Resumen Estadistico de alumnos matriculados en la Universidad Nacional de San Agustin a traves de un asistente de voz, usando Python.

Edsel Yael, Alvàn Ventura
Universidad Nacional de San Agustin
ealvan@unsa.edu.pe

Resumen

A lo largo de los años, se ha demostrado lo importante que es manejar grandes masas de datos, lo cual nos refiere a este articulo, en el que manejaremos datos publicos obtenidos de la Universidad Nacional de San Agustin a traves de un asistente de voz, usando Python, y tecnicas de Web Scrapping, para ese fin, con el objetivo de dar graficas estadisticas mediantes la voz, para luego interpretarlas y ver las possible explicaciones de los datos que estamos viendo.

Palabras clave—Graficas estadisitcas; Web Scrapping; Python; obtencion de datos.

I. INTRODUCTION (*Heading 1*)

En esta epoca de pandemia, los datos y sus predicciones son mas importantes que nunca para poder predecir y estimar el futuro a traves de la estadistica y ver que se puede mejorar. En este articulo se aplicara software Estadisitco para ver datos generals con respect a los alumnus amtriculados en el año 2020.

Muchos expertos en Pensamiento Computacional, afirman que hacer proyectos, que en si mismos son hechos por ser de utilidad y mejorar las capacidades de la programacion, ayudan a mejorar la habilidad del siglo XXI, segun Janet Wing, el Pensamiento Computacional.

II. MARCO TEÓRICO

En este articulo se vera el concepto de Web Scrapping, para la obtencion de datos de la pagina web (http://extranet.unsa.edu.pe/sisacad/visualiza_fechas_b.php), en esta pagina se vera como obtener los datos de todos los alumnos de todas las escuelas de la Universidad Nacional de San Agustin.

El "Web Scrapping" en español "Raspado Web", es una tecnica utilizada para obtener datos. Segun (*What Is Scraping | About Price & Web Scraping Tools | Imperva*, n.d.) el "web scraping" es el proceso de usar bots para extraer contenido y datos de un sitio web. A diferencia del screen scraping, que solo copia los píxeles que se muestran en pantalla, el web scraping extrae el código HTML subyacente y, con él, los datos almacenados en una base de datos. El raspador puede replicar el contenido completo del sitio web en otro lugar.

Ademas de esta herramienta, utilizaremos el software Plotly Dash, segun (*Plotly Dash: A Beginner's Guide to Building an Analytics Dashboard | by Saurabh Kothari | Analytics Vidhya | Medium | Analytics Vidhya*, n.d.), es un framework de Python de código abierto que se utiliza para crear aplicaciones web interactivas de visualización de datos. Está desarrollado por el equipo de plotly y fue lanzado a mediados de 2017. Está construido sobre Flask, Plotly.js, React.js.

Tambien se uso archivos json para el almacenamiento de la informacion, segun (*JSON*, n.d.), JSON es un formato de texto que es completamente independiente del lenguaje, pero utiliza convenciones que son familiares para los programadores de la familia de lenguajes C, incluidos C, C ++, C #, Java, JavaScript, Perl, Python y muchos otros. Ademas de archivos csv, estos son segun la pagina (*What Is a CSV File, and How Do I Open It?*, n.d.), un archivo de valores separados por comas (CSV), un archivo de texto sin formato que contiene una lista de datos.

Se utilizara un marco de datos, para el manejo de la informacion, para esto se utilizara el modulo pandas, segun (*Pandas - Python Data Analysis Library*, n.d.) es una herramienta de análisis y manipulación de datos de código abierto, rápida, potente, flexible y fácil de usar, construido sobre el lenguaje de programación Python.

Tambien se implementara un modo de reconocimiento de voz, el cual se hara mediante speech_recognition, el cual. Segun (Roig-Vila & Moreno-Isac, 2020) (*The Ultimate Guide To Speech Recognition With Python – Real Python*, n.d.), hace la conversion de sonido físico a una señal eléctrica con un micrófono y luego a datos digitales con un convertidor de analógico a digital. Una vez digitalizados, se pueden utilizar varios modelos para transcribir el audio a texto. Esto es basicamente lo que se hara para lograr un asistente de voz.

III. METODOLOGÍA

Primero lo que se hizo fue el "Web Scrapping", con las herramientas de BeartifulSoup y requests, que son modulos de python que nos ayudan a este fin, este nos sirvio para obtener los datos publicos de los alumnos de la Universidad Nacional de San Agustin, luego de obtenerlos mediante una peticion al servidor que los almacena, Se alojo en una archivo json toda esa informacion, la cual se proceso en archivos csv, para procesarlos con el modulo de python pandas, creando asi un

marco de datos en ingles “data frame”, que luego sera procesado por el modulo Plotly, para luego darnos las graficas estadisticas interactivas por voz, luego de hacer esto con los datos perninetes de nuestra base de datos de los alumnus matriculados en el año 2020, se procedio a implementar commandos de voz para lograr controlar que estadisiticas deben mostrarse, para lograr esto usamos “expresiones regulares”, las cuales nos permitiran hacer coincidir los commando de voz con la voz de una persona, con este proposito nosotros usamos el modulo speech recognition, el cual nos ayuda a traducir a traves de un microfono la voz de una persona.

