UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria de Barcelona Est

TRABAJO DE FINAL DE GRADO

Grado en Ingeniería Eléctrica

ANÁLISIS DE CONSUMO Y HÁBITOS DE LOS JÓVENES A TRAVÉS DEL BIG DATA

Memoria y Presupuesto

Autor: Daniel Romero González
Director: Joan Martínez Sánchez

Convocatoria: Junio 2020

Resumen

En este trabajo se realiza un análisis de organizaciones relacionadas con la juventud a través del Big Data extraído de Twitter. En primer lugar, se introduce el contexto del Big Data, el Business Intelligence y también de las organizaciones dirigidas al desarrollo de talento juvenil. Después se realiza un mapeo con las organizaciones que tienen cuentas de Twitter representativas y se recopilan. El siguiente paso es elegir la fuente de los datos, en este caso Twitter con su API, a través de un programa de creación propia en entorno Jupyter Notebook y la librería Tweepy para extraer los datos desde una lista en formato .txt. Seguidamente, cuando se tienen estos datos, se procede a la técnica de Business Intelligence denominada Market Basket Case a través del algoritmo A priori que es un análisis de relaciones, implementada en otro software de creación propia. Estos resultados se transformarán en matrices que, finalmente, serán representadas en mapas de calor. Con el análisis de los mapas de calor se llegarán a confirmar ciertas hipótesis de una forma numérica, como que los seguidores de las fundaciones tienden a seguir otras de la misma fundación, y que la fundación Universia tiene una cantidad considerable de seguidores en todo tipo de organizaciones dirigidas a jóvenes. Este tipo de hipótesis, cuando se demuestran numéricamente, se refuerza y puede hacer ahorrar miles de euros en publicidad y colaboraciones al hacerlas más eficaces y eficientes.

Resum

En aquest treball es realitza una anàlisi d'organitzacions relacionades amb la joventut a través de l'Big Data extret de Twitter. En primer lloc, s'introdueix el context del Big Data, el Business Intelligence i també de les organitzacions dirigides a el desenvolupament professional de joves. Després es realitza un mapeig amb les organitzacions que tenen comptes de Twitter representatives i es recopilen. El següent pas és triar la font de les dades, en aquest cas Twitter amb la seva API, a través d'un programa de creació pròpia en entorn Jupyter Notebook i la llibreria Tweepy per extreure les dades des d'una llista en format .txt. Seguidament, quan es tenen aquestes dades, es procedeix a la tècnica de Business Intelligence anomenada Market Basket Casi a través del algoritme Apriori, que és una anàlisi de relacions, implementat a un software de creació pròpia. Aquests resultats es transformaran en matrius que, finalment, han de ser representades en mapes de calor. Amb l'anàlisi dels mapes de calor s'arribaran a confirmar certes hipòtesis d'una forma numèrica, com que els seguidors de les fundacions tendeixen a seguir altres de la mateixa fundació, i que la fundació Universia té una quantitat considerable de seguidors a tot tipus de organitzacions dirigides a joves. Aquest tipus d'hipòtesis, quan es demostren numèricament, es reforça i pot fer estalviar milers d'euros en publicitat i col·laboracions a l'fer-les més eficaces i eficients.

Abstract

In this work, an analysis of youth-related organizations is carried out through Big Data extracted from Twitter. First, the context of Big Data, Business Intelligence and also of organizations aimed at the professional development of young people are introduced. Then a mapping is done with the organizations that have representative Twitter accounts and they are collected. The next step is to choose the source of the data, in this case Twitter via its API, through a self-created program in the Jupyter Notebook environment and Tweepy library to extract the data from a list in .txt format. Then, when these data are available, we proceed to the Business Intelligence technique called Market Basket Case through the Apriori algorithm, which is a relationship analysis, and it's implemented in a self-created program. These results will be transformed into matrices that, finally, will be represented in heat maps. With the analysis of the heat maps certain hypotheses will be confirmed in a numerical way, such as that the followers of foundations tend to follow others of the same foundation, and that the Universia foundation has a considerable number of followers in all kinds of youth organizations. This type of hypothesis, when demonstrated numerically, is reinforced and can save thousands of euros in advertising and collaborations by making them more effective and efficient.

Motivación

Una de las metas del autor de este proyecto es la creación de una incubadora de proyectos dirigidos a jóvenes. El paso previo de esto es ofrecer consultoría a este sector. De esta manera, conociendo y entendiendo el ecosistema, se pueden crear iniciativas que empoderen a la juventud con mayor probabilidad de éxito.

Analizando a través del Big Data un mercado, se pueden ofrecer servicios de consultoría: [1]

- En Marketing y ventas, al conocer tanto a sus propios seguidores como del resto de "players", y sus interacciones puede hacer que se detecten los patrones que tienen mayor eficacia. Por ejemplo: las palabras de mayor impacto, los sentimientos ante determinados temas, etc.
- Innovación y desarrollo de producto, al saber de primera mano qué funciona y con qué otras ideas se puede combinar una.
- En creación de alianzas, al conocer las cuentas que sigue una comunidad, pueden descubrirse nuevas organizaciones con las que llegar a acuerdos que agreguen valor a ambas partes.

Además, este trabajo puede utilizarse para la consultoría de muchos tipos de empresas.

Requerimientos previos

Para la realización de este trabajo se necesita:

- Conocimientos de programación en Python

Gracias a la asignatura de informática y de otras que requieren del aprendizaje de programación, como Estadística u OP.

- Conocimientos de Big Data

Gracias a la asignatura de Gestión de la Innovación se han podido tener conocimientos previos sobre el análisis de datos.
Para la minería de datos, se aprende a utilizar herramientas como Rapidminer y Knime, que permiten cruzar datos y visualizarse de una manera gráfica y entendible.
También, en la asignatura de Inteligencia Artificial Aplicada a la Ingeniería se ha tenido una base a partir de la cual ha sido posible la adaptación a los requerimientos de este trabajo.

- Conocimientos de Redes Sociales y de Marketing

Si se quiere analizar un mercado a través de las Redes Sociales, es obvio que se necesitan unos conocimientos sobre cómo funcionan estas.

- Conocimiento del ecosistema de la juventud

Dado que es el estudio de mercado que se va a realizar, a través de una investigación particular he explorado este ecosistema. Las organizaciones seleccionadas tienen principalmente el foco en el aprovechamiento y desarrollo del talento joven, pero también hay otro tipo de fines. Por ejemplo: autoescuelas, copisterías online, etc.

