دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

درس

تحلیل داده و مصورسازی

دکتر محمدامین صادقی – دکتر محمدرضا ابوالقاسمی

طراح تمرین: امیرحسین مصباح

زمان بارگزاری تمرین: شنبه ۹ مهر ماه ۱۴۰۱

تمرین شماره ۱

موضوع تمرین: آشنایی با کتابخانه‌های numpy، pandas و matplotlib

نیمسال اول سال تحصیلی ۱۴۰۱ - ۱۴۰۲

بخش اول – آشنایی با کتابخانه numpy

با توجه به اینکه در تمرین‌های آینده شما نیاز به مهارت کار با کتابخانه‌های مورد نیاز برای تحلیل داده در پایتون (مانند numpy، pandas، matplotlib و ...) را دارید، اهدف این تمرین آشنای با این کتابخانه‌ها و مرور متدهای پرکاربرد این کتابخانه‌ها می‌باشد. طبیعتا با در نظر گرفتن تعداد متدهای بسیار زیاد این کتابخانه‌ها ما در یک تمرین قادر به پوشش همه بخش‌ها نیستیم و کسب مهارت در این زمینه نیازمند تمرین بیشتر خود شما نیز است.

با توجه به این امر، در این بخش به بررسی و مرور کتابخانه پرکاربرد numpy می‌پردازیم. به سوالات زیر در نوتبوک آماده شده پاسخ دهید. لازم به ذکر است که برای کسب نمره کامل در این بخش نباید از حلقه‌های for، while و ... استفاده کنید. شما تنها مجاز به استفاده از کتابخانه numpy می‌باشید.

سوال اول:

میانگین و انحراف معیار اعداد ۱۰ تا ۱۰۰۰ را محاسبه کرده و چاپ کنید.

سوال دوم:

۱۰۰ نقطه رندوم ۱۰ بعدی ایجاد کنید. برای هر نقطه نزدیک‌ترین نقطه موجود از بین این نقاط دیگر را با در نظر گرفتن فاصله اقلیدسی پیدا کرده و اندیس نقطه مربوطه را چاپ کنید.

سوال سوم:

فاصله هر نقطه از سایر نقاط را حساب کرده و نمودار هیستوگرام فاصله‌ها را بکشید. لازم به ذکر است که نمودار شما باید دارای عنوان و لیبل برای محورهای افقی و عمودی باشد.

سوال چهارم:

با توجه به دیتای load شده در cell موجود در نوتبوک ارائه شده، به سوالات زیر جواب دهید:

- بخش اول: کدام patent دارای بیشترین norm است. (فاصله اقلیدسی از مبدا)
- بخش دوم: دو patentای را پیدا کنید که بیشترین فاصله را از هم دارند.
- بخش سوم: تابعی بنویسید که با گرفتن شماره patent به عنوان ورودی، نزدیک ترین همسایه آن را پیدا کرده و چاپ کند.

- بخش چهارم: تعداد patentهایی که با نزدیک‌ترین همسایه خود در یک دسته یا category هستند را به دست آورید.

- بخش پنجم: patentهای هر دسته را در نظر بگیرید. برای هر دسته فاصله pairwise را برای patentها حساب کنید. این فاصله‌ها را برای هر دسته در یک آرایه یا لیست ذخیر کنید، سپس میانگین و انحراف معیار فاصله‌های هر دسته را به دست آورید. با توجه به این معیار میتوان در مورد تراکم هر دسته تخمینی داشت؟ کدام دسته متراکم تر و کدام دسته پراکنده‌تر از بقیه دسته‌ها هستند؟

بخش دوم – کار با انواع فایل‌ها

هدف این بخش این است که شما تا حد امکان با انواع فرمت‌های موجود داده‌های مورد نیاز برای تحلیل داده آشنا شوید. در هر بخش موارد خواسته شده را در نوتبوک تهیه شده انجام دهید و به سوالات مورد نظر پاسخ دهید. در این بخش مجاز به استفاده از کتابخانه‌های موجود هستید.

فایل text

ابتدا فایل zen_of_python.txt را توسط کد پایتون خوانده و سپس به سوالات زیر جواب دهید:

۱- لغات موجود در این فایل را در یک لیست ذخیر کنید. لازم به ذکر است که در این لیست نباید لغت تکراری وجود داشته باشد.

۲- برای هر حرف یا به اصطلاح character تعداد تکرار آن را محاسبه کرده و چاپ کنید. همچنین این کار را برای تعداد تکرار هر لغت نیز انجام دهید.

۳- تابعی بنویسی که به ازای هر فایل ورودی با فرمت txt. تعداد لغات، کاراکتر و سطر فایل متنی را چاپ کرده و در انتهای آن فایل ورودی نوشته و ذخیره کند. توجه داشته باشید که این تابع فقط باید فایل با فرمت txt. را به عنوان ورودی گرفته و پردازش کند.

۴- فایل zen_of_python.txt را به عنوان ورودی به تابع نوشته شده در قسمت قبل دهید و نتیجه را ذخیره کرده و همراه با تمرین خود تحویل دهید.

فایل csv

فایل patents.csv را خوانده و به سوالات زیر پاسخ دهید:

۱- تعداد سطرهای و ستون‌های این دیتاست را چاپ کنید.

۲- اسم ویژگی‌های این دیتاست را چاپ کنید.

۳- ویژگی‌های آماری مانند میانگین، میانه، چارک اول و سوم، مینیمم و ماکسیمم و ... را برای ویژگی‌های عددی دیتاست به دست آورید.

۴- ۵ سطر اول دیتاست، ۵ سطر آخر و همچنین ۵ سطر به صورت رندوم از دیتاست را چاپ کنید.

