بسمه تعالی

Data cleaning

نحوه تمیز کردن داده :

فایل preprocessor.sh شامل دستورات tr زیر است
برای مشاهدا بهتر مراجعه شود به فایل

```
tr 'ی' 'ئ' < ../data/raw_celebrity_comments | tr '' ' ' | tr -sc "A-Za-zضصثقفغعهخحجچشسیبلاتنمکگظطزرذدپوائ'.,-" " " | tr -s [:space:] " " | tr [:upper:] [:lower:] | tr -s "A-Za-zضصثقفغعهخحجچشسیبلاتنمکگظطزرذدپو'.,-" > ../data/celebrity_comments
```

```
tr 'ی' 'ئ' < ../data/raw_friendly_comments | tr '' ' ' | tr -sc "A-Za-zضصثقفغعهخحجچشسیبلاتنمکگظطزرذدپو'.,-" "" | tr -s [:space:] " " | tr [:upper:] [:lower:] | tr -s "A-Za-zضصثقفغعهخحجچشسیبلاتنمکگظطزرذدپو'.,-" > ../data/friendly_comments
```

روش کار :
۱ : ی عربی را با فارسی جایگزین میکنیم
۲: نیم فاصله را با فاصله جایگزین میکنیم
۳: همه کاراکتر ها به غیر از فارسی و ابگلیسی را حذف میکنیم.
۴: فاصله های زیاد را با یکی جایگزین میکنیم (لزومی ندارد)
۵:حروف بزرگ انگلیسی را با کوچک جایگزین میکنیم .
۶:حروف تکراری را با یک حرف جایگزین میکنیم (بعضی کلمات انگلیسی شامل ۲ حرف پشت هم با معنی است اما به دلیل اینکه حالتی که هم خودشان و هم تک حرفشان معنی بدهد کم هستند از این حالت صرف نظر میکنیم و تکرار کم آنها انها را با نسخه یک حرفشان یکی میگیریم) مانند good , god

۷ : داده تمیز شده را در خروجی میریزیم
