

بسمه تعالی

مستندات توضیحات پروژه اول بوت کمپ تحلیل داده



آخرین ویرایش

۱۴۰۱/۱۱/۰۸

هدف این پروژه استخراج داده‌هایی از کافی‌شاپ‌های سراسر کشور و تحلیل آن‌هاست. برای انجام این پروژه لازم است اطلاعات کافی‌شاپ‌های کشور را به دست آورده و پس ذخیره‌سازی آن‌ها در پایگاه داده، اقدام به واکشی اطلاعات و تحلیل آن‌ها کنید.

وبسایت [فیدیلیو](#) به عنوان منبع اصلی این پروژه انتخاب شده است، بنابراین انتظار این است که با بررسی و استخراج اطلاعات از همین سایت بتوانید نیازمندی‌ها و خواسته‌های پروژه را برآورده کنید. این پروژه در چندین گام تعریف شده است که در لیست زیر، آن‌ها را می‌توانید مشاهده کنید.

- استخراج اطلاعات از سایت فیدیلیو؛ باید توجه کنید که خواسته پروژه این است که تنها اطلاعات کافی‌شاپ‌ها را استخراج کنید. اجباری در مورد استخراج اطلاعات رستوران‌ها یا شرینی‌فروشی‌ها وجود ندارد. در این مرحله از پروژه، هر آنچه از web scraping آموختید را به کار بگیرید.

- پس از استخراج اطلاعات، نیاز است آن‌ها را در یک پایگاه داده (DataBase) ذخیره کنید. طراحی شمای کلی جداول و ارتباط آن‌ها یکی دیگر از اهداف پروژه است. بنابراین ایجاد جدول‌ها و تعریف ویژگی‌های مختلف برای آنان به عهده خود شماست. اطلاعاتی که لازم است به ازای هر کافی‌شاپ استخراج کنید را در لیست زیر می‌بینید:

- اسم
- شهر
- آدرس
- شماره تلفن
- ساعات کاری (ساعت شروع و ساعت پایان)
- نسبت قیمت (در صفحه اصلی فیدلیو با علامت \$ مشخص شده است؛ همچنان عکس این مورد را در زیر مشاهده می‌کنید که دور آن دایره قرمز کشیده شده است!)



- امتیازات کافی‌شاپ (امتیاز غذا، سرویس، هزینه، محیط و ...)
- ویژگی‌هایی که هر کافه ارائه می‌دهد ( به عنوان مثال قلیان، اینترنت رایگان، موسیقی زنده، ارسال به منزل، پارکینگ، دستگاه POS و ...)

نیاز است از پایگاه داده‌ای که در اختیار هر گروه قرار داده می‌شود استفاده کنید. نحوه اتصال به پایگاه داده و توضیحات بیشتر را می‌توانید در بخش ضمایم همین مستند مطالعه کنید.

- پس از انجام دو گام بالا نیاز است به وسیله streamlit یا PowerBI برای اطلاعات کافی شاپ‌ها داشبورد بسازید. در داشبوردی که توسعه می‌دهید باید تحلیل‌هایی را قرار دهید که به نظر خودتان مفید و کاربردی هستند، به عنوان مثال می‌توانید موارد زیر را نیز در نظر داشته باشید.

- توزیع (نمودار میله‌ای) قیمت‌ها
  - توزیع (نمودار میله‌ای) امتیاز کافی شاپ‌ها
  - توزیع (هیستوگرام) دنبال کننده‌ها
  - تحلیل‌های مبتنی بر ساعت کاری کافی شاپ‌ها
  - تحلیل‌های مبتنی بر ویژگی‌ها و آپشن‌هایی که کافی شاپ‌ها ارائه می‌دهند (سیگار، قلیان، موسیقی و ...)
  - تحلیل امتیازات/ویژگی‌های هر کافه بر اساس شهر
  - امکان مقایسه آماری امتیازات دو کافه؛ مثلاً بر اساس امتیاز کلی، بتوانیم با 99% اطمینان بگوییم کافه A بهتر از کافه B است.
- می‌توانید از [این لینک](#) و [این لینک](#) و [این یکی لینک](#) نیز برای ایده گرفتن و ساخت داشبورد نمونه استفاده کنید.
- دقت کنید طراحی داشبورد و استخراج و مصورسازی مناسب داده‌ها بر اساس تحلیل شما بسیار در نمره‌دهی تاثیر دارد.