- Conocimientos empresariales

Gracias a las asignaturas de Empresa, Gestión de la Innovación y Liderazgo y dirección, se ha tenido un conocimiento de las empresas y de la innovación que ha posibilitado el entendimiento de las empresas y organizaciones. Esto además se complementa con un master online de negocios llamado ThePowerMBA, que en el transcurso del trabajo estoy también cursando.
De esta manera, los servicios y productos desarrollados tienen una aplicación práctica, y se pueden encajar en las necesidades empresariales cuando sea necesario.

Índice

RESUMEN I

RESUM II

ABSTRACT III

MOTIVACIÓN IV

REQUERIMIENTOS PREVIOS V

INTRODUCCIÓN IX

Objetivos del trabajo ix

Alcance del trabajo ix

CAPÍTULO 1: CONTEXTO 1

1.1 ¿Qué es el Big Data? 1

Casos de uso del Big Data 1

1.2 Sector de la juventud 2

1.3 Mapeo de organizaciones con cuenta de Twitter 2

1.3.1 Programas y cursos 3

1.3.2 Fundaciones 3

1.3.3 Comunidades 5

1.3.4 Asociaciones de estudiantes 5

1.3.5 Universidades privadas 7

1.3.6 Academias online 7

1.3.7 Cursos de idiomas en el extranjero 7

1.4 *BUSINESS INTELLIGENCE* 8

1.5 ESTRUCTURA DEL PROYECTO 11

3 EVALUACIÓN DE LAS REDES SOCIALES DONDE EXTRAER DATOS 12

3.1 Redes sociales más usadas 12

3.1.1 Facebook 13

3.1.2 Twitter 14

3.1.3 Instagram 16

3.2 La red social elegida: Twitter 19

3.3 EXTRACCIÓN DE DATOS DE TWITTER 20

2 REALIZACIÓN DE LAS EXTRACCIONES DE DATOS DE CUENTAS CON TWEEPY 22

3.1 Estructura de datos 22

3.2 Cómo son los archivos utilizados 23

4 MINADO DE DATOS Y CREACIÓN DEL MAPA DE CALOR 31

4.1 Market Basket Case 31

4.2 Funcionamiento del algoritmo del Market Basket Case 37

5 ANÁLISIS DE DATOS Y PROPUESTAS A LA TOMA DE DECISIONES 46

5.1 Análisis específico de los seguidores de las cuentas 46

5.2 Conclusiones y propuestas concretas 54

SUMARIO DEL TRABAJO 55

Conclusiones de uso práctico del estudio 56

Posibles evoluciones de este proyecto: 57

Ampliaciones de este tipo a otras situaciones o problemáticas 58

Valoración del aprendizaje y competencias adquiridas 59

PRESUPUESTO 61

BIBLIOGRAFÍA 63

Introducción

Objetivos del trabajo

El objetivo de este trabajo es encontrar las relaciones entre organizaciones a través de analizar datos reales para así obtener información fiable, más allá de las hipótesis que una organización pueda plantearse. Para ello, se ejecuta un mapeo de organizaciones, extraer sus datos y hacer un análisis de relaciones.

Alcance del trabajo

El alcance de este trabajo es el de las organizaciones con público juvenil representadas con sus cuentas de Twitter: fundaciones, organizaciones de estudiantes, programas formativos, educación online y comunidades.

Capítulo 1: Contexto

1.1 ¿Qué es el Big Data? [2]

El B*ig Data* está formado por conjuntos de datos de mayor tamaño y más complejos, especialmente procedentes de nuevas fuentes de datos. Estos conjuntos de datos son tan voluminosos que el software de procesamiento de datos convencional sencillamente no puede gestionarlos. Sin embargo, estos volúmenes masivos de datos pueden utilizarse para abordar problemas empresariales que antes no hubiera sido posible solucionar.

Lo que se hace con las técnicas de Big Data es analizar estos datos para crear modelos que permitan obtener información relevante y crear modelos que permitan predecir el comportamiento.

El aprendizaje automático o "machine learning" hace que unos modelos aprendan de los datos que van entrando, y que cada vez que entran nuevos datos este vaya mejorando.

Casos de uso del Big Data

Desarrollo de productos

El Big Data está presente en las compañías tecnológicas más importantes. Por ejemplo, Netflix la utiliza para prever la demanda de sus clientes. De esta manera, analizando el consumo de sus servicios anteriores y actuales, pueden saber con mayor certeza qué lanzamientos pueden tener mayor éxito. Al análisis de sus usuarios, también se suman los datos y analítica de redes sociales, tendencias, grupos de interés, etc.

- Experiencia de cliente [3]

Amazon ha invertido mucho en Big Data para mejorar al máximo la experiencia de usuario y para persuadirlo de cara a que gaste más dinero. Por ejemplo, cuando seleccionas un producto te sugiere automáticamente otros que también pueden interesarte.

- Impulso a la innovación [4]

Con el Big Data puede conseguirse información clave para innovar. Por ejemplo, estudiando la relación entre personas, instituciones, entidades y procesos, y después determinando nuevas maneras de utilizar la información.

Y los sectores donde se pueden aplicar también son muy numerosos.

Los artistas pueden conocer mejor lo que quiere su audiencia, puede predecirse cuanta energía se necesitará en una región para optimizar la producción eléctrica, en el sector de la moda saber qué características tiene que tener la ropa de la nueva temporada para vender más, etc.

El objeto de este trabajo es analizar el sector de los jóvenes, detectar oportunidades relacionadas y, si es posible, ofrecer un servicio que aporte valor a las organizaciones que se dirijan a ellos.

1.2 Sector de la juventud

No existe un sector "de los jóvenes" como tal, pero sí que hay una serie de actividades económicas e iniciativas que se dirigen a estos.
La juventud, según la ONU, son los jóvenes con edad entre 15 y 24 años, sean del país que sean. [5]
Sin embargo, estos grupos son muy heterogéneos. Por ejemplo, un jóven de 15 años no puede trabajar legalmente en España, mientras que muchos jóvenes de 24 años han acabado sus estudios y tienen una ocupación, sea relacionada con su vocación o no. También tienen hábitos distintos, por ejemplo, los canales de Youtube que consumen, los locales de ocio que frecuentan o tipo de películas que consumen.
Esa diferencia de poder adquisitivo, de obligaciones y de hábitos es una muestra de que este grupo es tan diverso que agruparlo todo en una categoría puede dar lugar a demasiadas imprecisiones.
Para un público juvenil, en este caso estudiantes, hay proyectos basados en el Big Data.
Por ejemplo [6], con la intención de reducir el fracaso escolar, hay universidades que están desarrollando sistemas de Inteligencia Artificial para el seguimiento y la orientación profesional.
David Bañeres, investigador en la UOC, afirma que permitirá recomendar a cada persona los estudios que sean más adecuados para ellos según sus gustos, capacidades y proyección laboral. También se podrá detectar un posible abandono, o avisar cuando hay más riesgo de suspender una asignatura.

