Big Data Course

Team Project Final Report

For students (instructor review required)

ⓒ2021 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung innovation Campus, you must receive written consent from copyright holder.

|  |
| --- |
| U-PICK Sistema de recomendación de Universidades Españolas |

27/07/23

CDC

Daniel Carmona Pedrajas

César Agulló Quirós

Christina Araujo Molano

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Simulated Scenario Description

2.2. Datasets Selection and Description

2.3. Data Ingestion Pipeline

2.4. Data Transformation Processing

2.5. Data Query and Insight

3. Results

3.1. Data Ingestion Scripts and Code

3.2. Data Transformation Scripts and Code

3.3. Description and Sample of Transformed Datasets

3.4. Data Visualization of Query Results

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

1.1. Background Information

El Sistema Universitario español, es uno en el que la primera elección de carrera es esencial: de esta depende el futuro del estudiante. Sin embargo, la tasa de estudiantes que abandonan su grado tras el primer año, ya sea debido a que las materias no se ajustan con sus objetivos e intereses, o porque la universidad no se alinea con sus valores. Esta realidad trae consigo una problemática, puesto que el hecho de que los estudiantes pierdan un año de estudios, no solo tiene consecuencias económicas (en la productividad potencial del país), sino que, de manera más directa, afecta a la autoestima de todos estos alumnos afectados.

1.2 Motivation and Objective

Nuestro objetivo con este proyecto es minimizar la ocurrencia de esta misma situación. Creemos que uno de los causantes de ello, es la falta de información del futuro estudiante, sea esta tanto cuantitativa como cualitativa. Por ello, nuestra propuesta se basa en un sistema que sea capaz de recoger, analizar y presentar toda esta información tan compleja, de naturaleza diversa en su origen y objetividad, para poder servir como herramienta al alumno, y así asegurar que se embarcarán en un camino con el que estarán satisfechos.

1.3 Members and Role Assignments

Data Engineer: Daniel Carmona

Data Analyst: César Agulló (Team Leader)

Data Scientist: Christina Araujo

1.4 Schedule and Milestones

Imagen que contiene Tabla

Descripción generada automáticamente

2. Project Execution

2.1 Simulated Scenario Description

Nuestro prototipo se probará con un estudiante que utiliza la herramienta para elegir universidad. De este modo, rellenará un cuestionario indicando sus preferencias y la importancia que le da a distintas cualidades en una institución. Estos datos se usarán para realizar una recomendación basada en los datos estáticos almacenados previamente en el sistemas, y proporcionar una lista de tres universidades con las que el ajuste es mayor, así como presentar la información relacionada (empleabilidad en sus grados de elección, la satisfacción estimada a partir de publicaciones en redes, notas de acceso, etc) . En este caso de prueba, el estudiante escoge, por ejemplo, los grados de Empresariales, Leyes e Historia como áreas de interés.

2.2. Datasets Selection and Description

Los datos utilizados proceden de distintas fuentes. Entre ellos contamos con datos en streaming, estructurados, y semiestructurados.

Datos en streaming:

El primer grupo de datos viene formado por los tweets. Estos contienen opiniones de universidades españolas, recogidas mediante la búsqueda en los hashtags de cada una de ellas. Puesto que actualmente la API de Twitter no se encuentra disponible para uso educativo de manera gratuita, estos tweets son simulados mediante un Script de Python y contienen los mismos campos que un tweet procedente de la API. Entre ellos, se encuentra un mensaje que conforma la parte no estructurada de nuestros datos. Estos se utilizarán para analizar las opiniones de cada una de las universidades, de modo que se pueda presentar al estudiante de manera concisa.

Datos estructurados:

Nuestros datos estructurados contienen datos estáticos sobre las universidades españolas. Estos no cambian, o tan solo lo hacen una vez al año (e.g. las notas de corte). Por ello, se extraen de diferentes documentos y páginas web para determinar las cualidades de cada universidad. Entre ellos se encuentran:

* Datos de localización extraídos usando la API de Google Maps
* Los datos básicos relacionados con cada grado, de QEDU (de estos forman parte la nota de corte, facultad, universidad, empleabilidad… de cada grado)
* Posiciones basadas en el H-Index de cada universidad, así como el número de investigadores asociados a cada campo, procedente de AD Scientific Index

2.3. Data Ingestion Pipeline

Los datos, una vez obtenidos, se encuentran en dos lugares. Los datos estructurados, se encuentran en csv que han de ser introducidos y tratados para su uso en la base de datos. De este modo, utilizando MariaDB, y por medio del conector de MySQL-Python, estos datos son introducidos en la base de datos de acuerdo al siguiente diagrama UML, de modo que se encuentren en un formato útil, y estén organizados para las futuras consultas que se harán sobre ellas.

Diagrama

Descripción generada automáticamente

Tabla

Descripción generada automáticamente

Estas tablas serán más tarde importadas a Hive mediante la herramienta Sqoop, de modo que todos nuestros datos (incluyendo los datos no estructurados) se encuentren en una base de datos NoSQL donde puedan ser procesados y analizados.

