

تمرین ۳- رتبه و چندک

۱- مقاله مونرو و پاترسون را مطالعه کنید و پرسش‌های زیر را پاسخ دهید. الف- مسئله اصلی مطرح شده در مقاله چیست؟ ب- چرا محاسبه دقیق چندک در یک گذر از داده‌ها با حافظه زیرخطی امکان‌پذیر نیست؟ ج- ایده کلی اثبات ارائه شده در مقاله چیست؟ (نیازی به بیان کامل اثبات نیست). د- این نتیجه چه تأثیری بر توسعه الگوریتم‌های تقریبی جریان داده داشته است؟

۲- مقدمات ادغام‌پذیری تلخیص در درس تعریف شد. اساس کارهای مربوط به ادغام‌پذیری کار آگاروال و همکاران است. مقاله مرتبط را مطالعه و نقش قابلیت ادغام در پردازش توزیع شده جریان داده را توضیح دهید و بیان کنید که این ویژگی چگونه در الگوریتم‌های تلخیص چ و تلخیص د مورد استفاده قرار گرفته است.

Agarwal, Pankaj K., et al. "Mergeable summaries." *ACM Transactions on Database Systems (TODS)* 38.4 (2013): 1-28.

۳- نقش پارامتر فشرده‌سازی را در تلخیص چ و تلخیص د بررسی کنید و اثر آن را بر موارد حافظه، سرعت، دقت، خطای نظری، قابلیت ادغام، و مناسب بودن برای چندک‌های انتهائی مانند ۹۵ و ۹۹ تحلیل کنید.

۴- قانون زیف Zipf's Law را مطالعه کنید و ارتباط آن را با بسامد و رتبه بررسی کنید.

۴- مراحل زیر را انجام دهید. الف- متن شاهنامه را توکن‌سازی کنید. ب- کلمات متمایز را استخراج و به ترتیب الفبای فارسی مرتب نمایید. ج- به هر کلمه یک شناسه عددی اختصاص دهید. د- متن را به جریان شناسه‌های عددی تبدیل کنید. ه- توزیع فراوانی شناسه‌ها را رسم و تحلیل کنید. منطبق بودن توزیع را بر قانون زیف بررسی کنید. و- الگوریتم‌های م، ر، ل، تلخیص چ، تلخیص د، و طرحواره شمارش-کمینه را پیاده‌سازی نمایید. ز- با استفاده از هر الگوریتم چندک‌های ۰.۵، ۰.۹، ۰.۹۵ و ۰.۹۹ را محاسبه کنید. رتبه حداقل پنج مقدار انتخابی را تخمین بزنید. ح- نتایج را با پاسخ دقیق مقایسه کنید. ط- حافظه، زمان اجرا و خطای هر روش را گزارش نمایید. ی- نتایج عملی را با کران‌های نظری ارائه شده در درس مقایسه و تحلیل کنید.

نمره اضافه- آزمایش‌های بخش (۴) را روی مجموعه داده NYC Taxi Trip Records (January 2009) و ستون trip_distance نیز انجام دهید و نتایج را با داده شاهنامه مقایسه کنید.

تمرین‌ها در گروه‌های یک یا دو نفره ممکن است. مراجع در صورت استفاده باید دقیق باشد و صرفاً به مدل‌های زبانی تکیه نشود.