کاوش داده‌گان انبوه

دکتر سامان هراتی زاده
دانشگاه تهران - دانشکده علوم و فنون
ترم اول سال تحصیلی – پاییز ۱۳۹۹

دستیاران آموزشی
مهدی صادقی (mahdisadeghi@ut.ac.ir)
شیوا پارسا راد (shiva.parsarad@ut.ac.ir)
ایمان برازنده (barazandeh.iman@ut.ac.ir)

مجموعه تمرین دوم : LSH، ابزار همه فن حریف!

مهلت تحویل این تمرین جمعه ۲۱ آذر ساعت ۲۳:۵۹ می‌باشد.

۱-(پژوهش:GPU vs. CPU یا NVIDIA vs. Intel) دو ویدوی زیر را تماشا کنید. ویدیوی اول یک ارائه از آقای شِریواستاوا[1]، استادیار دانشگاه RICE و ویدیوی دوم، مصاحبه‌ای با دانشجوی دکتری سابق ایشان و محقق فعلی دوره پسادکتری در دانشگاه استَنفورد، خانم چِن[2]، درباره مقاله آنها در کنفرانس MLSys 2020 است. در هر دو ویدیو درباره این مقاله صحبت شده است:

Statistical Estimations from Locality Sensitive Hashing

SLIDE: Smart Algorithms over Hardware Acceleration for Large-Scale Deep Learning

بدون اینکه درگیر جزئیات شوید، کار انجام شده در این مقاله را با تمرکز روی کاربرد LSH و دیگر توابع هَش در آن شرح دهید.

۲-تمرین ۳.۳.۲ . و ۳.۳.۳ از کتاب MMDS را حل کنید.

[1] Anshumali Shrivastava
[2] Beidi Chen

۳-(طراحی و پیاده سازی الگوریتم)

هدف از این تمرین پیاده سازی یک الگوریتم برای پیدا کردن داکیومنت‌های مشابه در دیتاستِی است که در اختیار شما قرار گرفته است. این الگوریتم باید شامل قدم‌های زیر باشد:

- همه داکیومنت‌ها به شینگل[۳] تبدیل شوند. علائم نگارشی به فاصله تبدیل شده و فاصله‌های متوالی تبدیل به یک فاصله شوند.
- ماتریس شینگل-داکیومنت را بسازید.
- با توجه به بخش ۳.۳.۵ کتاب، ماتریس MinHash Signature را بسازید.
- از توابع LSH برای یافتن داکیومنت‌های مشابه استفاده کنید.
- کاندیداهایی که شباهتشان یک آستانه بیشتر است را در خروجی چاپ کنید.

گزارش شما باید شامل موارد زیر نیز باشد:

- الگوریتم خود را با الگوریتم Brute Force برای یافتن داکیومنت‌های مشابه مقایسه کنید.
- می‌دانیم که معمولا طول شینگل‌ها را ۹ در نظر می‌گیرند. اگر از شینگل‌های بزرگتر یا کوچکتر استفاده کنیم چه تغییری در نتیجه ایجاد می‌شود.
- نتایج اجرای الگوریتم خود را با تعداد متفاوت توابع هَش (یا همان طول Signature) برای ساخت ماتریس MinHash Signature گزارش کنید.

[3] Shingle

نکات تکمیلی

- از زبان برنامه سازی پایتون کنید.
- کدها با نامگذاری مناسب در پوشه SOURCES قرار گیرد.
- گزارش کامل تمرین را با ساختار مناسب، بدون ابهام و به زبان ساده نوشته و با فرمت PDF در پوشه دیگری قرار دهید. برای نگارش گزارش‌ها نیز امتیاز مثبت و منفی در نظر گرفته شده است، پس برای نگارش گزارش خود وقت مناسبی اختصاص دهید.
- پاسخ تمرین خود را به ایمیل barazandeh.iman@ut.ac.ir ارسال نمایید. هر دو پوشه را در یک فایل ZIP با نام MMDS_STID_FullName_HW# مثلا MMDS_830498001_ImanBarazandeh_HW1 قرار داده و در ایمیلی با همین عنوان نیز ارسال کنید.
- در صورت مشخص شدن هر نوع تقلب انفرادی یا مشترک نمره تمرین صفر در نظر گرفته می‌شود.
- به ازای هر روز تاخیر ۲۵٪ از امتیاز تمرین از دست خواهید داد.
- هر گونه ابهام را با دستیاران آموزشی مطرح کنید.