دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده علوم کامپیوتر

گزارش پروژه 1

کار و تحقیق روی کتابخانه‌های داده‌کاوی پایتون

نگارش
ثمین مهدی پور

استاد راهنما
دکتر قطعی

استادمشاور
دکتر قطعی

مهر 1402

چکیده

در این گزارش به بررسی کتابخانه‌های زبان پایتون که در پروژه‌های داده‌کاوی مورد استفاده قرار می‌گیرند می‌پردازیم. با استفاده از این کتابخانه‌ها میتوان مراحل انجام یک پروژه مانند پیش‌پردازش را با دقت و سرعت بالا پیش برد. در هربخش کلیات کتابخانه، نحوه نصب و توابع تشکیل‌دهنده و چگونگی استفاده آن بصورت خلاصه ذکر شده است.

واژه‌های کلیدی:

پایتون، کتابخانه، داده‌کاوی، پیش‌پردازش

فهرست مطالب

صفحه

فهرست اشکال

صفحه

صفحه

فهرست جداول

فهرست علائم

علائم لاتین

h	ارتفاع
L	طول موج توربولانس
T	پریود توربولانس
U_0	سرعت تعادل وسیله پرنده
$u_{\mathrm{g}}^{\mathrm{B}}$	مولفه سرعت تندباد در راستای محور طولی دستگاه مختصات بدنی نسبت به اینرسی

علائم یونانی

$\Phi(\omega)$	چگالی طیفی قدرت توربولانس
σ	شدت توربولانس
ω	بسامد توربولانس
Ω	بسامد فاصله‌ای

بالانویس‌ها

B	دستگاه مختصات بدنی

زیرنویس‌ها

g	تندباد (گاست)

فصل اول

مقدمه

مقدمه

دنیای امروز هرلحظه وابسته به سرعت رشد تکنولوژی درحال تغییر است، در چنین شرایطی پیاده‌سازی پروژه های مبتنی بر هوش‌مصنوعی نیازمند سرعت و دقت بالا هستند. امکان پیاده‌سازی تمامی مراحل یک پروژه بصورت فردی وجود دارد اما چنین رویکردی نه تنها دقت کافی و استاندارد مدنظر را به همراه نخواهد داشت بلکه مدت زمانی بیشتر از میزان تعریفی برای پروژه را نیاز خواهد داشت. به منظور استانداردسازی و سرعت بخشی به پروژه‌های تعریفی در زبان پایتون کتابخانه‌هایی توسط شرکت و گروه های خبره آماده شده است که علاوه بر کیفیت بالا امکان پیاده‌سازی آسان و سریعی دارند. در پروژه های داده کاوی مراحل پردازش داده‌ها تا نتیجه‌گیری هرکدام نیازمند استفاده از کتابخانه های مختلفی هستند که تسلط و آگاهی از چگونگی به کاربردن آنها لازمه انجام هرپروژه ای است.

فصل دوم

معرفی کتابخانه‌ها

امروزه پایتون پرکاربردترین زبان برنامه‌نویسی است. وقتی نوبت به حل مسائل و چالش‌های علوم داده می‌رسد، پایتون باعث شگفتی کاربران خود می‌شود. در حال حاضر بیشتر دانشمندان داده از زبان برنامه‌نویسی پایتون استفاده می‌کنند. پایتون یک زبان آسان برای یادگیری، اشکال زدایی آسان، با استفاده‌ی گسترده، شی گرا، متن بازو با کارایی بسیار بالا است و برنامه‌نویسی با پایتون مزیت های بسیار زیادی دارد. در پایتون کتابخانه‌های فوق‌العاده‌ای برای علم داده ایجاد شده‌است که هر روز توسط برنامه نویسان برای حل مسائل استفاده می‌شوند. در ادامه به بررسی این کتابخانه‌ها می‌پردازیم:

NumPy[1] -1-2

NumPy یک کتابخانه برای زبان برنامه نویسی پایتون (Python) است. با استفاده از این کتابخانه امکان استفاده از آرایه‌ها و ماتریس‌های بزرگ چند بعدی فراهم می‌شود. هم‌چنین می‌توان از تابع‌های ریاضیاتی سطح بالا بر روی این آرایه‌ها استفاده کرد. از جمله دلایل مورد استفاده قرار گرفتن این کتابخانه در مسائل داده کاوی میتوان به موارد زیر اشاره کرد:

- ذخیره‌سازی مؤثر داده: NumPy روشی مؤثر برای ذخیره و مدیریت آرایه‌های بزرگ داده فراهم می‌کند. در داده‌کاوی که اغلب با مجموعه‌های داده بزرگ کار می شود، آرایه‌های NumPy کارآمدتر و سریع‌تر از لیست‌های پایتون برای پردازش عمل می‌کنند.
- عملیات آرایه‌ای: NumPy مجموعه‌ای گسترده از عملیات ریاضی و منطقی ارائه می‌دهد که می‌توان آنها را بر روی آرایه‌ها یا عناصر داخل آرایه‌ها اعمال کرد. این عملیات در پیش‌پردازش داده و مدیریت ویژگی‌ها نقشی اساسی ایفا می‌کنند. به عنوان مثال، شخص می‌تواند به راحتی عملیاتی مانند میانگین، میانه، انحراف معیار یا مجموع را بر روی آرایه‌های داده اعمال کند.