1.3 Mapeo de organizaciones con cuenta de Twitter

Dentro de la juventud, en este trabajo se exploran jóvenes pueden aportar valor especialmente a las organizaciones y empresas. Por ello, el foco está principalmente en los jóvenes talentos, y las organizaciones que así lo potencian.

Estas son principalmente fundaciones, programas educativos, organizaciones de estudiantes y comunidades.

Este mapeo se basa en las iniciativas que potencian el talento joven.

Los tipos son:

- Programas y cursos
- Fundaciones
- Comunidades
- Asociaciones
- Universidades privadas
- Formación complementaria
- Programas extranjero
- Plataformas online

1.3.1 Programas y cursos

Dentro de Programas y cursos, nos encontramos con:

- La Akademia

La Akademia es un curso gratuito para jóvenes de entre 18 y 22 años sobre educación emocional, autoconocimiento y vocación profesional.

- Celera

Celera es un programa donde se buscan a los mejores talentos jóvenes, se seleccionan 6 y se les capacita durante 3 años con herramientas para explotar al máximo su potencial. Está financiado por la fundación Rafael del Pino.

- Factoría de talento (Adecco)

Factoría de talento es un programa financiado por la empresa de Recursos Humanos Adecco, para desarrollar habilidades profesionales en jóvenes talentos.

- Jóvenes Juristas

Jóvenes Juristas es una organización sin ánimo de lucro que ofrece formación complementaria y oportunidades profesionales a estudiantes de derecho.

Sus cuentas de Twitter son:

factoriatalent0
laakademia_org
_celera
jovejuristacat

1.3.2 Fundaciones

Las fundaciones son entidades sin ánimo de lucro que, en este caso, ofrecen programas formativos a jóvenes. Estas en concreto se caracterizan por estar patrocinadas por empresas importantes.

Dentro de las fundaciones, nos encontramos con:

- Fundación Rafael del Pino

Rafael del Pino fue un empresario español que fundó Ferrovial. Actualmente (2020) su hijo preside la empresa. El año 2007 la revista Forbes lo ubicó en el puesto 79 de los hombres más ricos del mundo.

La fundación Rafael del Pino financia becas, programas educativos sobre liderazgo, innovación y jóvenes talentos, y también ha financiado Unleash, una conferencia de jóvenes talentos.

- Fundación Bankinter

Bankinter es un banco español que forma parte del IBEX 35.

En su fundación tiene un programa para estudiantes llamado Akademia, sobre innovación y desarrollado en colaboración con universidades.

- Fundació Princesa de Girona [7]

La Fundació Princesa de Girona, formada por más de 90 patronos de primer nivel, aspira a ser un referente a nivel español en apoyo a los jóvenes en su desarrollo personal y profesional. Celebra premios a jóvenes referentes (premios FPdGi) y programas para la mejora de la empleabilidad como Rescatadores de talento, y la transformación educativa de los jóvenes docentes. En 2020, destinará 2,8 millones de euros a programas en beneficio de la juventud.

- Universia

Universia es una red de universidades iberoamericanas, financiada por el banco Santander. Los ejes de actuación de Universia son la orientación académica, el empleo y la transformación digital universitaria.

Universia tiene un portal web muy completo para saber más sobre estos temas.

Sus cuentas de Twitter son:

Universia
frdelpino
FundacionBKT
FPdGi

1.3.3 Comunidades

Trivu

Trivu es una consultora que ayuda a las empresas a conectar con jóvenes para así captar talento y resolver problemas teniendo en cuenta a la juventud. Dispone de una comunidad muy potente a nivel español, donde se encuentran organizaciones de jóvenes y también talento muy potente.

Organizan actividades como hackatones, eventos y conferencias.

- Global Shapers

Global Shapers es una comunidad de jóvenes entre 20 y 30 años, auspiciada por el Foro Económico Mundial, que conecta a personas con la misión de generar un impacto social.

La cuenta de Global Shapers a analizar será la de Madrid, dado que la de Barcelona y la de Bilbao tienen muy pocos seguidores.

Las cuentas de comunidades son:
wearetrivu
ShapersMadrid

1.3.4 Asociaciones de estudiantes

Las asociaciones son entidades sin ánimo de lucro con un objetivo común.

Las asociaciones a analizar son aquellas que tienen ámbito internacional y una presencia importante en España, es decir, en varias ciudades y con un tamaño considerable.

Las asociaciones a analizar se dividen en distintos tipos:

Actividades en otros países

Una parte importante de estas actividades están subvencionadas por el programa Erasmus+ de la Unión Europea.

- ESN

Erasmus Student Network es una red de asociaciones encargadas de recibir a los alumnos Erasmus y organizar actividades para ellos.

- AEGEE

AEGEE es la organización de jóvenes más grande de Europa. Organizan actividades a nivel local, internacional y cada verano organizan “summer courses” co-financiados por la Unión Europea.

Prácticas en el extranjero

- IAESTE

IAESTE o asociación internacional para el intercambio de estudiantes para experiencia técnica es una organización sin ánimo de lucro que ofrece prácticas internacionales para estudiantes de especialidades técnicas.
Se centran sobretodo en arquitectos, ingenieros técnicos, ingenieros civiles y otras especialidades del sector.

- AIESEC

AIESEC es la organización de jóvenes más grande del mundo. Su misión es fomentar la paz en el mundo y desarrollar el máximo potencial del ser humano a través del liderazgo. Ofrecen voluntariados y prácticas internacionales, donde se desarrollan habilidades de liderazgo en cada uno de los participantes. Estos se orientan a cuplir los Objetivos de Desarrollo Sostenible de la ONU.
Las prácticas disponibles están relacionadas con los negocios, ventas, ingeniería, derecho y diversas disciplinas. Entonces, cualquier estudiante podría sacarle provecho.

Asociaciones de temáticas especializadas

- ELSA

ELSA o European Law Student Association es una organización europea de estudiantes de derecho. Complementan los estudios de derecho a través de programas y formaciones.

- BEST

BEST (Board of European Students of Technology) es una organización internacional de estudiantes de tecnología que tienen representación en las principales ciudades europeas. Organizan actividades como hackatones y formaciones subvencionadas por el programa Erasmus+ y relacionadas con el desarrollo de habilidades y aprendizaje. Estas se orientan a al desarrollo personal, profesional y ocio de estudiantes.
Como su estructura es por comités locales, se eligen dos cuentas significativas: la de Madrid y la de Barcelona.

