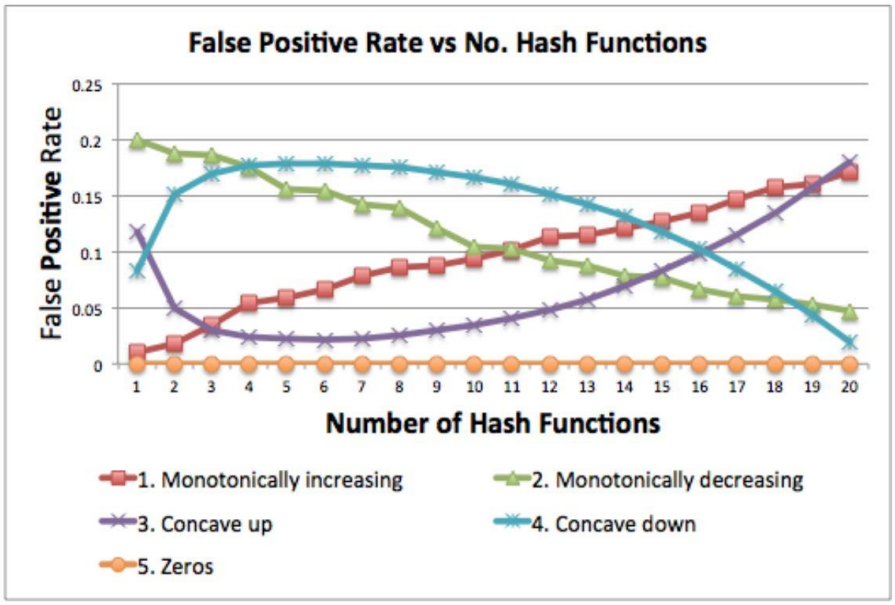
---

[۱] Edit distance
[۲] Clustroid
[۳] Cohesion

ج) یک توالی از اندازه‌ی باکت‌ها که با پاسخ شما به قسمت ب سازگار باشد، بیان کنید (ممکن است بیشتر از یک توالی را بتوان ذکر کرد، بیان یکی از آن‌ها کافی است).

Bloom Filter

۴ در این سوال پنج نمودار تابع ممکن، به صورتی که در شکل زیر مشاهده می‌کنید، داده شده است:



الف) مشخص کنید که کدام یک از نمودارها بهترین نمایش برای نرخ False positive فیلتر Bloom است. دلایل خود را ذکر کنید.

ب) مشخص کنید که کدام یک از نمودارها بهترین نمایش برای نرخ False negative فیلتر Bloom است. دلایل خود را ذکر کنید.

بخش دوم – سوالات پیاده‌سازی

DGIM

۱ DGIM یک الگوریتم کارآمد در پردازش جریان های بزرگ است. هنگامی که ذخیره جریان باینری جاری غیرممکن است، DGIM می تواند تعداد بیت‌های یک را در پنجره تخمین بزند. با این الگوریتم به صورت کامل در کلاس درس آشنا شدید. در این تمرین، در فایل stream_data_dgim.txt (جریان باینری) به شما داده شده است. الگوریتم DGIM را برای شمارش تعداد بیت‌های یک پیاده سازی کنید.

الف) اندازه‌ی پنجره را ۱۰۰۰ در نظر بگیرید و تعداد بیت‌های یک در پنجره‌ی جاری را به دست بیاورید.

ب) با همان اندازه‌ی پنجره‌ی ۱۰۰۰، تعداد بیت‌های یک در ۵۰۰ بیت آخر و ۲۰۰ بیت آخر را به دست آورید.

ج) برنامه‌ای بنویسید که به صورت دقیق تعداد یک‌های موجود در پنجره‌ی جاری را محاسبه کند. دقت و زمان اجرای الگوریتم DGIM خود را با این برنامه مقایسه کنید.

Recommendation System

۲ در این بخش می‌خواهیم با دو رویکرد مختلف collaborative filtering که در درس با آن‌ها آشنا شدید، سیستم‌های توصیه‌گر بسازیم. مجموعه‌داده‌ی مورد استفاده در این بخش، در دو فایل games.csv و ratings.csv در اختیار شما قرار گرفته است. فایل games.csv به ترتیب شامل ستون‌های: آیدی بازی، نام، تاریخ انتشار، توصیف مختصر بازی و نمره‌ی متاکریتیک آن بازی است. فایل ratings.csv نیز شامل امتیاز کاربر به بازی است، به اینصورت که در هر سطر آیدی بازی، آیدی کاربر و امتیاز داده شده آمده است.

با استفاده از روش‌های توصیه‌ی item-item collaborative filtering و user-user collaborative filtering برای کاربرهای با آیدی 5461 و 10140 از میان بازی‌هایی که به آن‌ها نمره نداده‌اند، تعداد ۵ بازی با بیشترین شباهت را پیشنهاد کنید. از معیار شباهت کسینوسی استفاده کنید.