- پاک‌سازی داده: در داده‌کاوی اغلب نیاز به پاک‌سازی و پیش‌پردازش داده داریم، و NumPy ابزارهایی برای مدیریت مقادیر نامعلوم، فیلتر کردن داده و اعمال تغییرات بر داده فراهم می‌کند.

[1] https://numpy.org/

با استفاده از توابع NumPy، می‌توان مقادیر نامعلوم را جایگزین کرد، ردیف‌ها را بر اساس شرایط فیلتر کرد یا داده را تغییر داد.

- جبر خطی: بسیاری از الگوریتم‌های داده‌کاوی مانند تجزیه مؤلفه‌های اصلی (PCA) و تجزیه مقدار تک (SVD) بر روی عملیات جبر خطی استوار هستند. NumPy برای عملیات ماتریسی و جبر روش‌هایی ارائه می‌کند.
- تولید اعداد تصادفی: NumPy شامل توابعی برای تولید اعداد تصادفی است که در نمونه‌برداری داده، ایجاد مجموعه‌داده‌های مصنوعی و استفاده از نمونه‌برداری برای تحلیل آماری مفید واقع می‌شود.
- ادغام با کتابخانه‌های دیگر: NumPy اغلب همراه با کتابخانه‌های دیگری که به طور معمول در داده‌کاوی مورد استفاده قرار می‌گیرند مانند Scikit-Learn، SciPy و Pandas به کار می‌رود. بسیاری از این کتابخانه‌ها ورودی داده را به صورت آرایه‌های NumPy در نظر میگیرند.

NumPy یک کتابخانه مهم در داده‌کاوی است چرا که ساختارهای داده کارآمد و مجموعه‌ای گسترده از توابع ریاضی و مدیریت آرایه‌ها را ارائه می‌کند. این کتابخانه پایه بسیاری از کتابخانه‌ها و ابزارهای داده‌کاوی دیگر در پایتون را تشکیل می‌دهد، که آن را به یک جزء ضروری در پیش‌پردازش، تحلیل و مدل‌سازی داده تبدیل می‌کند.

TensorFlow[1] -2-2

تنسورفلو کتابخانه‌ای برای انجام محاسبات عددی با کارایی بالاست. کتابخانه تنسورفلو در زمان نگارش این متن دارای ۳۵۰۰۰ نظر مخاطب و یک جامعه‌ی مشارکت‌کننده‌ی فعال ۱۵۰۰ نفری می‌باشد. که از آن در زمینه‌های مختلف علمی استفاده می‌شود. تنسورفلو اساسا یک چارچوب برای تعریف و انجام محاسبات است که شامل تنسورها (بردارها-ماتریس‌ها) است و بر اساس کلاس‌های خود امکان ایجاد اشیای محاسباتی را می‌دهد. استفاده از TensorFlow (یا TensorFlow 2.x با Keras) در داده کاوی و تحلیل داده می‌تواند بسیار مفید باشد، به ویژه زمانی که مسائلی از قبیل تشخیص الگو، پیش‌بینی و

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

تحلیل تصاویر یا متن، یادگیری عمیق و شبکه‌های عصبی تخصصی مورد نیاز باشند. TensorFlow به عنوان یک کتابخانه قدرتمند در زمینه یادگیری عمیق و محاسبات علمی عمومی شناخته می‌شود و در داده کاوی در موارد زیر می‌تواند مفید باشد:

- یادگیری عمیق : از نسخه 2 به بعد با Keras ادغام شده است و برای ایجاد و آموزش شبکه‌های عصبی عمیق بسیار ساده و کارآمد استفاده می‌شود. این می‌تواند در تشخیص الگوها، پیش‌بینی متغیرهای پیچیده و تحلیل تصاویر و ویدئوها کمک کند.
- پردازش تصویر: TensorFlow به عنوان یکی از ابزارهای اصلی برای پردازش تصویر و تشخیص اشیاء در تصاویر استفاده می‌شود. از شبکه‌های عصبی کانولوشنی (CNN) تا شبکه‌های معمولی (MLP) برای تصویربرداری و پردازش تصویر استفاده می‌شود.
- پردازش متن و پردازش زبان طبیعی: TensorFlow برای پردازش متن و تسلط بر مسائل NLP به کار می‌رود. از تبدیل‌های متنی به بردارهای عددی (Word Embeddings) تا آموزش مدل‌های زبانی مبتنی بر شبکه‌های عصبی (مانند مدل‌های BERT)، TensorFlow بسیار کارآمد است.
- یادگیری تقویتی: TensorFlow برای پیاده‌سازی و آموزش الگوریتم‌های یادگیری تقویتی مورد استفاده قرار می‌گیرد، که در مسائلی مانند بازی‌های رایانه‌ای و بهینه‌سازی مفید است.
- پردازش گفتار: برای تشخیص و ترجمه گفتار، TensorFlow با استفاده از مدل‌های شبکه‌های عصبی بازگشتی (RNN) و شبکه‌های معمولی (MLP) کاربرد دارد.
- پردازش گراف و داده‌های بزرگ: TensorFlow گاهی اوقات برای پردازش گراف‌های بزرگ و داده‌هایی با حجم زیاد استفاده می‌شود.

