دانشکده مهندسی برق

تحلیل داده‌های حجیم

پروژه پایانی

استاد: دکتر ایمان غلامپور

طراح پروژه: حسین شریفی، بهنام رئوفی

توضیحاتی درباره نحوه انجام پروژه درس:

- کدنویسی باید همراه با comment گذاری و با توضیحات کامل در گزارش باشد.
- گزارش نهایی در قالب فایل WORD و PDF ارائه گردد.
- استفاده از روش‌های داده‌کاوی الزامی است.
- برای بهبود نتایج، از تکنیک‌های یادگیری ماشین و بهینه‌سازی بهره ببرید.
- ارزیابی براساس خلاقیت، تحلیل عمیق، نمایش مناسب نتایج در گزارش و کدنویسی روشن و اصولی انجام می‌شود.
- ارائه ایده‌های نوآورانه که در لیست پیشنهادات ذکر نشده‌اند، مورد تشویق قرار می‌گیرد.
- امکان اضافه کردن ایده‌های جدید برای هر بخش وجود دارد.
- سعی کنید از Visualizationهای مناسب برای نمایش نتایج استفاده کنید.
- گزارش باید شامل نتایج کامل و خروجی‌های مرتبط باشد.
- خروجی‌ها به صورت فایل Jupyter Notebook قابل اجرا باشند و خروجی HTML اجباری می‌باشد.
- در صورت نیاز، از دانشجویان خواسته خواهد شد که ارائه‌ای حضوری یا مجازی درباره پروژه خود داشته باشند. این ارائه به‌منظور بررسی دقیق‌تر ایده‌ها، نحوه اجرا و پاسخ به سوالات احتمالی انجام می‌شود و نمره‌ی کل پروژه به تسلط بر این ارائه وابسته خواهد بود. زمان و نحوه ارائه متعاقباً اعلام خواهد شد.
- در صورتی که شباهتی میان پروژه‌ها مشاهده شود، نمره نهایی هر دو طرف صفر در نظر گرفته خواهد شد.

- در این پروژه، دانشجویان موظف‌اند موارد ۱، ۴ و ۷ را به‌صورت اجباری انجام دهند، درحالی‌که سایر موارد اختیاری هستند. هر دانشجو می‌تواند علاوه بر موارد الزامی، ایده‌های دلخواه را نیز اجرا کند تا پروژه‌ای کامل‌تر ارائه دهد. هدف این ساختار، ایجاد انعطاف‌پذیری در انتخاب و تشویق دانشجویان به خلاقیت است.

- جهت آشنایی با دیتاست در ابتدا فایل dataset.info مطالعه و براساس لینک ارائه شده دیتاست را دانلود کنید.

ایده‌های پیشنهادی:

۱. با استفاده از الگوریتم Pixie، گرافی از کاربران و تعاملات آن‌ها (مانند لایک، دیس‌لایک، ریتوییت و ...) بسازید. برای هر لینک در گراف، براساس فعالیت‌های کاربران، یک روش امتیازدهی (Scoring) طراحی کنید و الگوریتم Pixie را برای یافتن کاربران مشابه پیاده‌سازی کنید. در گزارش، روش امتیازدهی لینک‌ها و نحوه طراحی گراف را به‌طور کامل توضیح دهید.

۲. با توجه به امتیازدهی لینک‌ها که در قسمت قبل طراحی کرده‌اید، الگوریتمی ترکیبی با استفاده از روش‌های Content-based و Collaborative filtering پیاده‌سازی کنید تا کاربران با تعاملات مشابه را شناسایی کنید. نتایج این الگوریتم را با نتایج الگوریتم Pixie مقایسه کنید. مزایا و محدودیت‌های هر روش را به‌طور کامل توضیح دهید. با توجه به نتایج قسمت قبل معیار جاکارد برای کاربران محاسبه کنید و نتیجه را گزارش کنید.(برای بهبود نتیجه در این قسمت می‌توانید از ترکیب الگوریتم‌های Recommender System و توابع فاصله متنوع استفاده کنید.)

۳. با استفاده از Topic-Specific PageRank، الگوریتمی طراحی کنید که بر اساس یک کاربر ورودی و هشتگ‌های مشخص، توییت‌های با هشتگ مشابه به کاربر پیشنهاد دهد. نحوه ساخت گراف و تنظیم پارامترهای الگوریتم را توضیح دهید و نتایج را گزارش کنید.

۴. با استفاده از الگوریتم TrustRank و محاسبه Spam Mass، الگوریتمی طراحی و پیاده‌سازی کنید که بتواند توییت‌های اسپم را شناسایی کند. فرآیند انتخاب Seed Nodes و نحوه محاسبه Spam Mass را توضیح دهید و نتایج را گزارش کنید.

۵. با استفاده از HITS، الگوریتمی طراحی و پیاده‌سازی کنید که بتواند توییت‌ها و کاربران تأثیرگذار (Hubs and Authorities) را در این دیتاست شناسایی کند. نتایج را تحلیل کرده و ارتباط میان Hubs و Authorities را توضیح دهید.

۶. با استفاده از الگوریتم SVD، فاکتورهای پنهان ماتریس تعاملات کاربران و توییت‌ها را تحلیل و تفسیر کنید و این فاکتورها را با ویژگی‌های توییت‌ها مقایسه کنید.

۷. با استفاده از PySpark Structured Streaming ، توییت‌های جدید را به‌طور آنی پردازش کرده و هشتگ‌های هر توییت را شناسایی و شمارش کنید. همچنین، از ویژگی sentiment هر توییت برای تحلیل احساسات استفاده کرده و میانگین sentiment مربوط به هر هشتگ را به‌صورت Real-Time محاسبه و نمایش دهید. نحوه پردازش توییت‌ها و محاسبه میانگین‌های sentiment را توضیح داده و نتایج به‌روز شده را گزارش کنید.(می‌توانید اگر Sentiment، Negative باشد مقدار ۰، Positive باشد مقدار ۱ و اگر Neutral باشد مقدار ۰,۵ در نظر بگیرید)