La lista de cuentas de asociaciones es:

ESNSpain

aegeebarcelona

AEGEE_Zaragoza

AEGEEMalaga

IAESTE_Spain

aiesecspain

ELSA_Spain_

BESTMadridUPM
BESTUPC
BESTValenciaUPV

1.3.5 Universidades privadas

Teamlabs

TEAMLABS/ es una plataforma de creación de laboratorios de aprendizaje e innovación que apuesta por una metodología radicalmente nueva, inspirada en la metodología finlandesa Team Academy, que difumina las barreras entre los mundos académico y profesional.

Teamlabs está auspiciada por la Universidad de Mondragón. Tiene grados universitarios, postgrados y cursos para perfiles multidisciplinares donde, entre otras cosas, se desarrollan empresas y otras iniciativas.

Sus listas de cuentas son:

teamlabs
MondragonTA

1.3.6 Academias online

- ThePowerMBA

The Power MBA es una academia online donde se imparten masters sobre negocios y marketing, y otro tipo de formaciones complementarias. Uno de sus puntos más fuertes es la gran comunidad que tienen.

Su cuenta de Twitter es:

ThePowerMBA

1.3.7 Cursos de idiomas en el extranjero

- Education First

Education First es una empresa internacional presente en 50 países dedicada a la enseñanza de idiomas. Tienen una gran oferta de cursos en otros países, y se focaliza en los estudiantes.

Su cuenta de Twitter es:

EFEspana

1.4 *Business Intelligence* [8][9]

Antes se decía que la información es poder. Ahora, el poder es interpretar estos datos.

El *Business Intelligence* es el uso de estrategias y herramientas para transformar información en conocimiento con el fin de mejorar la toma de decisiones de una empresa.

Desde el punto de vista del Project Management Institute, la gestión de un proyecto es tan fundamental como la solución técnica. En la paradoja de Cobb del el año 1989, se plantea lo siguiente: "Sabemos por qué fallan los proyectos, sabemos cómo prevenir sus fallos - ¿Por qué siguen fallando?".

El año 2012, se busca una solución a esta forma de pensar, y se concluye que es imposible resolver la paradoja de Cobb por las siguientes razones:

- Todos los proyectos son arriesgados
- La mayoría de los proyectos incluyen riesgos inmanejables
- La gestión de riesgos no siempre se hace bien
- Las actas de constitución a menudo omiten los umbrales de riesgo
- Los proyectos deberían existir en un portafolio balanceado de riesgo
- La innovación se genera en el fallo
- Fallo al aprender

En todo tipo de proyecto existen estos riesgos, y la repetición de los errores hace concluir que no siempre se conocen las causas de estos errores.

En el artículo referenciado de Esther Hochsztain (Departamento de Métodos Cuantitativos Facultad de Ciencias Económicas y de Administración. UDELAR, Uruguay) y Andrómaca Tasistro (Agencia de Gobierno Electrónico y Sociedad de la Información (AGESIC), Uruguay) se presenta un "Framework" para a través del conocimiento y el análisis del riesgo, aumentar las probabilidades de éxito.

En la siguiente figura, se presentan los niveles a tener en cuenta:

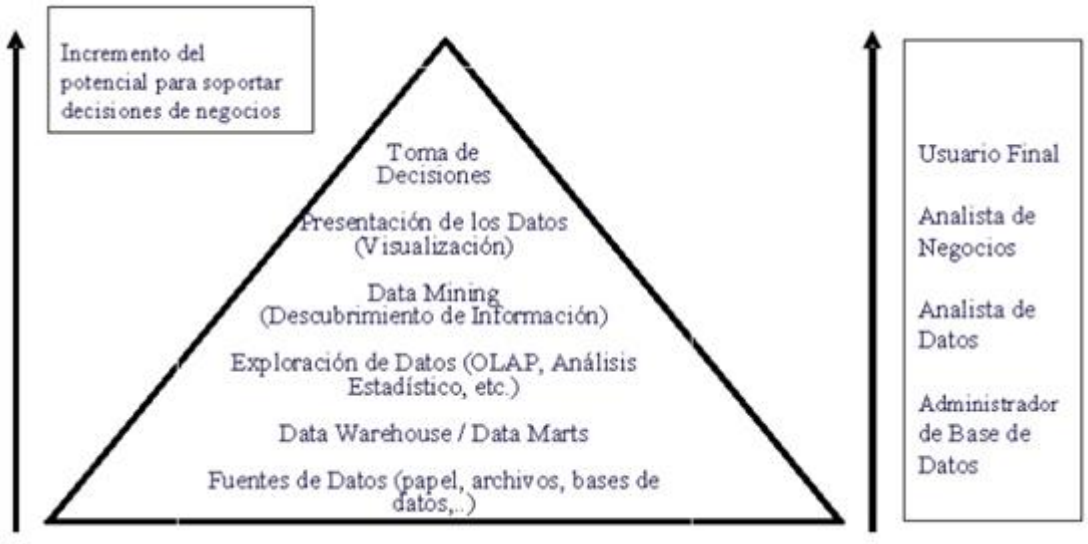

Figura 1.4.1: Niveles en la aplicación de técnicas de *Business Intelligence* [8]

Como se puede observar en este *"framework"*, los datos y su respectivo análisis son una pieza fundamental en la gestión de proyectos. De esta manera, se pueden detectar los errores con mayor precisión y, con ello, la toma de decisiones tendrá un menor riesgo.

En este proyecto se pueden encontrar los diferentes niveles encontradas en la figura 1.4.1:

- Fuentes de datos

Las fuentes de Datos son todos los datos que se utilizarán para mejorar la toma de decisiones. En este trabajo, se ha elegido Twitter porque a través de su API, y programado en Python, es posible extraer grandes cantidades de datos.

- Data Warehouse

Los Data Warehouse o almacenes de datos es la colección de datos que será interpretada posteriormente.

La estructura de este estudio se basa en archivos .csv, que para poder ser útil a la vez de estar todo en una misma carpeta, se utiliza un código que se distingue por sus variables.

- Exploración de datos

Es el análisis de los datos utilizando funciones estadísticas básicas.

- Data Mining

La minería de datos consiste en que, a través de técnicas avanzadas de análisis, se descubran patrones y nueva información invisible a un simple análisis estadístico.

Para este trabajo se utiliza tanto la programación en Python como el software Rapidminer.

- Presentación de los datos

Los resultados de estas técnicas deben de tener una utilidad y lo deben de poder entender de una manera eficaz la junta directiva y otras personas con poder de decisión en la organización.