SciPy[1] -3-2

SciPy (پایتون علمی) یک کتابخانه رایگان و منبع باز پایتون برای علوم داده است که به طور گسترده برای انجام محاسبات سطح بالا مورد استفاده قرار می‌گیرد. SciPy دارای ۱۹۰۰۰ نظر مخاطب در گیت‌هاب و یک جامعه‌ی مشارکت‌کننده‌ی فعال ۶۰۰ نفری می‌باشد. به طور گسترده برای محاسبات علمی و فنی مورد استفاده قرار می‌گیرد. زیرا NumPy را گسترش می‌دهد و روال‌های کاربر پسند و کارآمد زیادی را برای محاسبات علمی فراهم می‌کند. در ادامه توضیحاتی در مورد استفاده از SciPy در داده‌کاوی ارائه می‌شود:

- پردازش سیگنال و تصویر: SciPy دارای ابزارها و توابع بسیاری برای پردازش سیگنال‌ها و تصاویر است. این ویژگی در تجزیه و تحلیل سیگنال‌های صوتی، تصاویر دیجیتال، و پردازش تصویر مفید است.
- تحلیل آماری: SciPy ابزارهای متنوعی برای تحلیل آماری دارد، از جمله توزیع‌های احتمال، آزمون‌های آماری، توابع تخمین توزیع و آماره‌های توصیفی.
- بهینه‌سازی: SciPy دارای ابزارهای بهینه‌سازی برای حل مسائل بهینه‌سازی عددی است. این امکان را فراهم می‌کند تا در بهینه‌سازی پارامترها و مدل‌های مختلف در داده‌کاوی و یادگیری ماشین بهره‌برداری کنید.
- تحلیل ماتریس: SciPy ابزارها و توابع بسیاری برای تحلیل ماتریس‌ها و ماتریس‌های خطی دارد. این ویژگی در تجزیه و تحلیل داده‌های چند متغیره و پیاده‌سازی الگوریتم‌های مرتبط با تجزیه مؤلفه‌های اصلی (PCA) و تجزیه مقدار تک (SVD) مفید است.
- بهره‌برداری از توابع انتگرال‌گیری: SciPy ابزارها و توابع بسیاری برای محاسبه انتگرال‌ها و حل معادلات دیفرانسیلی دارد. این امکان را می‌دهد تا در مدل‌سازی و تحلیل داده‌ها با استفاده از معادلات ریاضی پیشرفته بهره‌برداری کنیم.

[1] https://scipy.org/

- کار با داده‌های تنسوری: SciPy از داده‌ساختارها و توابعی برای کار با داده‌های تنسوری پشتیبانی می‌کند. این مورد در وظایف مرتبط با یادگیری عمیق و شبکه‌های عصبی بسیار مفید است.

در کل، SciPy یک کتابخانه قدرتمند برای انجام محاسبات علمی و مهندسی در داده‌کاوی و تحلیل داده است. این کتابخانه به توسعه‌دهندگان اجازه می‌دهد تا به طور مؤثر از ابزارهای پیشرفته ریاضی و علمی برای استخراج اطلاعات از داده‌ها و ایجاد مدل‌های پیشرفته استفاده کنند.

Pandas[1] -4-2

Pandas (تحلیل داده با پایتون) یکی از ضروریت‌ها در چرخه‌ی حیات علوم داده است. این کتابخانه محبوب‌ترین و پراستفاده‌ترین کتابخانه‌ی پایتون برای علم داده، همراه با NumPy در matplotlib است. Pandas دارای 1700 نظر مخاطب در گیت‌هاب و یک جامعه‌ی مشارکت‌کننده‌ی فعال ۱۲۰۰ نفری می‌باشد. که به شدت برای تجزیه و تحلیل و پاک‌سازی داده‌ها استفاده می‌شود. Pandas ساختارهای داده‌ای سریع و انعطاف‌پذیری را فراهم می‌کند که برای کار با داده‌های ساخت‌یافته بسیار آسان و شهودی طراحی شده‌اند. در ادامه، برخی از کاربردها و قابلیت‌های Pandas در داده‌کاوی را توضیح می‌دهیم:

- خواندن و نوشتن داده: Pandas ابزارهای قدرتمندی برای خواندن داده از مختلف منابع مانند فایل‌های CSV، اکسل، پایگاه‌های داده SQL و حتی از وب را ارائه می‌دهد.
- ساختاردهی داده: Pandas داده‌های خود را در قالب دو ساختار اصلی به نام DataFrame و Series نگهداری می‌کند. DataFrame یک جدول داده با ستون‌ها و ردیف‌ها است و Series یک ستون یا ستون‌های خاص از یک DataFrame است
- پاک‌سازی و پیش‌پردازش داده: Pandas ابزارهای مختلفی برای پاک‌سازی داده‌ها و پیش‌پردازش آنها ارائه می‌دهد. این شامل حذف مقادیر نامعلوم، تبدیل داده‌ها، فیلتر کردن ردیف‌ها بر اساس شرایط مشخص و دیگر عملیات پاک‌سازی مرتبط با داده می‌شود.

