کارگاه آموزشی مرتب‌سازی داده‌ها

دانشکده‌ی مهندسی برق و کامپیوتر دانشگاه تهران

یادگیری ماشین - پاییز ۹۹

مقدمه

داده‌ها ممکن است همیشه به شکل جدول‌های مرتب با توضیحات دقیق و برچسب‌های تعریف‌شده برای ما در دسترس نباشند. گاهی لازم است که برای مسائلی که با آن‌ها مواجه هستیم داده جمع‌آوری کنیم، آن را در ساختارهایی که مطلوبمان است مرتب‌سازی کنیم و بعد آن‌ها را جهت تحلیل و استخراج دیدگاه‌های جدید با روش‌های مختلف مورد بررسی قرار دهیم.
در سایر بخش‌های درس یادگیری ماشین همیشه فرض کرده‌ایم که داده‌ها در ساختار مطلوب و با اطلاعات دقیق در دسترس ما هستند و ما پیاده‌سازی الگوریتم‌های متنوع یادگیری ماشین را روی آن‌ها تمرین کرده‌ایم. اما موضوع این کارگاه یک مرحله قبل‌تر از آن است. در این کارگاه قرار است به موضوعات زیر بپردازیم:

1. منابع مختلف داده چه هستند و چطور می‌توان از آن‌ها داده استخراج کرد؟
2. چه هزینه‌هایی باید برای جمع‌آوری داده بپردازیم و به چه نکاتی در این میان باید توجه کنیم؟
3. یک پروژه‌ی جمع‌آوری داده چه بخش‌هایی دارد و چه مراحلی در آن طی می‌شود؟
4. چگونه داده را مرتب کنیم که در مراحل بعدی منجر به استخراج یک دیدگاه جدید در مورد مسئله‌مان شود؟
5. چطور بررسی کنیم که آیا استخراج داده را درست انجام داده‌ایم یا خیر؟
6. و در آخر این‌که چطور داده‌های استخراج شده را به سایر لایه‌ها و افراد موجود در پروژه ارائه دهیم؟

قبل از این‌که ادامه‌ی این فایل را بخوانید، لطفا ابتدا ویدیویی که از طرف آقای دکتر امیرحسین شیرازی در اختیار شما قرار گرفته است (این لینک) را مشاهده کنید.

در ادامه‌ی این گزارش، ابتدا توضیحاتی در خصوص داده‌هایی که از سایت ورزش ۳ استخراج شده توضیح داده می‌شود و پس از آن چند لینک در خصوص این‌که چگونه باید از صفحات وب (html) داده استخراج کرد در اختیار شما قرار گرفته است. در آخر در خصوص آن‌چه که شما باید به عنوان گزارش این کارگاه تحویل دهید نکاتی بیان گردیده است.

داده‌های سایت ورزش۳

سایت ورزش۳ یک وبگاه خبری ورزشی است که در سال ۱۳۸۹ آغاز به کار کرد این سایت شامل: خبرهای ورزشی، رده‌بندی لیگ‌ها و تورنمنت‌های پرهوادار، پخش زنده و نتایج زنده بازی‌ها، ویدئوهای ورزشی، روزنامه‌های ورزشی، تصاویر بازیکنان و پیش‌بینی و... است. ورزش۳، چهارمین وبگاه پربازدید در ایران و پربازدیدترین وبگاه ورزشی ایرانی است. [برداشت از ویکی‌پدیای فارسی] ورزش۳ داده و آرشیوهای ارزشمندی از تاریخ فوتبال ایران و جهان دارد اما این داده‌ها ساختارمند نیستند. به همین دلیل داده‌های این سایت برای این کارگاه انتخاب شد.

داده‌هایی که از ورزش ۳ شما در این کارگاه در اختیار دارید:

- **جدول جام حذفی**
  - **Year** سال برگزاری جام حذفی در ایران
  - **Stage** استیج یا مرحله مانند یک شانزدهم نهایی
  - **html** صفحه بازی مورد نظر در مرحله مشخص شده مانند بازی ماشین‌سازی‌تبریز - پرسپولیس
    - در این صفحه اطلاعات مهمی از بازی مورد نظر می‌توانید استخراج کنید. دیتاهای متنی ارزشمندی نیز در این صفحه قرار دارد.
- **تیم‌ها**
  - **Competition** رقابت‌ها یا لیگ‌ها مانند آرشیو جدول های لیگ برتر ایران
  - **team** تیم مورد نظر
  - **html** مربوط به خود تیم مانند گل-گهرسیرجان
    - از این صفحه می‌توانید ویژگی‌های مهمی از تیم و تاریخچه به دست بیاورید. اطلاعات متنی ارزشمندی مانند خبرها نیز در این صفحه قرار دارد.
- **جدول‌ها**
  - **Competition** رقابت‌ها یا لیگ
  - **Year** سال برگزاری
  - **html** صفحه جدول رقابت مانند جدول لیگ برتر (99-00) - لیگ برتر ایران
    - در این صفحه اطلاعاتی از جایگاه تیم‌ها، امتیازها و ... در لیگ‌های مختلف وجود دارد.
- **نقل و انتقالات**
  - **transfers** نقل و انتقالات مربوط به یک فصل یا لیگ
  - **html** صفحه نقل و انتقالات مانند نقل و انتقالات لیگ برتر 99-00
    - در این صفحه می‌توانید ویژگی‌های ارزشمندی از وضعیت بازیکنان و تیم‌ها در هر فصل به دست بیاورید.
- **جام جهانی**
  - **Year** سال برگزاری
  - **Stage** استیج مانند یک هشتم نهایی
  - **html** صفحه بازی مورد نظر مانند بازی برزیل - شیلی
    - در این صفحه اطلاعات مهمی از بازی مورد نظر می‌توانید استخراج کنید. دیتاهای متنی ارزشمندی نیز در این صفحه قرار دارد.

