سوالات تمرین شماره یک BIG DATA

تاریخ تحویل : 1401/9/2 نمره : 1.5

1. فایل JaneAusten.txt شامل کلیه نوشته های جین آستین، نویسنده انگلیس، می باشد که توسط پروژه گوتنبرگ جمع آوری شده است. با استفاده از SPARK برنامه ای بنویسید که،

- الف: تعداد کل کلمات این متن را مشخص نمایید.
- ب: تعداد کلمات بدون تکرار چقدر است؟
- ج: ده کلمه ای که بیشترین تعداد تکرار را دارند کدامند و هر کدام چند بار تکرار شده اند؟ همچنین با تغییر تعداد هسته هایی که برنامه (ب) بر روی انها اجرا می شود از یک هسته تا حداقل چهار هسته نمودار زمان اجرای ان را رسم نمایید. خروجی گزارش شامل کد، نتایج، نمودار و تحلیل ان می باشد.

2. سه فایل C1,C2,C3 هر کدام حاوی اطلاعات دو ستونی متنی می باشند که هر سطر ان مختصات یک نقطه دو بعدی را نشان می دهد. با استفاده از SPARK دو برنامه بنویسید که با استفاده از دو روش خوشه بندی K-means++ وBisecting K-means ،

- الف : برای هر فایل داده نمودار هزینه کلاستر را برای k=2 to25 رسم نمایید.
- ب: تعداد بهینه خوشه ها چقدر است؟

- ج: در تعداد بهینه خوشه و در روش خوشه بندی با هزینه کمتر نقاط مرکزی هر کلاستر را برای هر یک از فایل های داده مشخص نمایید؟
- د: درتعداد بهینه خوشه نمودار زمان اجرای دو روش خوشه بندی را بر روی یک تا 4 هسته برای هر یک از فایل ها رسم نموده و تحلیل نمایید. گزارش شامل کد و نتایج و نمودارهای هر مرحله و تحلیل جواب ها می باشد.