[1] https://pandas.pydata.org/

- انتخاب و فیلتر کردن داده: Pandas این امکان را می‌دهد داده‌ها را بر اساس معیارهای مختلف انتخاب و فیلتر کنیم. می‌توان ستون‌ها و ردیف‌های مورد نظر خود را انتخاب کرده و داده‌های خود را به شکل دقیق‌تری برای تحلیل انتخاب کرد.
- تجزیه و تحلیل داده: Pandas ابزارهای متنوعی برای تجزیه و تحلیل داده‌ها ارائه می‌دهد. می‌توان انواع مختلفی از آمارها، تجزیه و تحلیل متغیرها و ایجاد نمودارهای تصویری برای درک بهتر داده‌ها استفاده کرد.
- ادغام و ترکیب داده‌ها: می‌توان داده‌ها را از منابع مختلف خواند و آنها را با یکدیگر ترکیب کرد یا ادغام کرد.

در کل، Pandas یک ابزار بسیار قدرتمند و اساسی در داده‌کاوی و تحلیل داده‌های ساختار یافته است که به ترتیب و سازماندهی داده‌ها، پیش‌پردازش، تجزیه و تحلیل و نمایش داده‌ها کمک می‌کند. این کتابخانه برای تحلیل داده‌ها و ایجاد مدل‌های پیش‌بینی بسیار مفید است و یکی از ابزارهای اصلی در دانش داده و علوم داده به حساب می‌آید.

Matplotlib[1] -5-2

کتابخانه Matplotlib یکی از ابزارهای قدرتمند در داده‌کاوی و تجزیه و تحلیل داده‌ها در پایتون است. این کتابخانه امکان می‌دهد تا داده‌های خود را به صورت گرافیکی نمایش دهیم و از تصاویر و نمودارها برای درک بهتر داده‌ها و ارائه نتایج به صورت بصری استفاده کنیم. در ادامه توضیحاتی در مورد استفاده از کتابخانه Matplotlib در داده‌کاوی ارائه می‌شود:

- رسم نمودارها: Matplotlib امکان می‌دهد تا انواع مختلف نمودارها را رسم کرد، از جمله نمودارهای خطی، نمودارهای نقطه‌ای، نمودارهای میله‌ای، نمودارهای دایره‌ای و بسیاری دیگر. این نمودارها معمولاً برای نمایش توزیع داده‌ها، روابط بین متغیرها و الگوهای داده‌ای استفاده می‌شوند.

[1] https://matplotlib.org/

- تاریخچه‌ها و زمان‌بندیاگر داده‌های شما سری زمانی باشند، Matplotlib ابزارهـای قدرتمنـدی برای رسم تاریخچه‌ها و نمودارهای زمانی ارائه می‌دهد. این کاربرد خصوصاً در تحلیـل داده‌هـای مالی، اقتصادی و زمانی مفید است.
- نمایش داده‌های چند بعدی: اگر داده‌ها دارای بیش از یک ویژگـی (بعـد) باشـند، Matplotlib می‌تواندر رسم نمودارهای داده‌های چند بعدی کمک کنـد. ایـن نمودارهـا معمـولاً بـه منظـور کاوش داده‌ها و تجزیه و تحلیل روابط بین ویژگی‌ها به کار می‌روند.
- نمودارهای پیچیده: Matplotlib اجازه می‌دهـد تـا نمودارهـای پیچیـده‌تری مثـل نمودارهـای شبکه‌ای، نمودارهای سه بعدی و نمودارهای پیکانی را رسم کنیم. ایـن ابزارهـا ممکـن اسـت در تجزیه و تحلیل داده‌های پیچیده و ساختارهای پیچیده مفید باشند.
- نمودارهای توزیع و تغییر: Matplotlib با ارائه ابزارهایی مثل نمودارهای توزیع احتمالاتی (مانند نمودارهای هیستوگرام) و نمودارهای تغییر (مانند نمودارهای Q-Q می‌توانـد در تحلیـل توزیـع داده‌ها و تشخیص اشکالات مفید عمل کند.
- نمایش توصیف‌ها و توضیحات: با استفاده از Matplotlib، می‌توان توصیف‌ها، توضیحات و اعداد را بر روی نمودارها نمایش داد تا داده‌ها را بهتر توضیح داده شود.

در کل، Matplotlib یک ابزار قدرتمند و چندمنظوره است که محور تجزیه و تحلیل داده‌ها را به صورت بصری تسهیل می‌کند و در تفسیر و ارتقاء فهم داده‌ها در داده‌کاوی بسیار مفید است.