چگونه از فایل‌های html داده استخراج کنیم؟

- Web Scraping with Python Using Beautiful Soup
- How To Scrape Web Pages with Beautiful Soup and Python 3
- Tidy Data

کارهایی که شما باید انجام دهید

فرض کنید که شما عضوی از محصولی هستید که قرار است بر اساس داده‌های گذشته‌ی تورنومنت‌های مختلف فوتبالی، قهرمان لیگ هر کشور را مشخص کند. یک تیم پیش از شما روی این موضوع وقت گذاشته است که داده‌ها را از سایت ورزش۳ استخراج کند و آن‌ها را به صورت فایل‌های html در اختیار شما قرار داده است. حال وظیفه‌ی تیم شما این است که این فایل‌ها را مرتب‌سازی کنید و به شکل جداولی با توضیحات مشخص به تیم بعدی تحویل دهید که بتوانند از آن برای پیش‌بینی قهرمان با استفاده از الگوریتم‌های یادگیری ماشین استفاده کنند. همچنین شما باید گزارشی آماده کنید که نشان دهد چرا اطلاعاتی که ارائه داده‌اید قابل اعتماد هستند و می‌توان روی آن‌ها تحلیل انجام داد. توجه داشته باشید که رویکرد شما در پاسخ به این سؤالات مهم است.

شما باید سه دسته فایل آماده کنید: ۱. گزارش کارگاه ۲. توضیحات داده‌های استخراج شده برای تیم بعدی ۳. داده‌های استخراج شده

سؤالات گزارش کارگاه

1. برای پیش‌بینی قهرمان یک لیگ، از داده‌هایی که تیم قبلی به شما داده است چه اطلاعاتی را می‌توانیم به دست آوریم؟ شما تصمیم گرفته‌اید که چه ویژگی‌هایی را استخراج کنید و چه داده‌هایی را کنار بگذارید؟
2. هزینه‌ی استخراج هر کدام از ویژگی‌هایی که به آن‌ها در سؤال قبل پاسخ داده‌اید چقدر است؟
3. چه توجیهی برای این هزینه دارید و در آینده چه کمکی می‌تواند به ما بکند؟
4. چگونه می‌سنجید که داده‌های استخراجی شما قابل اعتماد هستند؟ با توجه به ویدیوی آموزشی فرآیند کیفیت‌سنجی را چطور انجام داده‌اید؟ چه متریک‌هایی تعریف کرده‌اید؟
5. به نظر شما چطور می‌توان این پروژه را در ابعاد بزرگ‌تر و در یک تیم انجام داد؟ چه وظایفی و چه فرآیندهای کاری نیاز است برای مرتب‌سازی داده‌ها تعریف شود؟
6. نتایج حاصل از فعالیت‌های تیم خود را چطور به تیم بعدی ارائه می‌کنید؟ تصمیم‌های خود را توضیح دهید.

گزارش خروجی

این گزارش حاصل کار شما در خصوص کار با داده است. شما قرار است آن را به تیم بعدی تحویل دهید بنابراین انتظار می‌رود که در آن توضیح دهید داده‌هایی که استخراج کرده‌اید چه هستند، به چه کار می‌آیند و تحلیل کنید که چرا قابل اعتماد هستند.

داده‌های استخراج شده

مطابق با گزارش خروجی، کدها و نوت‌بوک‌های خود را در این بخش تحویل دهید و همچنین جداولی یا هر قالب دیگری از داده که استخراج کرده‌اید را در این بخش آپلود کنید.

نکات پایانی

- درک مسائل و رویکردهای آموزش داده شده مهم است. بیش از حد زمان نگذارید و سعی کنید چیزهایی که از شما خواسته شده است را پاسخ دهید.
- ممکن است کار با داده‌های html در ابتدای کار برای شما دشوار باشد. حتما اگر در این خصوص یا مباحث دیگر ابهام داشتید با دستیاران آموزشی مطرح کنید و از آن‌ها کمک بگیرید.
- نمره‌ی به دست آمده از این کارگاه جزء بخش‌های اصلی درس نیست و به عنوان نمره‌ی امتیازی در درس محاسبه خواهد شد.
- حتما ویدیوی پایانی کارگاه را که به عنوان جمع‌بندی در اختیار شما قرار خواهد گرفت مشاهده کنید.