La propia presentación del proyecto delante del tribunal es un ejemplo de ello.

- Toma de decisiones

Con toda la información extraída, se toman las decisiones correspondientes.

Por ejemplo, supongamos que el 10% de los seguidores de la cuenta de Twitter de AIESEC siguen a Tony Robbins. Pues se creará más contenido sobre este autor, se intentará invitarlo a conferencias de AIESEC, y se les hará más menciones. De esta manera, los seguidores ganarán afinidad con la marca AIESEC, y los seguidores de Tony Robbins podrán conocer AIESEC, ganando así seguidores y potenciales clientes.

1.5 Estructura del proyecto

De forma gráfica, este trabajo consiste en las siguientes partes:

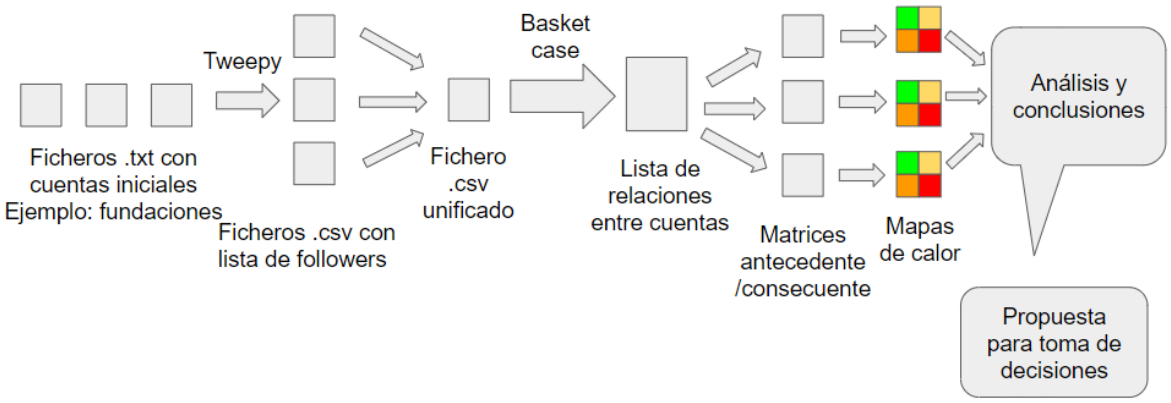

Figura 1.5.1: Diagrama de los procedimientos del proyecto

Las cuatro fases en las que se divide la adquisición de conocimiento desde las cuentas de Twitter son las siguientes:

Fase 1: Selección de cuentas y redes

Fase 2: Realización de las extracciones de datos de cuentas con Tweepy

Fase 3: Minado de datos y creación del mapa de calor

Fase 4: Análisis de datos y propuestas a la toma de decisiones

3 Evaluación de las redes sociales donde extraer datos

3.1 Redes sociales más usadas

Según un estudio realizado por “We are social” y Hootsuite, estas son las redes sociales más utilizadas a nivel mundial:

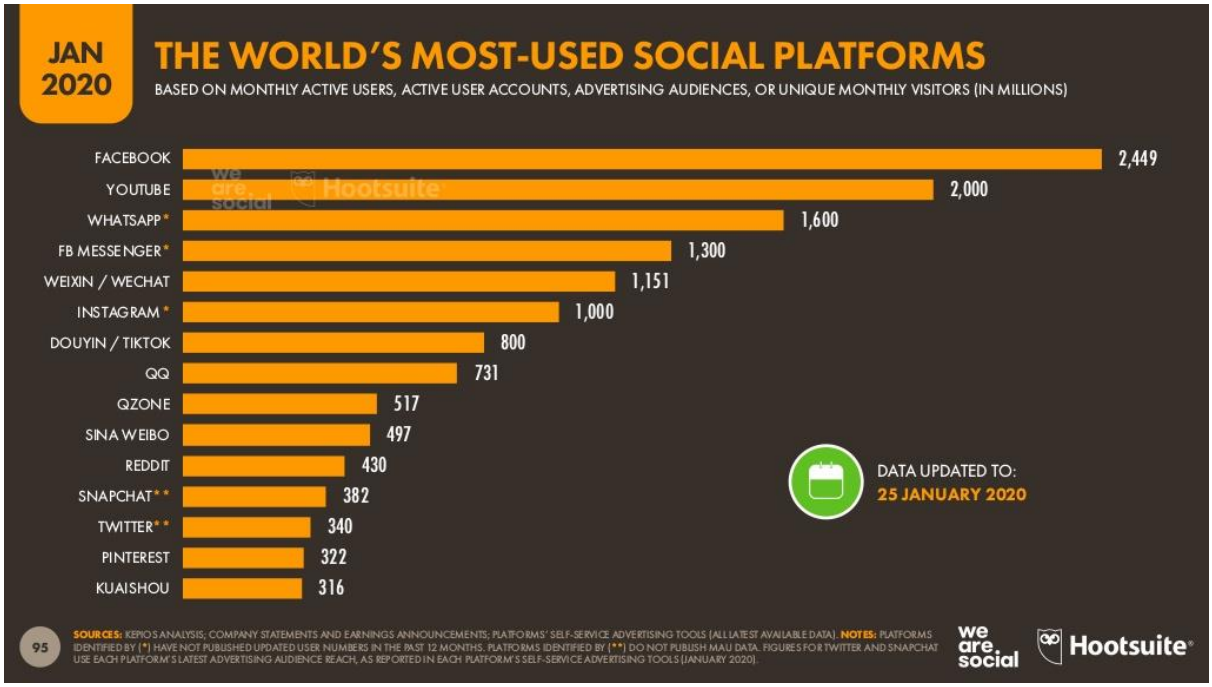

Figura 3,1: Estudio de las redes sociales más utilizadas en el mundo. [10]

Según el informe de la consultora de comunicación The Social Media Family, en España las redes sociales más utilizadas han evolucionado de la siguiente manera:

	2014	2015	2016	2017	2018	2019	Evolución
Facebook	20 millones	22 millones	24 millones	23 millones	24 millones	22 millones	-8,33%
Twitter	3,5 millones	4,4 millones	4,5 millones	4,9 millones	4,9 millones	4,4 millones	-10,20%
Instagram		7,4 millones	9,6 millones	13 millones	15 millones	16 millones	6,66%

Usuarios de redes sociales en España en diferentes años. [11]

Aquí se puede observar cómo la tendencia de Twitter y de Facebook es de reducir sus usuarios, y de Instagram de aumentarlos.