Keras[1] -6-2

Keras یک کتابخانه بسیار قدرتمند برای ایجاد و آموزش شبکه‌های عصبی مصنوعی در پایتون است که برای وظایف داده کاوی نیز بسیار مفید است. در ادامه به نحوه اسـتفاده از Keras در داده کـاوی اشـاره خواهد شد:

[1] https://keras.io/

- تعریف معماری شبکه عصبی: با استفاده از Keras، می توان معماری مورد نظر خود برای شبکه عصبی را تعریف کرد که شامل تعیین تعداد لایه‌ها، تعداد نورون‌ها در هر لایه، و نوع لایه‌ها مانند لایه‌های متفرقه (Dense)، روندهای بازگشتی (Recurrent) و تنظیمات دیگر می‌شود.
- تعیین توابع فعال‌سازی: با Keras، می توان توابع فعال‌سازی مربوط به هر لایه را انتخاب کرد. این توابع معمولاً به عنوان توابع غیرخطی در شبکه عصبی عمل می‌کنند.
- تنظیم معیارهای عملکرد: در وظایف داده کاوی، معیارهای عملکرد مهمی مثل دقت (Accuracy)، دقت میانگین معکوس (F1-score) و خطا میانگین مربعات (Mean Squared Error) هستند. می‌توان معیارهای مورد نیاز خود را در Keras تنظیم کرد.
- آموزش مدل: با استفاده از داده‌های آموزش، می‌توان مدل شبکه عصبی را با تابع آموزشی مانند "compile" و "fit" در Keras آموزش داد. در این مرحله، مدل وزن‌های خود را با تطابق بین پیش‌بینی‌های مدل و مقادیر واقعی در داده‌های آموزش به‌روز می‌کند.
- ارزیابی مدل: بعد از آموزش، می توان مدل را با استفاده از داده‌ها ارزیابی کرد. این مرحله شامل اندازه‌گیری معیارهای عملکرد و تحلیل نتایج مدل می‌شود.
- پیش‌بینی با مدل آموزش دیده: پس از آموزش، می‌توان مدل خود را برای پیش‌بینی خروجی بر روی داده‌های تست یا داده‌های جدید استفاده کرد.
- تنظیم پارامترهای مدل: Keras امکان تنظیم پارامترهای مدل مانند نرخ یادگیری، تعداد دوره‌های آموزش، و برخی سایر پارامترهای مهم را فراهم می‌کند.
- ذخیره و بازیابی مدل: مدل‌های Keras را می‌توان ذخیره کرد و در آینده بازیابی کرد تا نیازی به مجدد آموزش نداشته باشیم.

در کل، Keras یک ابزار قدرتمند برای توسعه و آموزش مدل‌های شبکه عصبی برای وظایف داده کاوی و پیش‌بینی است. این کتابخانه انعطاف‌پذیری بالایی دارد و از آسانی استفاده می‌شود، بنابراین محبوبیت زیادی در جامعه علم داده و یادگیری ماشین دارد.

Scikit-Learn[1] -7-2

Scikit-Learn یک کتابخانه معروف در زمینه یادگیری ماشین و داده‌کاوی در پایتون است که ابزارها و الگوریتم‌های متنوعی برای تحلیل و پیش‌بینی داده‌ها ارائه می‌دهد. این کتابخانه به عنوان یکی از مهم‌ترین ابزارها در داده‌کاوی برای انجام وظایف متنوعی مورد استفاده قرار می‌گیرد. در زیر به برخی از کاربردهای مهم Scikit-Learn در داده‌کاوی اشاره خواهم کرد:

- آموزش مدل‌های پیش‌بینی: Scikit-Learn انواع مدل‌های یادگیری ماشین را ارائه می‌دهد، از جمله مدل‌های کلاسیک مانند رگرسیون خطی، درخت تصمیم، و ماشین‌های پشتیبانی‌کننده تا مدل‌های عمیق مانند شبکه‌های عصبی. این ابزارها اجازه می‌دهند تا با استفاده از داده‌های خود مدل‌های پیش‌بینی را آموزش دهیم.
- انتخاب ویژگی: Scikit-Learn ابزارهایی برای انتخاب ویژگی‌های مهم در داده‌های خود ارائه می‌دهد. این کاربرد مفید است زیرا معمولاً تعداد ویژگی‌ها در داده‌ها بسیار بزرگ است و نیاز به انتخاب ویژگی‌های مهم و حذف ویژگی‌های غیرمفید داریم.
- خوشه‌بندی: با استفاده از روش‌های خوشه‌بندی ارائه شده در Scikit-Learn، می‌توان داده‌ها را به گروه‌های مشابه تقسیم کرد. این کاربرد در تحلیل گروه‌های مشابه در داده‌ها یا در دسته‌بندی داده‌ها مفید است.
- ارزیابی عملکرد مدل‌ها: Scikit-Learn ابزارهایی برای ارزیابی عملکرد مدل‌های یادگیری ماشین در داده‌های تست ارائه می‌دهد. این ارزیابی‌ها شامل معیارهایی مانند دقت، بازیابی، و F1-score می‌شوند.
- پیش‌پردازش داده: قبل از استفاده از مدل‌های یادگیری ماشین، معمولاً نیاز به پیش‌پردازش داده‌ها داریم. Scikit-Learn ابزارهایی برای انجام پیش‌پردازش مانند مقیاس‌دهی و تبدیل داده‌ها ارائه می‌دهد.