Nuestros datos en streaming, en cambio, son generados y enviados mediante un socket al agente Flume. Este será el encargado de llevar los tweets a HDFS, e insertarlos en una tabla de Hive manejada externamente. Los campos de esta tabla serán: el usuario que ha publicado el tweet, el número de likes, y el mensaje asociado. De esta manera, el número de likes podría potencialmente ser usado para ponderar los mensajes a la hora de analizar la opinión general, usando este número de impresiones como peso (puesto que un mayor número de likes puede indicar que una gran cantidad de gente concuerda con la opinión expresada).

2.4. Data Transformation Processing

Nuestros datos son procesados en tres modelos diferentes, usando PySpark. El primero de ellos es un análisis sobre las opiniones de los tweets almacenados en la base de datos. Para ello, se utiliza un set de datos previamente etiquetado para entrenar un modelo de regresión logística. La salida de este modelo es una clasificación binaria que indica si la opinión es positiva o negativa. De este modo, y mediante el uso del hashtag contenido en el mensaje, es posible calcular un ratio de satisfacción para cada una de las universidad, presentarlo al usuario, y así que forme parte del proceso de toma de decisión.

Nuestro segundo modelo se trata de un modelo con el objetivo de predecir las notas de corte en el año de ingreso. Puesto que las recomendaciones provistas serían de poca utilidad si el alumno no tuviera posibilidades para acceder a este título, esta predicción es importante para asegurar que este caso no ocurra, y con ello el beneficio que proporciona la herramienta.

El tercer modelo es el que realiza las recomendaciones en sí. Mediante los campos que ha rellenado el estudiante en el cuestionario, se crea un vector, que contiene las áreas en las que está interesado el estudiante con ponderaciones teniendo en cuenta el orden en el que se ha proporcionado. Así, usando el número de investigadores que hay en cada universidad en esos campos, es posible determinar la universidad que más se ajusta a los intereses proporcionados por el estudiante.

De este modo, y con estos tres modelos podemos garantizar la utilidad y veracidad de nuestros sistema.

2.5. Data Query and Insight

Una vez se ha calculado la lista de las tres universidades recomendadas, y teniendo en cuenta el grado que ha indicado el usuario, es necesario proporcionar información de utilidad para el usuario. De este modo, usando tanto los datos estáticos como el resultado del análisis de opiniones somos capaces de proporcionar gráficas que muestran de manera rápida y simple la información de más relevancia para que el usuario pueda tomar una decisión con respecto a su futuro. Algunas de estas gráficas son: la empleabilidad en los grados asociados a cada una de las universidades, el grado de satisfacción, la predicción de notas, el número de investigadores en las áreas de interés, las posiciones en el H-Index.

Estos datos se proporcionan mediante consultas a la base de datos, y gracias al procesamiento realizado en el paso anterior.

3. Results

3.1 Data Ingestion Scripts and Code

Ingesta de datos estáticos a MariaDB:

<https://github.com/Dacarpe03/SIC_Big_Data_CDC/blob/e020d364bde22db329ff39e11a90feeddaf0e237/Scripts/QEDU%20a%20MariaDB.ipynb>

Comandos Sqoop:

<https://github.com/Dacarpe03/SIC_Big_Data_CDC/blob/e020d364bde22db329ff39e11a90feeddaf0e237/Scripts/export_MariaDB_hive.sh>

Generador de Tweets:

<https://github.com/Dacarpe03/SIC_Big_Data_CDC/blob/e020d364bde22db329ff39e11a90feeddaf0e237/Scripts/fake_tweet_generator.py>

<https://github.com/Dacarpe03/SIC_Big_Data_CDC/blob/e020d364bde22db329ff39e11a90feeddaf0e237/Scripts/web_scrapping.py>

3.2 Data Transformation Scripts and Code

Creación del modelo del análisis de sentimientos:

<https://github.com/Dacarpe03/SIC_Big_Data_CDC/blob/e020d364bde22db329ff39e11a90feeddaf0e237/Scripts/AnalisisSentimientoBueno.ipynb>

<https://github.com/Dacarpe03/SIC_Big_Data_CDC/blob/e020d364bde22db329ff39e11a90feeddaf0e237/Scripts/pig_tweets_preprocessing.pig>

Recomendación de Universidades Españolas:

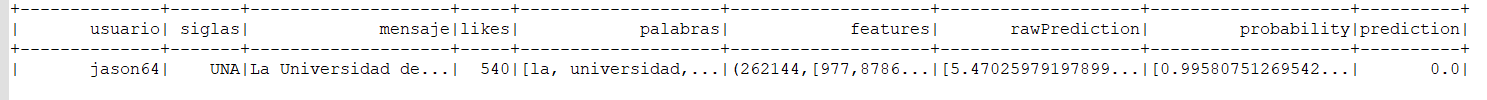
<https://github.com/Dacarpe03/SIC_Big_Data_CDC/blob/e020d364bde22db329ff39e11a90feeddaf0e237/Scripts/recomendacion_universidades.ipynb>