3.1.1 Facebook

Facebook es una red social donde se pueden hacer publicaciones de todo tipo: imágenes, videos, textos, enlaces, etc. Está diseñada y preparada para tanto ordenador como móvil.

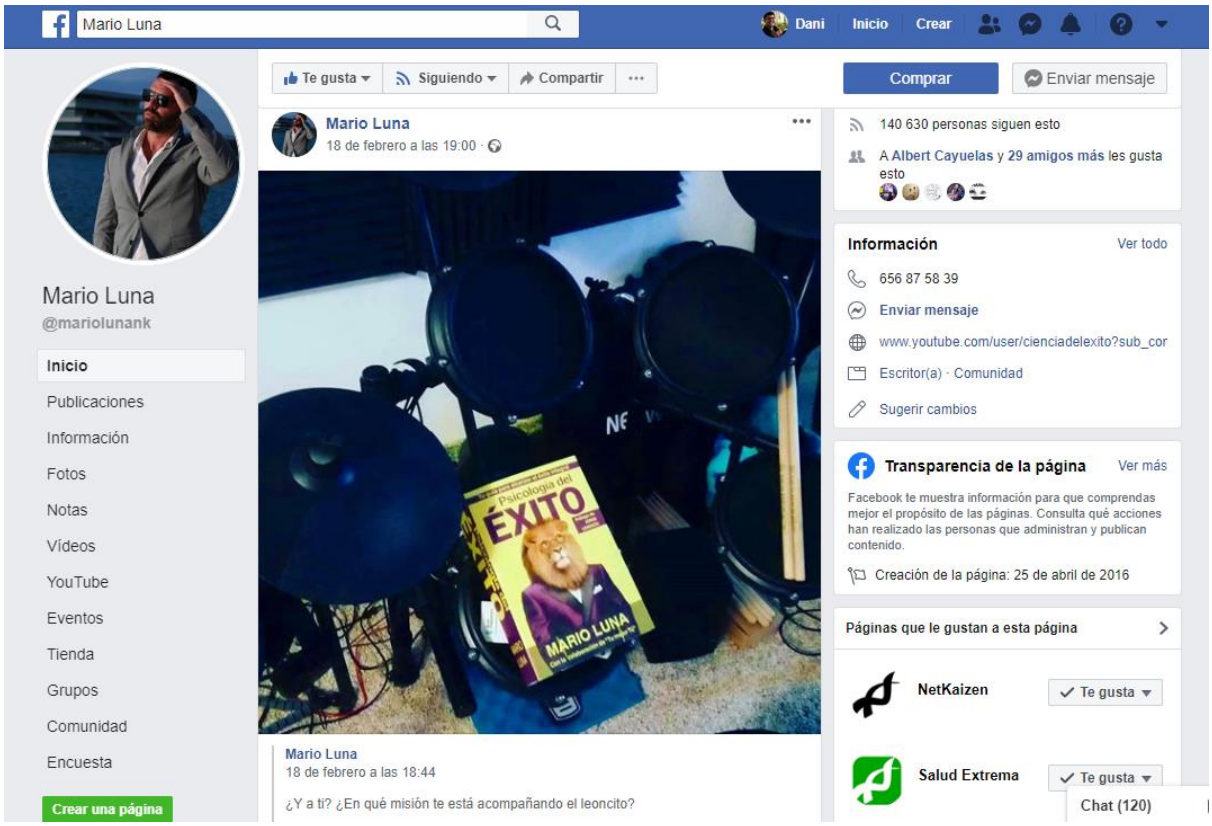

Figura 3.2 Interfaz gráfica de Facebook

Facebook permite la descarga de datos públicos y de tus propias páginas y publicaciones.

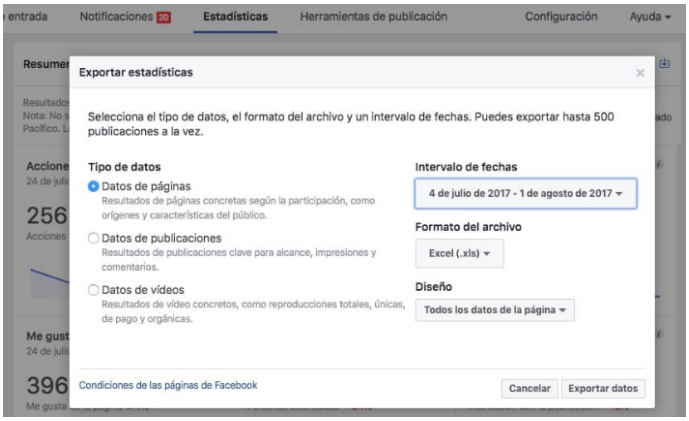

Figura 3.3 Ventana de API de Facebook

La principal desventaja de Facebook es que hay una gran cantidad de usuarios que tienen datos privados. Entonces, por ejemplo, extraer uno a uno la lista de "Me gustas" se hace imposible.

También, Facebook proporciona datos ya tratados, es decir, estadísticas y segmentaciones. Entonces, se complica la opción de extraer datos para tratarlos y hacer el minado correspondiente.

Y otra característica de Facebook es que las cuentas de usuario y las páginas son distintas. Esta diferenciación dificulta la extracción estandarizada de datos.

3.1.2 Twitter

Twitter es una red social que se distingue por, en cada publicación escrita, tener un límite de 280 caracteres. De esta manera, los usuarios se expresan de una forma abreviada y concisa.

En enero de 2020 tenía a nivel mundial 340 millones de usuarios activos y es la 11ª a nivel mundial, según un estudio realizado por Hootsuite.

Figura 3.4 Interfaz grádica de Twitter en la descripción del perfil

Figura 3.5 Interfaz gráfica de Twitter en el muro

Twitter también se caracteriza por el uso de etiquetas (hashtags) "#", donde los usuarios agrupan las publicaciones por temáticas. Se recopilan en Tendencias (Trending Topics) las etiquetas más utilizadas durante un período de tiempo.

Para extraer datos de Twitter se puede utilizar su API a través de programación en Python con la librería Tweepy. Para ello, hay que inscribirse en el formulario de Twitter, identificarte y justificar el uso que le vas a dar. Después de eso, si Twitter te lo acepta, puedes crear tu programa para utilizar la API y extraer los datos que necesites.

Utilizando la librería Tweepy, se pueden extraer en formato txt, la lista de atributos de los usuarios: Nombre, Estatus, Número de amigos/seguidos, Número de seguidores, Localización, Idioma, Descripción, etc. Puede incluso extraerse esta lista de atributos de todos los seguidores y amigos/seguidos de una cuenta. Además, como norma general, los propios usuarios hacen pública y abierta esta información, por lo que se puede utilizar sin ningún problema.