[1] https://scikit-learn.org/stable/

- انجام تقسیم داده: برای آموزش و ارزیابی مدل‌ها، باید داده‌ها را به دسته‌های آموزش و تست تقسیم کرد. Scikit-Learn ابزارهایی برای تقسیم داده‌ها به صورت تصادفی یا به نسبت مشخص ارائه می‌دهد.
- مدیریت کران‌های داده: اغلب در داده‌کاوی، نیاز به مدیریت کران‌های داده داریم. Scikit-Learn ابزارهایی برای کران‌گذاری داده‌ها و پردازش داده‌های پرت و نویز دارد.

در کل، Scikit-Learn یک ابزار قدرتمند و گسترده در داده‌کاوی و یادگیری ماشین است که به تحلیل و استفاده از داده‌ها برای استخراج اطلاعات ارزشمند کمک می‌کند. این ابزارها به توسعه‌دهندگان و پژوهشگران در زمینه‌های مختلف امکان می‌دهند تا به راحتی مدل‌های پیش‌بینی بسازند و داده‌ها را تحلیل کنند.

PyTorch[1] -8-2

PyTorch یک کتابخانه معروف و قدرتمند در زمینه یادگیری عمیق و شبکه‌های عصبی است که برای وظایف مختلف داده کاوی نیز مورد استفاده قرار می‌گیرد. در زیر توضیحی در مورد استفاده از PyTorch در داده کاوی ارائه می‌شود:

- آموزش مدل‌های پیش‌بینی: PyTorch به عنوان یک کتابخانه قوی برای طراحی و آموزش مدل‌های یادگیری عمیق مورد استفاده قرار می‌گیرد. این مدل‌ها می‌توانند برای پیش‌بینی و تحلیل داده‌ها در وظایف مختلف داده کاوی مانند تصویربرداری، ترجمه متن، تشخیص الگوها و انتشارات اخبار مورد استفاده قرار گیرند.
- تشخیص الگوها و تصویربرداری: PyTorch به عنوان یکی از کتابخانه‌های برتر برای پردازش تصویر مورد استفاده قرار می‌گیرد. با استفاده از شبکه‌های عصبی کانولوشنی (CNN) در PyTorch می‌توانید الگوها را در تصاویر تشخیص داد.
- پردازش زبان طبیعی : برای وظایف مرتبط با پردازش متن و NLP، PyTorch به عنوان یک ابزار مورد استفاده قرار می‌گیرد.

[1] https://pytorch.org/

- خوشه‌بندی و کاوش داده: PyTorch می‌تواند در وظایف خوشه‌بندی و کاوش داده مورد استفاده قرار گیرد.
- تحلیل داده‌های سلسله مراتبی: PyTorch به عنوان یک ابزار برای مدل‌سازی داده‌های سلسله مراتبی و ساختارهای پیچیده مانند گراف‌ها و درخت‌ها نیز مورد استفاده قرار می‌گیرد. این امکان را فراهم می‌کند تا الگوها و روابط پنهان در داده‌ها را مدل کنیم.

PyTorch به عنوان یک کتابخانه انعطاف‌پذیر و توانمند برای یادگیری عمیق و پردازش داده‌ها، در وظایف مختلف داده کاوی مورد استفاده قرار می‌گیرد و امکاناتی جهت طراحی، آموزش، و ارزیابی مدل‌های پیش‌بینی و تحلیل داده فراهم می‌کند.

OpenCV[1] -9-2

یک کتابخانه منبع باز برای پردازش تصویر است که ابزارها و توابع متنوعی را برای کار با تصاویر و ویدیوها ارائه می‌دهد. این کتابخانه به صورت گسترده در بسیاری از حوزه‌ها از جمله داده کاوی و تجزیه و تحلیل تصاویر و ویدیوها به کار می‌رود. در زیر تعدادی از کاربردهای OpenCV در داده کاوی را توضیح داده میشود:

- پردازش تصاویر: OpenCV به تحلیل و پردازش تصاویر در داده کاوی کمک می‌کند. این کاربرد در شناسایی الگوها، شمارش اشیاء، تشخیص چهره‌ها، تحلیل تصاویر پزشکی، و غیره مورد استفاده قرار می‌گیرد.
- استخراج ویژگی‌ها: OpenCV ابزارها و توابعی برای استخراج ویژگی‌های تصویری مانند رنگ، شکل، حاشیه‌یابی، و تبدیل‌های هندسی ارائه می‌دهد. این ویژگی‌ها می‌توانند به عنوان ورودی برای الگوریتم‌های داده کاوی مورد استفاده قرار گیرند.
- تشخیص الگو: OpenCV امکان تشخیص الگوها و قالب‌ها در تصاویر را فراهم می‌کند. این کاربرد در تشخیص نمادها، لوگوها، یا شئ‌های خاص در تصاویر مورد استفاده قرار می‌گیرد.