IV. Analisis

Los datos que obtuvimos tras ver los graficos estadisticos, de la base de datos de los alumnus matriculados en el año 2020 fueron:

1) La Carrera con mas estudiantes:
La Carrera con mas estudiantes de la Universidad Nacional de San Agustin, es la escuela professional de Educacion, superando a la Carrera de Administracion en 241 estudiantes, es la Carrera con mayor estudiantes de la Universidad, su cantidad es de 1457 alumnos matriculados en el año 2020. Cabe destacar que educacion es la que mas vacantes tiene por proceso de admission, lo cual explica en parte el gran nuemro de estudiantes de esta Carrera. La escuela professional con menos estudiantes de la Universidad Nacional de San Agutin(UNSA), es la escuela professional de Filosofia, esta Carrera tiene 120 estudiantes matriculados este año, la penultima escuela professional con menos estudiantes es la Ingenieria Pesquera, la cual tiene 46 estudiantes mas que la escuela professional de Filosofia. La explicacion aproximada a esto es que no hay muchos interesados en la Carrera universitaria de Filosofia.
En la Figura(1) se muestra la grafica representada por el numero de estudiantes por escuela professional.

2) Distribucion por Grupos:
La distribucion por grupos en la UNSA, es por el orden de merito, primordialemnte los del grupo 1, tienen la opcion de elegir primero sus horarios, luego los alumnus del grupo 2, pueden elegir sus horarios, a partir de una fecha publicada en la pagina http://matricula.unsa.edu.pe/ en el apartado de Cronograma de Matriculas, luego los del grupo 3, son los ultimos en matricularse, mayormente en este punto algunos cursos llegaron a su limite de alumnus permitido por la Universidad, excepcionalmente se pueden crear nuevos grupos para dicho curso.
Como pueden ver la distribucion es equitativa, cada vez que se asignan grupos a los alumnus, se asignan dividiendo al total de alumnos en tres grupos de igual tamaños, siendo asi, que el grupo 1 tiene a un total de 7761 alumnos matriculados, en el grupo 2, se encuentran 7753 alumnos, y por ultimo, en el grupo 3 estan 7751 alumnos, esto hace un total de 23265 alumnos matriculados en total en el año 2020.
En la Figura(2), se muestra la distribucion por grupos con respect al numero de alumnos.

3) Numero de alumnos por Año de Ingreso:
En este año 2020, se registro 28 alumnos matriculados que ingresaron en el año 2020, comparado con la cantidad de alumnus matriculados en el año 2019, que son un total de 75 alumnos, hay varias razones para esto,una es la pandemia, debido a que redujo los ingresos economicos de la mayoria de la poblacion en el Peru, otra es que todavia alumnos estan esperando a matricularse el 2021, pero ingresaron en el año 2020, lo que explica en parte la cantidad de alumnos que ingresaron en el 2020.

En la Figura(3), se muestra el grafico de barras por año de ingreso.

4) El nombre mas comun:
El nombre mas comun de entre los alumnos matriculados en el año 2020, es “Milagros”, con un total de 320 estudiantes que llevan ese nombre, mientras que el segundo mas comun es “Fernando”, el cual es usado por 254 alumnos.
Le sigue nombres como “Jesus”, ”Luis”, ”Antonio”, ”Angel”,”Carlos”, “Alberto”, “Alexander”, “Alonso”, etc. Estos nombres estan proximos a 200 estudiantes.
En la Figura(4), se muestra la distribucion de nombres, por el numero de alumnos.

5) El numero de Varones y Mujeres:
Debido a nuestra limitada informacion publica de los alumnos matriculados en el año 2020 de la UNSA, se procedio a estimar cuantos alumnos de genero Maculino y Genero Femenino hay en la Universidad, para esto se procedio a usar el modulo gender_guesser, el cual nos permite saber si el nombre es de un estudiante Varon, o de una Mujer. Claramente va a ver errores, pues hay nombres que se usan tanto para varones como para mujeres, por lo que cuando ocurra esto se declara en la grafica con la cateforia “Ambos generos”, sino se reconoce como “Masculino” , “Femenino” o “Ambos generos”, entonces si categoria sera “No determinado”, pues es un nombre que la inteligencia artificial no reconoce.
En la Figura(5), se muestra la distribucion de nombres identificados como varones, mujeres, ambos generos y genero indeterminado.

