

بخش اول قسمت اول. انتخاب مجموعه داده و توضیحات کامل در مورد آن

مجموعه داده سیستم اشتراک گذاری دوچرخه

پس زمینه

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

سیستم اشتراک دوچرخه نسل جدیدی از سیستم سنتی اجاره دوچرخه است جایی که در آن کلیه روند های عضوگیری، اجاره، بازگشت اتوماتیک دوچرخه صورت می گیرد. کاربر از طریق این سیستم قادر است که به راحتی یک دوچرخه را از یک مکان خاص اجاره و به مکان دیگر برگرداند. در حال حاضر، حدود 500 سیستم اشتراک دوچرخه در سراسر جهان وجود دارد که مجموعاً از 500 هزار دوچرخه تشکیل شده است. امروزه، به این سیستم ها علاقه زیادی نشان داده می شود به این دلیل که نقش مهمی در ترافیک و زیست محیط و موضوع های سلامت دارند.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

جدا از کاربرد های این سیستم، مشخصه های این داده های تولید شده برای تحقیقات قابل تامل است. اطلاعات مدت زمان سفر، تاریخ خروج، مکان نهایی به طور صریح در این سیستم های حمل و نقلی دیگر مانند اتوبوس، مترو رکورد و ضبط می شود. این ویژگی ها سیستم اشتراک دوچرخه را به شبکه حسگر مجازی تبدیل می کند که می تواند برای سنجش تحرک در شهر استفاده شود. از این رو، انتظار می رود بسیاری از رویدادهای مهم در شهر توسط این سیستم نظارت بر این داده ها تشخیص داده شوند.

مجموعه داده

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in <http://capitalbikeshare.com/system-data>. We aggregated the data on two hourly and daily basis and then extracted and added the corresponding weather and seasonal information. Weather information are extracted from <http://www.freemeteo.com>.

فرآیندهای سیستم اجاره دوچرخه بسیار با زمینه فصلی و محیطی مرتبط است، برای نمونه شرایط آب و هوایی، بارش، روز هفته، فصل، ساعتی از روز و ... می تواند در رفتار اجاره تاثیرگذار باشد. هسته مجموعه داده به لاگ سال های 2011 و 2012 توسط یک سیستم اجاره دوچرخه در واشنگتن دی سی آمریکا مربوط می شود و در لینکی که در بالا قرار داده شده اطلاعات بیشتر موجود است. (ناشر دیتاست) داده ها را بر اساس دو دسته ساعتی و روزانه جمع آوری کرده و سپس داده های مربوط به اطلاعات آب و هوا و فصلی را از لینکی که در بالا استخراج و اضافه شده است

شناسایی هدف

- Regression :

رگرسیون (کار مربوطه این تکلیف) :

Predication of bike rental count hourly or daily based on the environmental and seasonal settings.

پیش بینی شمار اجاره دوچرخه ها به صورت روزانه یا ساعتی براساس زمینه فصلی یا محیطی

- Event and Anomaly Detection :

تشخیص رویدادها :

Count of rented bikes are also correlated to some events in the town which easily are traceable via search engines. For instance, query like "2012-10-30 washington d.c." in Google returns related results to Hurricane Sandy.

شمار دوچرخه های اجاره داده شده مربوط می باشد به یکسری رویدادهایی در شهر هستند که به راحتی توسط موتور جستجو قابل ردیابی هستند. در آخر یک نمونه آورده شده است که ارتباط این موضوع را روشن می کند (این موضوع ارتباطی به کار ما ندارد).

فایلها

- **hour.csv** : bike sharing counts aggregated on hourly basis. Records: 17379 hours

شمار دوچرخه های به اشتراک گذاشته شده بطور ساعتی جمع آوری شده است. رکورد ها : 17379 ساعت

- **day.csv** : bike sharing counts aggregated on daily basis. Records: 731 days

✓ شمار دوچرخه های به اشتراک گذاشته شده بطور روزانه جمع آوری شده است. رکورد ها : 731 روز

مشخصه های مجموعه داده

Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

هر دو فایل *روز و *ساعت دارای فیلد های زیر هستند به استثنا فیلد ساعت که در فایل *روز موجود نیست.

Feature name	Feature description
- season : season (1:springer, 2:summer, 3:fall, 4:winter)	فصل : (1: بهار، 2: تابستان، 3: پاییز، 4: زمستان)
- yr : year (0: 2011, 1:2012)	سال : (0: 2011، 1: 2012)
- mnth : month (1 to 12)	ماه : (1 تا 12)
- hr : hour (0 to 23)	ساعت : (0 تا 23)
- holiday : weather day is holiday or not	روز تعطیل : تعطیلی آب هوایی
- weekday : day of the week	روز هفته
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.	روزکاری : اگر روز تعطیل یا آخر هفته نیست 1 در غیر این صورت 0
+ weathersit : - 1: Clear, Few clouds, Partly cloudy - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog	+ وضعیت آب و هوا : - 1: صاف، کمی ابری ، نیمه ابری - 2: غبار + ابری، غبار + ابرهای پاره ای، غبار + نیمه ابری، غباری - 3: برف سبک، باران سبک + رعد و برق + ابرهای پراکنده، باران سبک + ابر پراکنده - 4: باران سنگین + گلوله های یخ + رعد و برق + غبار، برف + مه

- temp : Normalized temperature in Celsius. The values are divided to 41 (max)	میزان دما : دما بر حسب سلسیوس نرمالایز شده، تمام مقادیر بر 41 (بیشترین دما) تقسیم شده اند
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)	میزان دما حس شده : دما حس شده بر حسب سلسیوس نرمالایز شده، تمام مقادیر بر 50 (بیشترین دما) تقسیم شده اند
- hum: Normalized humidity. The values are divided to 100 (max)	میزان رطوبت : رطوبت ها نرمالایز شده و تمام مقادیر بر 100 (بیشترین رطوبت) تقسیم شده اند
- windspeed: Normalized wind speed. The values are divided to 67 (max)	سرعت باد : سرعت باد نرمالایز شده و تمام مقادیر بر 67 (بیشترین سرعت باد) تقسیم شده اند
- casual: count of casual users	کابران معمولی (تصادفی یا ثبت نام نکردند) : تعداد کاربران معمولی
- registered: count of registered users	کابران ثبت نام شده : تعداد کاربران ثبت نام شده
- cnt: count of total rental bikes including both casual and registered	تعداد کل : شمار مجموع دوچرخه های اجاره داده شد شامل کابران معمولی و ثبت نام شده