دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس

تحلیل داده و مصورسازی

دکتر محمدامین صادقی

طراح تمرین: محمد نیلی

زمان بارگزاری تمرین: ۸ مهر ماه

تمرین شماره ۱

کار با انواع فایل و داده

نیمسال اول سال تحصیلی ۱۴۰۰ - ۱۴۰۱

مقدمه

پیش از شـــروع به انجام تمرین، کتابچه هندزآن را مطالعه و Taskهای آن را در داخل آن انجام دهید.

بخش اول – خواندن انواع فایل

در بخش قبل به راه اندازی و نصـــب پکیج های مختلف پرداختیم. در این قســـمت قصـــد داریم تا با خواندن انواع فایل در بخش های جداگانه آشنا بشویم.

فایل txt

از سایت Gutenberg [1]به دلخواه یک کتاب را انتخاب کرده و مراحل زیر را انجام دهید .

(برای شـــروع میتوانید کتاب های با حجم کم را دانلود کرده و یا کتاب مورد نظر را خودتان خلاصـــه کنید، بدین صورت که حداقل ۵۰ سطر از کتاب را انتخاب کنید و سایر بخش ها را دستی پاک کنید.)

- تنها ۵ خط اول کتاب را چاپ کنید.
- کلمات هر خط را جدا کرده و در یک متغیر ذخیره کنید.

فایل csv

داده های مربوط به واکســـن کرونا به تفکیک هر کشـــور (با نام country_vaccination.csv ضـــمیمه شده است) را بخوانید و مراحل زیر را انجام دهید .

- تمامی داده ها را خوانده و داده ها را تمیز کنید (حذف داده های Nan و تکراری)
- تعداد داده های حذف شده و تکراری را به تفکیک سطر و ستون گزارش کنید.

برای این بخش ۴ خروجی لازم اسـت گزارش شـود (داده های حذف شـده یا تکراری به ازای سـطر یا ستون)

[1] https://www.gutenberg.org/

فایل log

از لینک پیوست [1]شده داده های مربوط به log پیدایش لینوکس را دانلود کرده و به سوالات زیر پاسخ دهید دقت شود که در پاسخ به سوالات تنها مجاز به استفاده از فایل با پسوند log هستیم. (در صورت مشکل در دانلود، فایل لازم ضمیمه شده است)

- با توجه به نوع و حجم داده ها ، به بهترین روش داده ها را بخوانید.
- لیست author های منحصر به فرد را استخراج کنید.

بخش دوم – گزارش آماره ها

با استفاده از فایل های خوانده شده در بخش قبل، آماره های خواسته شده را گزارش کنید.

فایل txt

- تعداد کل کلمات را گزارش کنید.
- در صورت امکان تعداد کلمات منحصر به فرد را گزارش کنید.

فایل csv

- آماره های کلی تمامی ستون ها را استخراج کنید.
- ستون جدید (first_vac) به عنوان اولین واکسن را از روی ستون vaccines استخراج کنید.
- کدام کشورها بیشترین و کمترین میزان تزریق واکسن را داشتند ؟
- کدام کشور ها تعداد روز های بیشتری بدون تزریق واکسن گذراندند ؟
- کدام کشور ها بالای ۲۰ ملیون دوز به ازای یک روز را تجربه کردند؟
- نرخ اولین واکسن را به تفکیک هر کشور گزارش کنید.

(به کمک ستون جدید FIRST_VAC این کار را انجام دهید)

فایل log

- نرخ و تعداد مشارکت به ازای هر author را گزارش کنید.
- فعال ترین author ها (۵ فرد برتر) را مشخص کنید

[1] https://www.kaggle.com/arpitdw/exploring-the-evolution-of-linux/data

- به دلخواه روی فعالیت author ها یک نمودار رسم کنید. نتایج و بینشی که از آن بدست آوردید و دلیل انتخاب نمودار خود را به اختصار شرح دهید.

بخش سوم – دستکاری انواع داده ها (امتیازی*)

در این بخش قصد داریم که کمی بیشتر با انواع داده ها کار کنیم.

فایل csv

داده های میزان ابتلا به کرونا و مرگ و میر ناشی از کرونا را به ازای یک کشور دلخواه (داده های مربوط به سال ۲۰۲۱) دانلود کنید و به سوالات زیر پاسخ دهید.

- داده های دو جدول واکسن و ابتلا به کرونا را با توجه به تاریخ ادغام کنید.
- به ازای یک کشور دلخواه ، نمودار تزریق / ابتلا به کرونا را رسم کنید.

(محور افقی تاریخ و محور عمودی تعداد ، توجه شود نیاز به مقایسه ی نرخ ابتلا و تزریق داریم.)

نکات پیاده سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز جمعه ۲۳ مهر ماه خواهد بود.
- انجام این تمرین به صورت یک نفره می باشد.
- خروجی مورد انتظار تمرین و هندزآن ، ۲ فایل اجرا شده جداگانه کتابچه جوپیتر با یکی از فرمتهای HTML(.html) و یا Notebook(.ipynb) می‌باشد.
- هرگونه توضیحات و گزارش نویسی را به صورت Markdown در داخل کتابچه جوپیتر انجام دهید.
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمائید.

HW1_[Lastname]_[StudentNumber].zip

- در صورت وجود سوال و یا ابهام میتوانید از طریق رایانامه زیر با دستیار آموزشی در ارتباط باشید:

mohammad.nili@ut.ac.ir