V. Resultados

Como se vio, en los datos mostrados, se puede obtener datos utiles a traves de lso datos porporcionados de la Universidad Nacional de San Agustin, lo cual nos muestra el estado de las carreras universitarias en esta Universidad, se puede analizar algunos datos interesantes sobre los alumnos. Tambien se puede aprender mucho en el aprendizaje de la programacion con Proyectos como este, el cual nos ayuda ver que recursos que tenemos en la computadora, como el microfono, internet, etc. Que nos ayuda en el aprendizaje de como funcionan los perifericos en general.
Sin bien se capturo los datos de un servidor de una universidad, animo hacerlo con datos publicos y de interes general, para luego sacar conclusiones utiles de muchos otros sitios webs.

VI. TRABAJOS FUTUROS

Si bien se obtuvo datos de la url http://matricula.unsa.edu.pe/ , se pude hacer con muchas mas paginas web.

En un futuro se podria hacer con los ingresantes del proceso de admision de la UNSA en el año 2021, O con otras universidades, y sacar medidas mas estadisticas utiles para predecir.

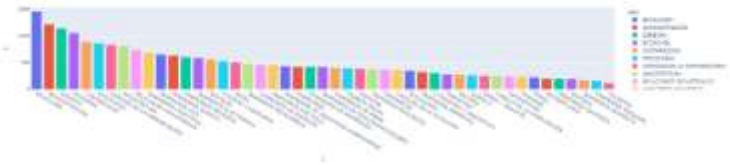

Figura (1)

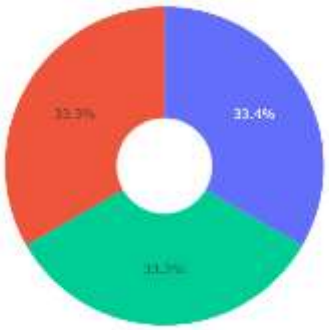

Figura (2)

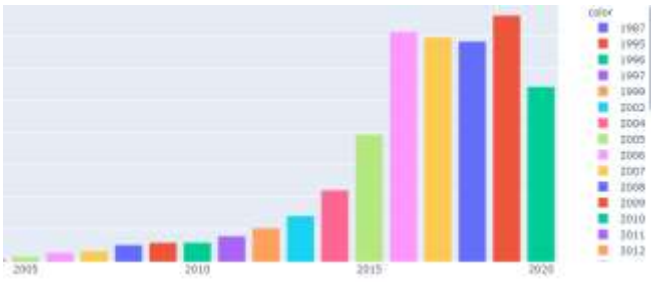

Figura (3)

Figura (4)

REFERENCES

1. *JSON*. (n.d.). Retrieved December 30, 2020, from https://www.json.org/json-en.html
2. *pandas - Python Data Analysis Library*. (n.d.). Retrieved December 30, 2020, from https://pandas.pydata.org/
3. *Plotly Dash: A beginner's guide to building an analytics dashboard | by Saurabh Kothari | Analytics Vidhya | Medium | Analytics Vidhya*. (n.d.). Retrieved December 30, 2020, from https://medium.com/analytics-vidhya/plotly-dash-a-beginners-guide-to-building-an-analytics-dashboard-cedf297e01f1
4. Roig-Vila, R., & Moreno-Isac, V. (2020). El pensamiento computacional en Educación. Análisis bibliométrico y temático. *Revista de Educación a Distancia (RED)*, *20*(63). https://doi.org/10.6018/red.402621
5. *The Ultimate Guide To Speech Recognition With Python – Real Python*. (n.d.). Retrieved December 30, 2020, from https://realpython.com/python-speech-recognition/
6. *What Is a CSV File, and How Do I Open It?* (n.d.). Retrieved December 30, 2020, from https://www.howtogeek.com/348960/what-is-a-csv-file-and-how-do-i-open-it/
7. *What Is Scraping | About Price & Web Scraping Tools | Imperva*. (n.d.). Retrieved December 30, 2020, from https://www.imperva.com/learn/application-security/web-scraping-attack/

ANEXOS

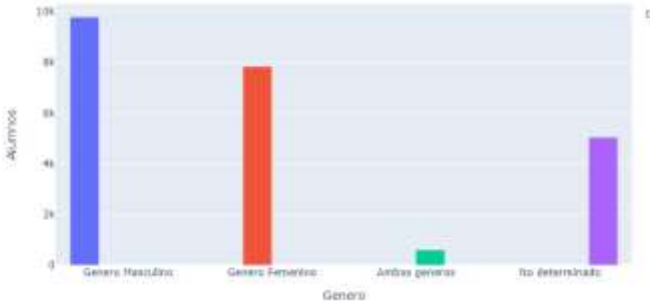
10k
8k
6k
4k
2k
0
Alumnos
Genero Masculino
Genero Femenino
Ambos generos
No determinado
Genero

Figura (5)

.