[1] https://opencv.org/

- پردازش تصویر در زمینه ماشین لرنینگ: OpenCV به عنوان یکی از ابزارهای مهم برای پیش‌پردازش تصاویر و ویدیوها در مسائل مرتبط با ماشین لرنینگ و یادگیری عمیق مورد استفاده قرار می‌گیرد. این شامل تغییر اندازه تصاویر، افزایش داده‌ها (data augmentation)، و استخراج ویژگی‌ها می‌شود.
- پردازش تصاویر ویدیوOpenCV به تحلیل و پردازش داده‌های ویدیویی کمک می‌کند. این کاربرد در تشخیص حرکت، ترکیب تصاویر، و آنالیز ویدیوهای پیشرفته مورد استفاده قرار می‌گیرد.
- استخراج اطلاعات از تصاویر: OpenCV می‌تواند به استخراج اطلاعات مفهومی از تصاویر کمک کند. این اطلاعات می‌توانند به عنوان ورودی برای مسائل داده کاوی مانند دسته‌بندی تصاویر بر اساس محتوا یا تحلیل موضوعی تصاویر مورد استفاده قرار گیرند.

در کل، OpenCV به عنوان یک ابزار قدرتمند در پردازش تصویر و بینایی ماشین از داده‌های تصویری برای مسائل داده کاوی بهره‌برداری می‌کند و در انجام تعدادی از کاربردهای مفید در داده کاوی و تحلیل داده تأثیرگذار است.

Seaborn[1]-10-2

Seaborn یک کتابخانه پایتون است که به تجزیه و تحلیل داده و داده‌کاوی کمک می‌کند. این کتابخانه بر پایه کتابخانه Matplotlib ساخته شده است و امکانات بیشتری برای تولید نمودارهای زیبا و اطلاعاتی فراهم می‌کند.

Seaborn به ویژه برای تجزیه و تحلیل داده‌های توزیع‌شده و متغیرهای مختلف مفید است و مزیت‌های زیر را دارد:

- نمودارهای زیبا: Seaborn امکان ایجاد نمودارهای زیبا و آمیزه را فراهم می‌کند که به راحتی قابل فهم هستند. این نمودارها اغلب از طریق توابع ساده‌ای ایجاد می‌شوند.

[1] https://seaborn.pydata.org/

- پشتیبانی از داده‌های توزیع‌شده: اگر با داده‌هایی که توزیع‌های مختلف دارند کار می‌شود، Seaborn دارای توابع خاصی برای تولید نمودارهای توزیع این داده‌هاست که به تحلیل و تفسیر آنها کمک می‌کند.
- پایداری بالا: Seaborn از تنظیمات پیش‌فرض بهره می‌برد که باعث می‌شود نمودارهای تولید شده آماده‌تر باشند و نیاز به تنظیمات دستی کمتری داشته باشند.
- پشتیبانی از داده‌های Pandas: Seaborn به خوبی با ساختار داده‌های Pandas سازگار است و امکان تغییر نمایش داده‌های Pandas را فراهم می‌کند.
- نمایش روابط: این کتابخانه امکان نمایش روابط بین متغیرها را با استفاده از نمودارهای هیستوگرام چندمتغیره و نمودارهای رگرسیون فراهم می‌کند.

به طور کلی، Seaborn یک ابزار قدرتمند برای تجزیه و تحلیل داده‌ها و تولید نمودارهای اطلاعاتی و زیبا در داده‌کاوی است که باعث می‌شود فرایند تفسیر داده‌ها و استخراج اطلاعات مهم از آنها ساده‌تر و کارآمدتر باشد.

Plotly[1]-11-2

Plotly یک کتابخانه تصویرسازی تعاملی برای زبان برنامه‌نویسی پایتون است که امکان می‌دهد نمودارها و گراف‌های تعاملی و زیبا بسازیم. این کتابخانه برای داده‌کاوی و تجزیه و تحلیل داده بسیار مفید است زیرا امکان نمایش داده‌ها و الگوهای مختلف در داده‌ها را فراهم می‌کند. در زیر تعدادی از مزایای استفاده از Plotly در داده‌کاوی آورده شده است:

- تصاویر تعاملی: با استفاده از Plotly، می‌توانید نمودارها و گراف‌ها را به صورت تعاملی ساخته و تنظیم کنید. این امکان می‌دهد با نمودارها تعامل کرده و جزئیات را بررسی کنیم.