## امتیازات اضافه

- اضافه کردن سورس‌های مختلف داده، به عنوان امتیاز اضافه محسوب می‌شود. به عنوان مثال می‌توانید از اسنپ‌فود، استارباکس، Google places API و هر منبع دیگری استفاده کنید. چالش این مرحله، بخش یکپارچه‌سازی (integration) اطلاعات از منابع مختلف است. به عنوان مثال ممکن است اسنپ‌فود ویژگی‌هایی مانند قلیان که هر کافه ارائه می‌دهد را نداشته باشد، یا استخراج آن‌ها مشکل باشد.
- طراحی داشبورد به صورتی که کاربر بتواند با آن تعامل کند، یکی از امتیازات ویژه محسوب می‌شود. به عنوان مثال کاربر بتواند لیست کافی شاپ‌هایی را مشاهده کند که امکان سیگار کشیدن در آن‌ها فراهم است و انتخاب «سیگار کشیدن» یا باقی موارد به دلخواه کاربر باشد.
- استقرار پروژه خود در سروری که دسترسی به آن برای همه کاربران اینترنت امکان‌پذیر باشد (Deployment)

- ذخیره‌سازی و تحلیل داده‌های غیر از کافی‌شاپ‌ها. به عنوان مثال می‌توانید گام‌هایی که برای کافی‌شاپ تشریح شد را برای رستوران‌ها نیز پیاده‌سازی کنید.

## همکاری منتورها و داوری پروژه

در این پروژه، منتور هر گروه در نقش Product Manager ظاهر شده و نحوه پیشبرد پروژه را مدیریت خواهد کرد. ضمن هماهنگی‌های لازم برای انجام پروژه، می‌توانید برای اشکالات فنی نیز از منتورها کمک بگیرید. منظور از مدیریت پروژه، مشخص کردن تسک‌ها، انتصاب آن‌ها و پیگیری برای انجام شدنشان و کمک به اعضای تیم برای پیاده‌سازی تسک‌هاست؛ اما منتورها در فرایند کدزنی نقشی نخواهند داشت.

داوری پروژه به این صورت خواهد بود که پس از اتمام پروژه، همه گروه‌ها نتیجه تلاش خود را به یکی از منتور ارائه خواهند داد. در نهایت داور با مشورت با منتور گروه، نمره نهایی پروژه را در کوئرا ثبت خواهد کرد. امتیازدهی این پروژه شامل معیارهای متنوعی نظیر کیفیت محصول خروجی، همکاری گروهی، امتیازات اضافه و ... است.

آنچه لازم است در کوئرا آپلود کنید، zip شده تمامی کدهایی است که برای پروژه نوشته‌اید. از زیپ‌کردن پایگاه داده لوکال یا فایل‌های csv خودداری کنید تا حجم فایل ارسالی شما زیاد نشود!

## ضمایم

### نحوه اتصال به پایگاه داده

برای هر گروه یک پایگاه داده در سروری به آدرس 37.32.5.76 و شماره پورت 3306 ایجاد شده است. کد زیر نشان می‌دهد چگونه می‌توان به پایگاه داده‌ای که در سرور قرار دارد، متصل شد.

```
from sqlalchemy import create_engine
user = 'USERNAME FOR YOUR GROUP'
password = 'PASSWORD FOR YOUR GROUP'
host = "37.32.5.76:3306" # THIS FIELD IS IDENTICAL FOR ALL GROUPS => DON'T CHANGE IT :)
db = 'DATABASE NAME OF YOUR GROUP'
engine = create_engine(f"mysql+mysqlconnector://{user}:{password}@{host}/{db}")
```

برای اجرا شدن کد بالا نیاز است به وسیله package manager ای که استفاده می‌کنید (pip یا هر package manager دیگری) دو کتابخانه sqlalchemy و mysqlconnector را نصب کنید. این پایگاه داده MySQL بوده و نیاز است فرمت نوشتاری کوئری‌ها را بر اساس همین دیتابیس رعایت کنید (با استفاده از فرمت MSSQL کوئری‌های شما اجرا نخواهند شد). در کد بالا، متغیرهای db, password, user به ازای هر گروه، متفاوت است. بنابراین این سه متغیر از طرف منتورهای هر گروه به اعضا منتقل خواهد شد.

توجه داشته باشید که یوزری که از منتور دریافت خواهید کرد، دارای بالاترین سطح دسترسی به پایگاه داده همان گروه است. بنابراین در اجرای کوئری‌ها دقت لازم را داشته باشید تا اتفاق ناگواری برای پایگاه داده گروه شما رخ ندهد!