عنوان پروژه: استخراج دانش از CodeForces.com

علی اسمعیلی سراجی – 94222102

فایل­های داده:

* فایل maindb.csv: این فایل شامل فیلدهای username داخل سایت که یکتاست(handler)، زمان ثبت نام کاربر به سال(registrationTime)، کشور(country)، موسسه ای که کاربر در آن مشغول به کار است(organ)، زبان برنامه­نویسی(language)، امتیاز درون سایت(rating)، تعداد سوال­های اکسپت شده(Accepted)، و تعداد سابمیت­های اشتباه است(Wrong).
* فایل data.json: فایل json، شامل اطلاعات عمومی کاربرانی که تا ماه پیش حداقل یک تراکنش درون سایت انجام داده اند. از جمله نام و نام خانوادگی، امتیاز درون سایت، شهر، کشور و غیره...
* فایل users.csv: فایلی شامل تنها username کاربرهای درون سایت
* فایل readed.csv: درون این فایل تمامی یوزر هایی که اطلاعات آن­ها به طور کامل دانلود شده است، نگه داری می­شود.
* فایل contest.list.json: فایلی که به طور مستقیم از آدرس

/api/contest.list

گرفته شده است. این فایل json ای شامل مشخصات تمام کانتست­های برگزار شده تا به حال است. از جمله طول آزمون، زمان شروع آزمون، درجه دشواری آزمون و غیره.

* فایل diff\_matrix.csv: هر سطر این فایل، شامل سه ستون اندیس کاربر اول در فایل users.csv، اندیس کاربر دوم، و میزان اختلاف دو کاربر است.
* فایل db.csv: این فایل شامل نام کاربری، حداکثر امتیاز کاربر تا به کنون، امتیاز کنونی کاربر، کشور، موسسه و ایمیل و تعداد friends درون سایت است.
* فولدر user\_submition: درون این فولدر به ازای هر یوزر، فایلی از نوع json وجود دارد که به طور مستقیم از آدرس

/api/user.status?handle=<username>

دانلود شده.

* فولدر user\_rating: درون این فولدر به ازای هر یوزر، شماره تمامی کانتست هایی که در آن شرکت کرده، و میزان تغییرات امتیاز در هر کانتست، به صورت فایل json ذخیره شده. این فایل­ها به طور مستقیم از آدرس

/api/user.rating?handle=<username>

گرفته شده.

* فولدر user\_info: درون این فولدر به ازای هر یوزر، اطلاعاتی همچون email، آواتار، امتیاز کنونی، کشور، شهر، موسسه، نام و نام خانوادگی ذخیره شده است.

این فایل­ها به طور مستقیم از آدرس

/api/user.info?handles=<username>

گرفته شده.

* فولدر rated\_submition: درون این فولدر به ازای هر یوزر، فایلی از نوع csv وجود دارد که به ازای هر کانتستی که کاربر درون آن شرکت کرده، امتیاز قبل از کانتست و بعد از کانتست، و البته تعداد سوالات حل شده تا به قبل از آن کانست در سطوح A، B، C، D، و E و بالاتر، آمده است.

\*\*) داده­هایی از جمله ایمیل و کشور و موسسه که به صورت اختیاری توسط کاربر درون سایت پر می­شود، در صورت ناموجود بودن، با دش (- ) پر شده اند.

فایل­های اجرایی و script ها:

* فایل csv\_maker.py: که کار اصلی آن، تبدیل data.json، به users.csv است. به طوری که username ها، به تریتیب نزولی درون users.csv قرار می­گیرند.

این فایل، db.csv را هم که شامل اطلاعات عمومی کاربران، مانند email و کشور بود، راه هم به صورت csv و ترتیب نزولی می­سازد.

کد توضیح خاصی ندارد و تنها از توابع کتابخانه­ای json درون python، برای بازنگاری آن استفاده می­کند.

* فایل user\_download.go: این فایل، برنامه­ای به زبان GoLang است که کار اصلی آن خواندن اسامی username از فایل users.csv و دانلود و جاگذاری در فایل مربوطه در پوشه­های user\_rating، user\_submition و user\_info است.

زبان Go به دلیل، امکانات راحت و سریع جهت موازی کردن برنامه، جهت دانلود سریع تر اطلاعات از سمت سرور انتخاب شده. با مقداردهی متغیر MACHINE\_COUNT درون Terminal، می­توان تعداد ماشین­های موازی­ای که به طور همزمان از سایت دانلود می­کنند را تعیین کرد.

از این جهت که فایل­های درون پوشه user\_submition، تمامی سابمیت­های یوزر، و اندازه هرکدام، در حدود 1.5 مگابایت است، اطلاعات تمامی یوزرها را نمی­توان در یک نوبت دانلود کرد. به همین دلیل، اسامی یوزرهایی را که تا کنون به طور کامل دانلود کردیم را درون فایل readed.csv قرار می­دهیم که دانلود، در دفه­های بعدی از ادامه شروع شود.

* فایل main\_flatter.py: این فایل اسکریپتی به زبان پایتون است، که کار اصلی آن ساختن فایل maindb.csv است. registerTime، country، organ، rating همگی از فایل مربوطه درون پوشه user\_info، و Accepted، Wrong و language از خواندن تمامی submition ها از فولدر user\_submition استخراج می­شود. زبان برنامه نویسی مورد استفاده، از روی بیشترین کامپایلر استفاده شده استخراج می­شود.
* فایل contest\_flatter.py: اسکریپتی به زبان پایتون، که کار اصلی آن ساختن فایل­های داده درون rated\_submition است.

تغییرات امتیاز قبل و بعد از هر کانتست، از طریق فایل مربوطه درون user\_rating، تعداد سوال­های حل شده تا به قبل از کانتست هم از طریق بررسی از فولدر user\_submition بدست می­آید. زمان کانتست هم که برای بدست آوردن سوال­هایی که تا قبل از آن حل شده استفاده می­شود، از فایل contest.list.json استخراج می­شود.

* فایل run.sh: اسکریپتی به زبان shell است که کار اصلی آن، فراخوانی به ترتیب اسکریپت­های دیگر است.
* فایل diff\_matrix: برای محاسبه diff\_matrix.csv استفاده می­شود.

تعداد Accept ها و Wrong ها در محاسبه اختلاف، به روش کسینوسی است.

می­توان با بررسی همبستگی بین تعداد سوالات حل شده در سطوح مختلف، مقدار سابقه در کد زدن و از این دست موارد با میزان پیشرفت و یا پسرفت، میزان تاثیر آن­ها را بررسی کرد.

برای اضافه کردن داده­های جانبی هم، می­توان از email و api هایی که سایت هایی مثل linkedIn و github ارائه می­کنند استفاده کرد و سابقه برنامه­نویسی یوزر را استخراج کرد.