۵- نوع داده هر ستون را چاپ کنید.

۶- مقدار داده‌های ویژگی title را سطرهای ۱۰۲۴ تا ۲۰۴۸ چاپ کنید. (خود دیتاپوینت ۲۰۴۸ نیز چاپ شود). برای اینکار از دو متد loc وiloc کتابخانه pandas استفاده کنید. تفاوت این دو متد در چیست؟

۷- بررسی کنید که آیا این دیتاست دارای مقدار NaN می‌باشد یا خیر؟

۸- بررسی کنید که آیا این دیتاست دارای دیتاپوینت (سطر) تکراری است یا خیر؟

۹- با استفاده از lambda function و متدهای مناسب در کتابخانه pandas تعداد کلمات موجود در ویژگی title را برای هر سطر حساب کرده و در این مقدار را برای هر دیتاپوینت در یک ویژگی جدید به نام title_length ذخیره کنید.

۱۰- نمودار هیستوگرام ویژگی جدید ایجاد شده در قسمت قبل را با استفاده از متدهای موجود در کتابخانه pandas رسم کنید.

فایل log

با خواندن فایل git_log.log موارد خواسته شده در بخش زیر را انجام دهید:

۱- تعداد کل commitهای موجود در این فایل را چاپ کنید.

۲- برای هر دولوپر تعداد کامیت‌هایی که انجام داده است و همچنین ایمیل و تاریخ آخرین کامیت را چاپ کنید.

۳- با parse کردن فایل لاگی که در اختیارات قرار داده شده یک دیتافرم pandas تشکیل دهید که دارای ستون‌های developer (اسم برنامه نویس)، email (ایمیل برنامه نویس)، commit_count (تعداد کامیت‌های هر برنامه نویس) و last_commit_date (تاریخ آخرین commit برنامه نویس) را شامل شود.

۴- با استفاده از متدهای کتابخانه pandas نام برنامه نویس‌هایی که کمترین و بیشترین commit را داشته‌اند را چاپ کنید.

فایل json

هر کدام از فایل‌های json موجود در پوشه videos شامل اطلاعات یک ویدیو در youtube می‌باشد. در این قسمت میخواهیم با خواندن فایل‌های json یک دیتافریم pandas از این اطلاعات تشکیل دهیم. با توجه به این موضوع موارد خواسته شده زیر را انجام دهید.

۱- با parse کردن هر کدام از فایل‌های json یک دیتا فریم pandas تشکیل دهید. که دارای ویژگی‌های زیر باشد:

- Title: عنوان ویدیو.
- Lang: زبان ویدیو.
- Record_date: تاریخ ضبط ویدیو. (نوع داده این ستون حتما باید datetime باشد.)
- url: لینک ویدیو.
- Description: توضیحات مربوط به ویدیو.
- Category: دسته مربوط به ویدیو.
- Tags: تگ‌هایی که در ویدیو استفاده شده است.
- Speakers: افرادی که در ویدیو حضور دارند.
- Duration: مدت زمان ویدیو.

در صورت نبود هر یک از ویژگی‌ها در هر کدام از فایل‌ها json مقدار آن را NaN قرار دهید.

۲- در صورت وجود مقدار NaN در هر کدام از ویژگی‌ها تعداد آن را چاپ کنید و همچنین با توجه به نوع داده هر ویژگی و با استفاده از یک روش مناسب مقادیر NaN را با مقدار مناسب پر کند.

۳- عنوان ویدیوهایی که سال ۲۰۱۶ منتشر شده‌اند را به دست آورید.

۴- مقادیر میانگین، مینیمم، ماکسیمم و میانه ویژگی duration هر دسته به دست آورید.

۵- با استفاده از متدهای مناسب کتابخانه pandas یک ویژگی جدید برای این دیتافریم با نام label ایجاد کنید به این صورت که اگر مدت زمان ویدیو کمتر از ۱۰۰۰ بود مقدار این ویژگی برابر با ۱، اگر مدت زمان ویدیو بیشتر از ۱۰۰۰ و کمتر از ۲۰۰۰ بود مقدار این ویژگی برابر با ۲ و اگر مدت زمان ویدیو بیشتر از ۲۰۰۰ بود مقدار این ویژگی برابر با ۳ باشد.

۶- باتوجه به ویژگی ایجاد شده در قسمت قبل، عنوان اولین و آخرین ویدیو را در هر دسته را با توجه به تاریخ به دست آورید.

۷- (امتیازی) ابر کلمات یا همان word cloud توضیحات ویدیوها را با استفاده از کتابخانه‌های مورد نیاز برای ۵۰ کلمه پرتکرار به دست آورید. لازم به ذکر است که برای به دست آوردن word cloud باید کلمات stop word را از توضیحات حذف کنید.

نکات پیاده سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز شنبه ۲۲ مهرماه ماه خواهد بود.
- انجام این تمرین به صورت یک نفره می باشد.
- خروجي مورد انتظار تمرین فایل jupyter ضمیمه شده و فایل zen_of_python.txt که بر روی آن تغییرات گفته شده را اعمال کرده‌اید، می‌باشد.
- هر گونه توضیحات و گزارشنویسي را به صورت Markdown داخل کتابچه jupyter انجام دهید.
- هر گونه تشابه میان تمرین‌های تحویل داده شده به عنوان تقلب در نظر گرفته می‌شود.
- در صورت استفاده از کدهای آماده، لینک مورد استفاده حتما ذکر شود.
- لطفا گزارش، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمائید.

HW1_[Lastname]_[StudentNumber].zip

برای مثال: HW1_mesbah_12345678.zip

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

امیرحسین مصباح – amir.mesbah@ut.ac.ir

شاد و سلامت باشید ☺