[1] https://plotly.com/

- پشتیبانی از متن تعبیه شده و علائم: Plotly امکان اضافه کردن متن تعبیه شده به نمودارها و گراف‌ها را فراهم می‌کند.
- پشتیبانی از متغیرهای چندگانه: شما می‌توانید چندین متغیر را در یک نمودار یا گراف نمایش دهید و این امکان را می دهد که روابط میان داده‌ها را بررسی کنیم.
- گراف‌های متنوع: Plotly انواع مختلفی از نمودارها و گراف‌ها را پشتیبانی می‌کند، از جمله نمودارهای خطی، نمودارهای نقطه‌ای، نمودارهای میله‌ای، نمودارهای دایره‌ای و...
- پشتیبانی از داده‌های ساختار یافته: می توان داده‌های خود را به شکل فریم داده‌های Pandas به Plotly منتقل کرده و با داده‌های ساختار یافته کار کرد.
- تولید تصاویر و فایل‌های تصویری: می توان تصاویر با کیفیت بالا از نمودارها و گراف‌ها ایجاد کرده و آنها را به عنوان تصاویر یا PDF و PNG ذخیره نمود.

فصل سوم

جمع‌بندی و نتیجه‌گیری و پیشنهادات

بطور کلی، کتابخانه‌هایی که ارائه می‌شود، در انواع مختلفی از وظایف مرتبط با داده کاوی و تحلیل داده استفاده می‌شوند. در زیر یک براورد کلی از کاربردهای این کتابخانه‌ها آورده شده است:

1. NumPy:

- پیش‌پردازش داده: NumPy برای انجام عملیات ماتریسی، آماری و آرایه‌ای مورد استفاده قرار می‌گیرد.

2. SciPy:

- پیش‌پردازش داده: SciPy توابع و ابزارهای پیش‌پردازش داده برای علوم و مهندسی ارائه می‌دهد.

3. TensorFlow و PyTorch:

- پردازش داده و مدل‌سازی: این کتابخانه‌ها به منظور ساختن، آموزش و اجرای مدل‌های عمیق برای مسائل یادگیری ماشین و یادگیری عمیق مورد استفاده قرار می‌گیرند.

4. Keras:

- پردازش داده و مدل‌سازی: Keras به عنوان یک واسط برای TensorFlow و PyTorch استفاده می‌شود و سادگی و زیبایی در ساخت مدل‌های عمیق را فراهم می‌کند.

5. OpenCV:

- پردازش تصویر و ویدیو: OpenCV برای پردازش تصاویر، تشخیص الگو، تقویت تصویر و استخراج ویژگی‌های تصویری به کار می‌رود.

6. Plotly و Matplotlib:

- تصویرسازی داده: این کتابخانه‌ها برای ساخت نمودارها و گراف‌های تعاملی و زیبا جهت نمایش داده‌ها و الگوهای مختلف در داده‌ها استفاده می‌شوند. Plotly به ویژه در تصویرسازی تعاملی برجسته است.

7. Seaborn:

- تصویرسازی داده: Seaborn یک کتابخانه تصویرسازی برای داده‌های آماری و تجزیه و تحلیل داده است و به شما امکان می‌دهد نمودارها و گراف‌های آماری با استفاده از Matplotlib ایجاد کنید.

8. Pandas:

- پیش‌پردازش داده: Pandas برای مدیریت و تحلیل داده‌های جدولی (داده‌های ساختار یافته)، شامل تفکیک داده، تمیزکاری، ترکیب داده‌ها و محاسبات آماری مورد استفاده قرار می‌گیرد.

9. Scikit-Learn

- مدل‌سازی داده: این کتابخانه ابزارهای یادگیری ماشین، کلاسیفیکیشن، رگرسیون، خوشه‌بندی و ارزیابی مدل‌ها را ارائه می‌دهد و در تجزیه و تحلیل داده‌ها و پیش‌بینی بسیار مفید است.

همچنین باید توجه داشت که در پروژه‌های داده کاوی، معمولاً ترکیبی از این کتابخانه‌ها و ابزارها به کار گرفته می‌شود تا وظایف مختلفی از جمله پیش‌پردازش داده، تصویرسازی، مدل‌سازی، و ارزیابی را انجام دهید. انتخاب کتابخانه‌ها بستگی به نیازهای خاص پروژه و تسلط شما به این کتابخانه‌ها دارد.

منابع و مراجع

[1] https://hamruyesh.com/top-10-python-libraries-for-data-science-for-2021-guide/

[2] https://plotly.com/

[3] https://pytorch.org/

[4] https://www.tensorflow.org/

[5] https://matplotlib.org/

[6] https://scikit-learn.org/stable/

[7] https://scipy.org/

[8] https://numpy.org/

[9] https://keras.io/

[10] https://pandas.pydata.org/

[11] https://seaborn.pydata.org/

Abstract

In this report, we examine the Python language libraries that are used in data mining projects. Using these libraries, you can advance the steps of a project like pre-processing with high accuracy and speed. In each section, the general aspects of the library, how to install and constituent functions and how to use it are briefly mentioned.

Key Words: Python, library, data mining, preprocessing

Amirkabir University of Technology
(Tehran Polytechnic)

Computer Science

Project 1

By
Samin Mahdipour

Supervisor
Dr.Ghatee

October 2023