Tipología y ciclo de vida de los datos – PRÁCTICA 1

Diego Contreras Jiménez

Dataset

Clasificación, probabilidad y presencia mediática de La Liga.

Descripción

El conjunto de datos generado como parte de esta actividad contiene la clasificación actual de la liga profesional de fútbol española “LaLiga”, las probabilidades en los resultados de la próxima jornada y un indicador de la presencia mediática basado en los tweets de la última semana.

Imagen

Contexto

LaLiga es la máxima categoría masculina de clubs de futbol en España. Requiere poca presentación pues es conocida a nivel mundial, tiene millones de seguidores y mueve una cantidad muy elevada de dinero.

La clasificación habitual que se presenta jornada tras jornada es ya bien conocida por los aficionados, incluye el órden de posición, total de partidos, victorias, empates, derrotas, puntos, goles a favor y en contra. A esta se le pretende enriquecer con otro tipo de información que viene de distintos sitios:

- Casas de apuestas

  Desde hace muchos años exiten loterías y casas de apuestas que permiten apostar dinero basándose en los resultados de los partidos de LaLiga.

  En este caso incorporaremos la información que provee una casa de apuestas según sea la cuota de pago por el acierto del resultado de un partido.

- Twitter

  Twitter es una de las redes sociales más importantes en este momento a nivel mundial y permite la consulta y estudio de su información a través de su API pública.

Gracias a esta plataforma podremos incorporar a nuestros datos un indicador de presencia mediática en la última semana obtenido a partir de los hashtags de #LaLiga y del propio equipo.

Contenido

Campos:

- Posicion: Posición en la clasificación.
- Equipo: Nombre del equipo.
- Partidos: Número de partidos jugados.
- Victorias: Número de partidos ganados.
- Empates: Número de partidos empatados.
- Derrotas: Número de partidos perdidos.
- Puntos: Número de partidos ganados.
- GolesAFavor: Número de goles a favor.
- GolesEnContra: Número de goles en contra.
- Diferencia: Diferencia entre goles a favor y goles en contra.
- Hashtag: Hashtag usado en Twitter.
- Gana: Probabilidad de ganar el próximo partido.
- Empata: Probabilidad de empatar el próximo partido.
- Pierde: Probabilidad de perderel próximo partido.
- Tweets: Marcador que muestra la actualidad mediática en la última semana

Obtención:

La obtención de estos se ha realizado en tres sitios web mediante métodos o técnicas diferentes:

- API: http://api.football-data.org
- Scraping: https://www.predictz.com/predictions/spain/primera-liga/
- Twitter: https://twitter.com/

Estos datos pertenecen al estado actual de los equipos de primera división en LaLiga, y por lo tanto deben ser actualizados en cada nueva jornada, siendo esta su propia vigencia.

Agradecimientos

El propietario de este conjunto de datos es Diego Contreras Jiménez.

Inspiración

Este conjunto de datos es interesante porque a la clásica clasificación de LaLiga le aporta información extra, como son las probabilidades del resultado en el próximo partido de casas de apuestas y la relevancia mediática que tiene cada equipo en Twitter en la actualidad.

Estos datos pueden ser usados para diversos fines, tanto para tener actualizada la evolución del estado de La Liga o para hacer minería de datos y así obtener predicccciones que puedan ser interesantes.

Tipología y ciclo de vida de los datos – PRÁCTICA 1

Diego Contreras Jiménez

Licencia

La licencia escogida para la publicación de este conjunto de datos ha sido:

CC0 Public Domain License

Sin derechos de propiedad intelectual

> *La persona que ha asociado una obra a éste documento ha dedicado la misma al dominio publico, liberándola de forma mundial y en la medida que lo permita la ley, de todos sus derechos de propiedad intelectual, incluyendo todos los derechos conexos.*
>
> *Puede copiar, modificar, distribuir la obra y hacer comunicación pública, incluso para fines comerciales, sin pedir permiso.*

El motivo de seleccionar esta licencia para el dataset se debe a que han sido obtenidos para una práctica de la asignatura **Tipología y ciclo de vida de los datos** de la Universidad UOC, y considero que esta licencia permite y promueve a otros estudiantes e investigadores hacer uso completo y sin restricciones de estos para fines educativos.

Enlaces

Código

https://github.com/dieconji/laliga/tree/master/src

Dataset

https://github.com/dieconji/laliga/tree/master/dataset