3.1.3 Instagram

Instagram es una red social que se basa en compartir imágenes. Está diseñada para móviles, por lo que su experiencia de usuario se basa en una app para Android y para iPhone.

La interfaz gráfica de una cuenta es la siguiente:

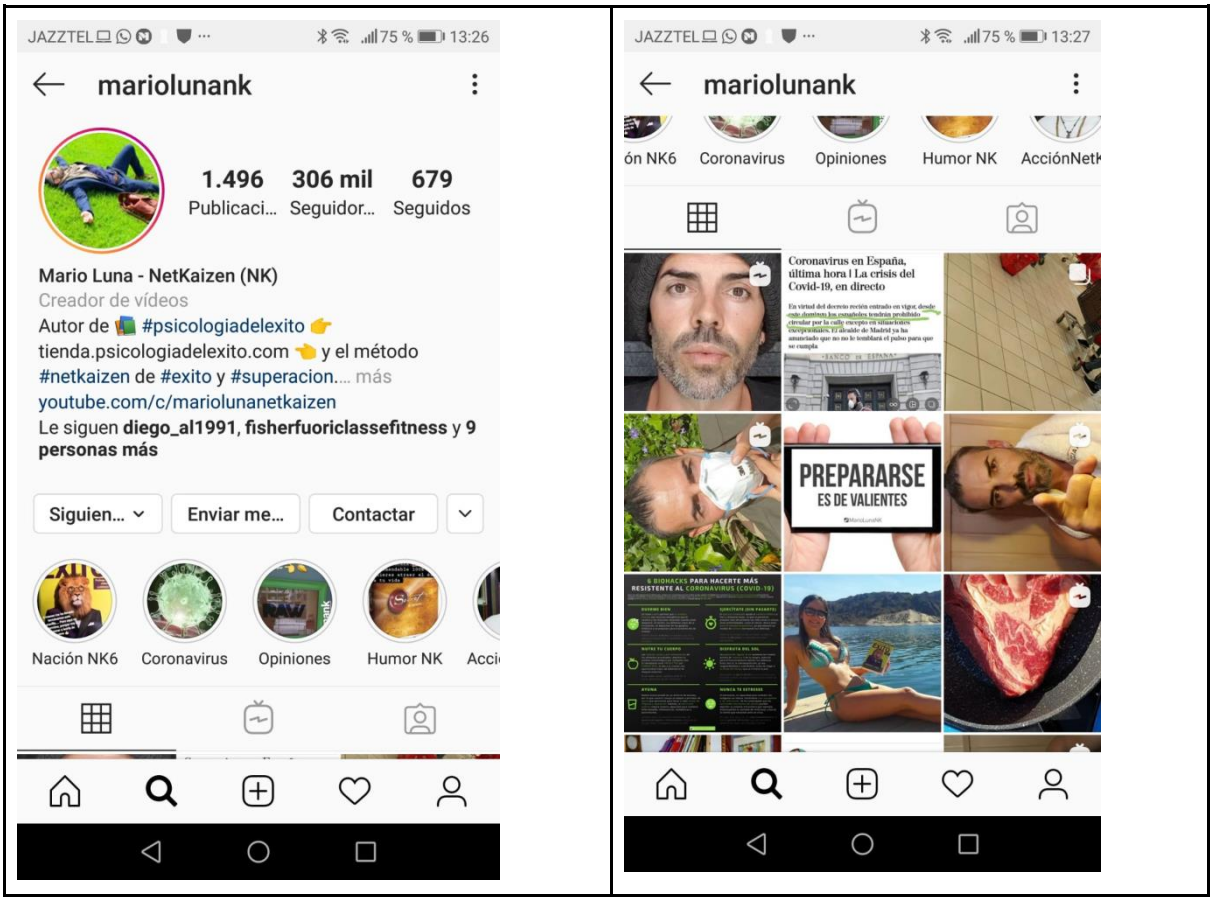

Figura 3.6 y 3.7 Interfaz gráfica de Instagram

Las publicaciones pueden ser:

- Stories, que desaparecen públicamente a las 24h
- Colección de Stories
- Publicaciones comunes

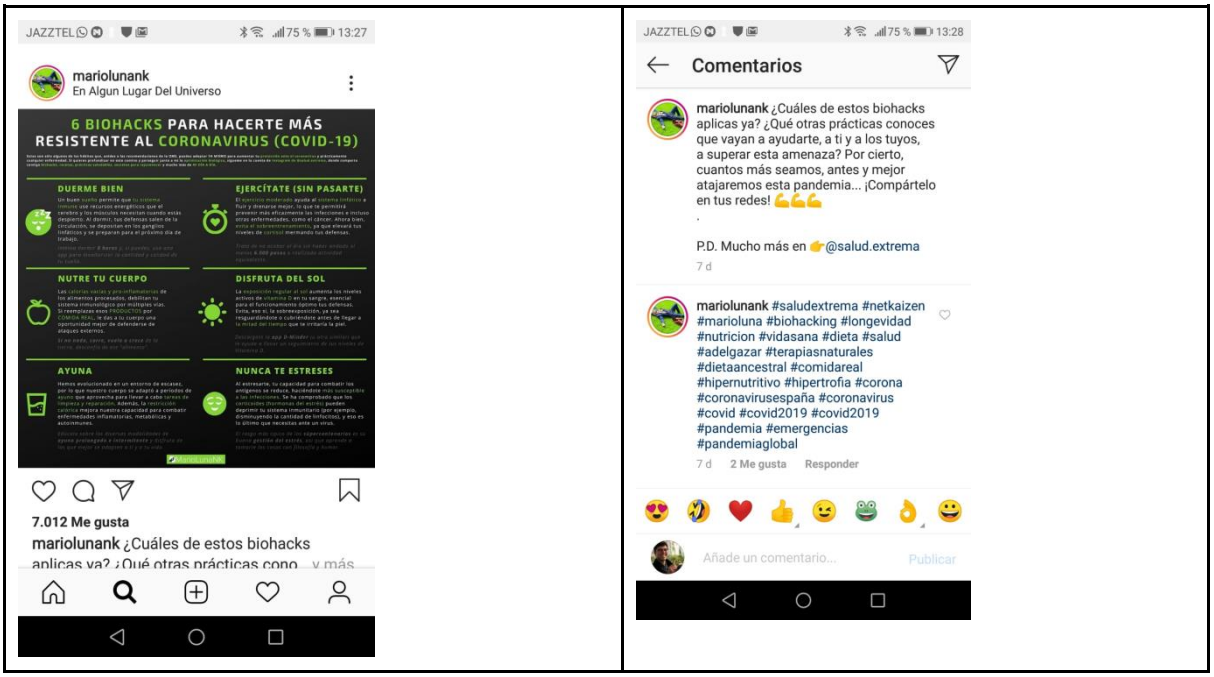

Figura 3.8 y 3.9 Interfaz gráfica de publicaciones (3.8) y comentarios (3.9) de Instagram

Instagram únicamente dispone de una API gráfica, y desactivó su API anterior. Este tipo de cambios generan una incerteza para desarrollar una extracción de datos, dado que la herramienta programada podría quedar obsoleto en poco tiempo.