Predicción de notas de corte:

<https://github.com/Dacarpe03/SIC_Big_Data_CDC/blob/e020d364bde22db329ff39e11a90feeddaf0e237/Scripts/regresion_notas_corte.ipynb>

3.3 Description and Sample of Transformed Datasets

El primer Dataset que sufre una transformación es el procedente de la tabla Tweets. Tras la transformación se crean cinco nuevas columnas: *palabras, features, rawPrediction, probability, y prediction*. La más importante será la etiquitada *prediction* puesto que este es el output de la regresión logística. Los valores que puede tomar son 0 o 1, indicando respectivamente si la opinión es negativa o positiva. A continuación, una fila de ejemplo de los valores resultantes:



El modelo de predicción de notas, parte de los datos de las notas de los años 2022 y 2021 de los grados en cuestión. A partir de un dataframe coteniendo esta información, transforma estos datos en una predicción para el año siguiente por medio de una regresión lineal. La siguiente fila es un ejemplo del resultado de esta precicción, siendo la columna *prediction* la salida útil.

Imagen que contiene Forma

Descripción generada automáticamente

El tercer modelo tiene una salida más simple, mostrando sencillamente las tres universidades de mayor ajuste en orden.

Tus universidades recomendadas ordenadas por afinidad son:

1. Universidad Autónoma de Barcelona (UAB) -- Cerdanyola del Vallès

2. Universidad de León (UNILEON) -- León

3. Universidad de Málaga (UMA) -- Málaga

3.4. Data Visualization of Query Results

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico, Gráfico circular

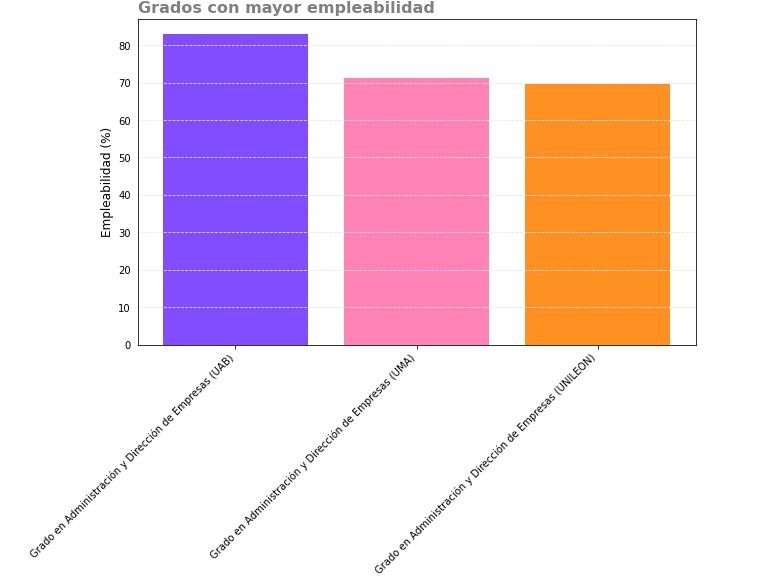
Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico de líneas

Descripción generada automáticamente



4. Projected Impact

4.1. Accomplishments and Benefits

Nuestro sistema, aún sufriendo de limitaciones, consigue mediante proporcionar un resultado tangible. Con ello, simplifica y centraliza el análisis de las mejores opciones, que debe ser tanto de los datos objetivos disponibles como de las opiniones personales y con un valor inestimable de estudiantes que ya han vivido la experiencia y pueden ofrecer una perspectiva única, ahorrando una gran cantidad de tiempo y esfuerzo al futuro alumno de grado, y cuantificando esa información de manera que sus decisiones sean en base a datos razonados.

4.2 Future Improvements

En el futuro, se podría eliminar alguna de las limitaciones que actualmente presisten en el modelo y también expandir el alcance de este análisis de modo que este sea más exhaustivo. En el aspecto de limitaciones, es importante integrar la herramienta con la API de Twitter, u otro servicio, para obtener opiniones reales (se ha intentado con Reddit y Mastodon pero ninguna de estas opciones ha dado resultado). Si esta situación cambia en el futuro, aportaría mucho valor al proyecto. Además, el análisis de datos estáticos se limita a número de investigadores, notas, y posiciones, pero podría expandirse para incluir el número de materias optativas, tasas de éxito, profesorado, y otras variables que afectan la calidad de los estudios ofertados en cada institución.

5. Team Member Review and Comment

|  |
| --- |
|  |

|  |  |
| --- | --- |
| NAME | REVIEW and COMMENT |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

6. Instructor Review and Comment

|  |  |  |
| --- | --- | --- |
| CATEGORY | SCORE | REVIEW and COMMENT |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |