دانشکده علوم مهندسی

استاد:
دکتر علی فهیم

تمرین ششم

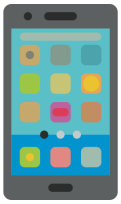
گزارش کار

تاریخ:
1399/10/27

دیتا کاوی - پاییز ۱۳۹۹

نام و نام خانوادگی:
مینو احمدی

مقدمه



در این تمرین ابتدا به محاسبه زیر رشته های یک دسته از رشته ها می پردازیم. سپس closed frequent substring را برای ۲۰ رشته اول دیتابیس محاسبه می کنیم.

نحوه اجرای برنامه

python HW6.py sequencedb.txt seqin.txt

خروجی برنامه

> for each substring in sequin.txt:

‘substring - support’

‘closed_frequent_substring:’

> for each substring in closed_frequent_substring:

‘substring - support

قسمت اول: ساختن درخت Suffix

برای ساختن درخت از ساختار Node استفاده شده. هر Node یک نام و یک اشاره گر به پدر و لیستی از اشاره گر ها به فرزندانش و یک ست از فریکونت دارد. روی کل دیتابیس حرکت کرده و به ازای هر رشته تمام Suffix های آن را به تابع add_child مربوط به Node ریشه می دهیم. تابع ذکر شده در صورت ابتدا پدر این زیر رشته را پیدا می کند و سپس آن را به بچه های آن اضافه می کند(اگر نیاز بود Node را می شکند و Node جدید با بچه قدیم و زیر رشته جدید می سازد.) در انتها شماره رشته به کلیه Node ها از روت تا خود بچه ایجاد شده اضافه می شود.

قسمت دوم: پیدا کردن Closed Frequent Substring

برای این قسمت کافیست یک بار روی درخت Suffix حرکت کنیم و هر جا زیر رشته جدیدی دیدیم آن را به Closed اضافه کنیم(با تعاریف مربوطه).

دو نوع تعریف برای این قسمت مشاهده شد. اولی در اینترنت که نتیجه اش برای شما ارسال شد و نادرست بود(در این حالت فقط رشته هایی که ساپورت یکسانی با رشته مد نظر داشتند حذف می شدند) و دومی که کد بر این اساس است و فکر می کنم مد نظر شما بود(در این حالت رشته هایی که ساپورت کوچکتر یا مساوی با رشته مد نظر داشتند حذف می شدند).