**INSTITUTO TECNOLÓGICO DE COSTA RICA**

**INGENIERÍA EN COMPUTACIÓN**

**Introducción al desarrollo de aplicaciones web**

**Investigación Web Scrapping**

**Efren Jiménez Delgado.**

**ELADIO CRUZ VÁSQUEZ**

**Mayo, 2017**

**Tabla de contenido**

[1. Resumen 3](#_Toc482647515)

[2. Introducción 3](#_Toc482647516)

[3. Resultados 3](#_Toc482647517)

[4. Conclusiones 5](#_Toc482647518)

[5. Recomendaciones 5](#_Toc482647519)

[6. Bibliografía 5](#_Toc482647520)

# 1. Resumen

Esta tercera etapa del proyecto sobre Web Scrapping se tratará de explicar acerca de los resultados obtenidos sobre las páginas que se habían seleccionado anteriormente.

Para realizar esta actividad se cuentan con diferentes herramientas que se utilizaron como librerías y lenguajes de programación que se utilizaron para este trabajo de investigación tales como Postgress, Python, entre otras.

# 2. Introducción

En esta investigación, se estará hablando acerca de los resultados que se obtuvieron en el Web Scraping, de las páginas que cada uno seleccionó.

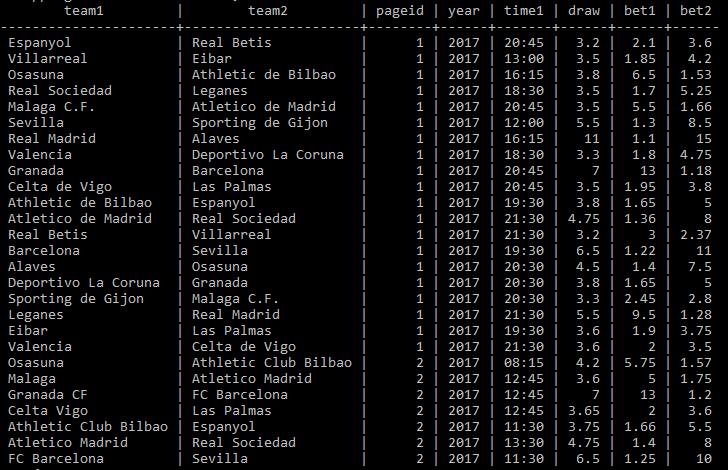
En el enunciado anterior se comentó que la principal función de este proyecto es la extraer información de diferentes sitios web con el fin de contrastarla y así obtener algún beneficio que, haciendo manualmente esta actividad sería más complicado.

Entonces, en esta etapa final se comentará los datos que se pudieron obtener, así como las plataformas de hosting que se emplearon en la elaboración del proyecto.

# 3. Resultados

Para realizar esta última etapa de la investigación, se contó con una base de datos Postgress que se logró hostear en la plataforma de Heroku, ya que es gratis y permite crear varias conexiones simultaneas, así como gran cantidad de datos.

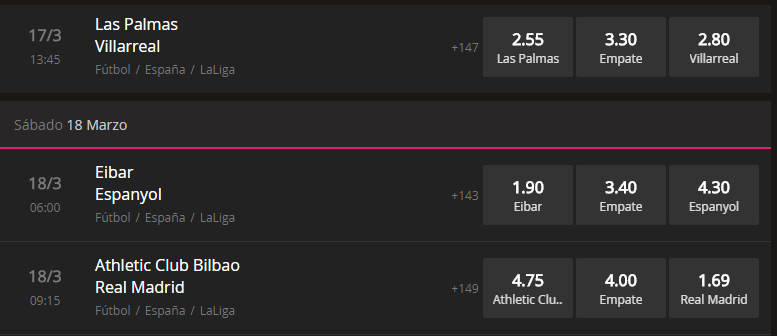
Los resultados obtenidos fueron que se pudo realizar una correcta extracción de los datos y como se explicó en el enunciado anterior se almacenaron de la siguiente manera



Además, se tuvo que crear una nueva tabla con datos de auditoria, que al igual que la anterior también se logra almacenar correctamente la información.

Finalmente, se adjuntan las imágenes del enunciado pasado donde se logra observar como estaba distribuida la información.





# 4. Conclusiones

Para concluir se puede mencionar lo mismo que el documento anterior:

* Se debe indagar anteriormente a realizar el trabajo, los datos sobre la legalidad que tenga los sitios web, esto con el fin de no llegar a cometer algún delito grave.
* Se debería de enseñar esta práctica, ya que se puede lograr grandes beneficios al aplicar correctamente esta técnica.

# 5. Recomendaciones

* Se debe de investigar los términos legales de los sitios web para no cometer ningún delito.
* Se debería de indagar los mejores lenguajes de programación que permitan desarrollar una herramienta útil.

# 6. Bibliografía

Carey, H. J., & Manic, M. (2016). HTML web content extraction using paragraph tags. *2016 IEEE 25th International Symposium on Industrial Electronics (ISIE)*, 1099-1105.

Mahto, D. K., & Singh, L. (2016). A dive into Web Scraper world. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 689-693.

Malik, S. K., & Rizvi, S. (2011). Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation. *2011 International Conference on Computational Intelligence and Communication Networks*, 465-469.