- در خروجی نام بازی‌ها را به همراه امتیاز شباهت آن‌ها (به ترتیب نزولی امتیاز شباهت) بیاورید.
- در صورتی که دو برنامه امتیاز یکسان داشتند، برنامه با ایندکس کمتر را انتخاب کنید.

۳ در این سوال می‌خواهیم با استفاده از داده‌های موجود برای این بخش و با استفاده از اسپارک، الگوریتم تکرار شونده‌ی K-Means را پیاده سازی نماییم. فایل داده مورد نیاز برای این بخش حاوی ۴۶۰۱ سطر است، که هر سطر آن بیانگر سندی است که با استفاده از یک بردار ویژگی ۵۸ بُعدی بازنمایی شده است. فایل‌های c1 و c2 نیز به ترتیب حاوی سنترویدهای اولیه k خوشه هستند که در c1 به صورت رندوم تعیین شده و در c2 این سنترویدها تا حد ممکن، با در نظر گرفتن معیار فاصله ی اقلیدسی، از یکدیگر فاصله دارند.

برای تمام قسمت‌های این بخش، حداکثر تعداد تکرارها را برابر با ۲۰ و تعداد خوشه‌ها را برابر با ۱۰ در نظر بگیرید.

الف) با در نظر گرفتن معیار فاصله‌ی اقلیدسی، برای هر تکرار تابع هزینه را محاسبه نمایید. این عمل بدین معنی است که میبایست برای تکرار نخست از مقادیر پیش فرض یکی از فایل‌های c1 و c2 استفاده کنید. الگوریتم k-means را روی داده‌های فایل data با استفاده از مقادیر فایل‌های c1 و c2 اجرا کرده سپس مقادیر بدست آمده تابع هزینه را به صورت نمودار بر حسب تکرار از ۱ تا ۲۰ برای هر یک از دو مورد c1 و c2 رسم نمایید.

ب) درصد تغییر هزینه الگوریتم k-means بین اجرای صفرم و اجرای دهم را طبق عبارت $\frac{cost[0]-cost[10]}{cost[0]}$، با استفاده از مقادیر اولیه سنترویدها در c1 و c2، با یکدیگر مقایسه نمایید (معیار فاصله را در این سوال اقلیدسی در نظر بگیرید). توضیح دهید که کدام یک مقداردهی اولیه بهتری داشته است.

ج) مورد الف را این بار با در نظر گرفتن فاصله‌ی منهتن به عنوان تابع هزینه تکرار نمایید.

د) درصد تغییرات بین اجرای صفرم و دهم را این بار با در نظر گرفتن فاصله‌ی منهتن بدست آورده و نتایج را برای دو فایل c1 و c2 مقایسه نمایید.

الگوریتم تکرارشونده‌ی k-means برای این تمرین در زیر ارائه شده است.

Algorithm 1 Iterative k-Means Algorithm

1: **procedure** Iterative k-Means
2: Select *k* points as initial centroids of the *k* clusters.
3: **for** iterations := 1 to *MAX_ITER* **do**
4: **for** cach point *p* in the dataset **do**
5: Assign point *p* to the cluster with closest centroid
6: **end for**
7: Calculate the cost for this iteration.
8: **for** each cluster *c* **do**

نکات مربوط به تحویل تمرین

- مجموعه‌های داده و فایل‌های مرتبط با تمرین را می‌توانید از طریق سامانه درس دانلود کنید.
- **کد:** دقت داشته باشید که استفاده از کتابخانه‌های آماده برای بخش‌های خواسته‌شده، در پیاده‌سازی مجاز نیست.
- **گزارش**: ملاک اصلی انجام تمرین گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. برای این تمرین یک فایل گزارش در قالب pdf تهیه کنید و در آن برای هر سوال، تصاویر ورودی، تصاویر خروجی و توضیحات مربوط به آن را ذکر کنید. سعی کنید توضیحات کامل و جامعی تهیه کنید.
- **تذکر ۱:** مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و شدیدا برخورد خواهد شد.
- **تذکر ۲:** برای سهولت در انجام تمرینات، توصیه می‌شود که پلتفرم کولب گوگل استفاده نمایید.
- **تذکر ۳:** در نظر داشته باشید کد های شما باید قابلیت اجرا در هنگام ارائه را داشته باشند. همچنین بر روی کدهای خود مسلط باشید.
- **کانال درس:** اطلاعیه‌های مربوط به درس کانال زیر قرار می‌گیرند:
- https://t.me/+cLCmyX2sIPVjN2I0
- **راهنمایی:** در صورت نیاز میتوانید سوالات خود را در خصوص تمرینات، از طریق ایمیل زیر بپرسید.
- E-mail: bigdata.aut.1401@gmail.com

- **ارسال:** پاسخ سوالات تشریحی، فایل های کد و گزارش خود را در یک فایل فشرده قرار داده و با نام با فرمت HW2_StudentID ارسال نمایید.