Igualmente, la API de Instagram está enfocada en los resultados de las cuentas Business de Instagram y para solo sus propias cuentas. Entonces, desde Instagram no se facilitan herramientas para analizar de manera exhaustiva los seguidores de otras cuentas.

Es posible crear sistemas a través de Web Scrapping, pero con cualquier pequeño cambio en la programación ya dejaría de funcionar.

A través del Web Scrapping, se pueden extraer, solo de cuentas públicas y que hayan aceptado la petición de seguimiento de tu cuenta.

Otro problema de Instagram es que, como se basa en imágenes, para analizar estas se necesitaría de mucha memoria para almacenar imágenes, y de una capacidad de procesamiento de imágenes muy potente.

Aún así, se podrían extraer los siguientes datos que pueden ser interesantes:

- Nombre
- Descripción
- Lista de seguidores
- Lista de seguidos

- Descripción de las publicaciones
- Lista de cuentas que dieron “like” a las publicaciones
- Comentarios de las publicaciones
- Lista de cuentas que publicaron en un Hashtag

3.2 La red social elegida: Twitter

De las otras dos redes sociales se les puede extraer mucha información valiosa, y analizar sus datos puede dar como resultado un impacto mayor por su cantidad de usuarios.

Sin embargo, se escoge Twitter por las siguientes razones:

- Cantidad de usuarios suficiente

No es la red social más utilizada, pero es suficiente para hacer el estudio de mercado de este trabajo.

- Facilidad de extracción de datos

Para extraer los datos se necesita programar en Python, y una vez creado, esta aplicación puede obtener la lista de seguidores, lista de seguidos, descripciones, publicaciones, etc. De una forma rápida y sencilla se pueden conseguir datos muy valiosos.

- Su uso principal es en texto

Twitter es una plataforma de microblogging, es decir, que sus usuarios suelen expresarse escribiendo. Entonces, para obtener información como Sentiment Analysis, palabras más comunes, etc, es eficaz.

La ventaja del minado de datos con técnicas de análisis se podrían aplicar igualmente a cuentas y publicaciones de Facebook, Instagram e incluso a Linkedin, donde hay un mercado de Recursos Humanos muy potente. Sin embargo, para concentrar esfuerzos, Twitter será la fuente de datos del trabajo.

3.3 Extracción de datos de Twitter

Una API (Application Programming Interface) es un conjunto de definiciones y protocolos que sirve para comunicar unos servicios con otros sin necesidad de formar unos parte de otros. Esto simplifica la creación de aplicaciones, ahorrando tiempo y dinero en su desarrollo.

Fuente: https://www.redhat.com/es/topics/api/what-are-application-programming-interfaces

Para utilizar la API de Twitter hay que pedir autorización a la red social.

En primer lugar se necesita tener un usuario de Twitter. En este caso, utilizaré el mío personal.

Después, hay que entrar en https://developer.twitter.com/en/apps , rellenar el formulario y explicar los usos que se le darán. En este caso, se debe de indicar la universidad, el nombre del alumno, del tutor y qué se harán con los datos. Todo ello, escrito en inglés.

Una vez conseguida la autorización, se crea una app y de ahí se extraen las 4 claves necesarias para utilizar la API: Consumer API key, Consumer API key secret, Access token y Access token secret.

La API de Twitter tiene limitaciones, como un máximo de cuentas cada vez que se hace una llamada concreta:

En la siguiente tabla se muestran algunos ejemplos de peticiones y cantidad de información obtenida.

API	Petición	Max. datos por petición (página)	Cada 15 minutos
REST	GET statuses/user_timeline	200 tuits	900 * 200 = 180.000 tuits
REST	GET users_show	1 perfil	1 * 900 = 900 perfiles
REST	GET follwers_list	200 perfiles	200 * 15 = 3.000 perfiles
REST	GET followers_ids	5.000 ids de usuario	5.000 * 15 = 75.000 Ids
Search	GET search/tweets	100 tuits	180 * 100 = 18.000 tuits
Streaming	POST statuses_filter	–	Máximo de 45.000 tuits

Figura 3.10: Límites en la API de Twitter

Por ello, en el software desarrollado se incluirá la posibilidad de esperar el tiempo necesario cuando se alcance el límite, y repetir las llamadas hasta conseguir todos los datos requeridos para este estudio.

- **Tweepy**

Tweepy es una librería de Python que utiliza la API de Twitter. A través de los comandos adecuados, se extraerá la información de los usuarios seleccionados y de sus seguidores.

- **Jupyter notebook**

Jupyter Notebook es un entorno informático interactivo que permite programar en Python y además generar un documento.

Este se ejecuta en la consola de comandos de Windows y aparece la interfaz en el navegador web predeterminado, siendo así más ligero que otros softwares para programar en Python.

2 Realización de las extracciones de datos de cuentas con Tweepy

3.1 Estructura de datos

Los datos extraídos deben de estar ordenador de manera que trabajar con ellos sea sencillo y rápido. Para ello, los archivos se ordenan de la siguiente manera:

- Todo en la misma carpeta

Todos los archivos se encuentran en la misma carpeta, en el mismo lugar donde se encuentra el documento de Jupyter. De esta manera, es accesible en cualquier momento.

- Tipos de archivos

En este trabajo, los archivos con datos son principalmente tres:

- 0: Lista de cuentas de Twitter a analizar

En este caso, las cuentas están recopiladas en varios archivos .txt, como 0Asociaciones_estudiantes.txt o 0Fundaciones.txt, aunque se han fusionado todas en el archivo 0Conjunto_inicial.txt para simplificar el procedimiento.

- 1: Lista de seguidores de la cuenta 0 de Twitter a analizar

Cada cuenta de la que se han extraído los seguidores tiene su propio archivo llamado 1SD_[NOMBRE DE LA CUENTA]_TAG_[NOMBRE ARCHIVO TIPO 1].csv

En el caso de extraer todas las cuentas unificadas, cuando se extraiga la cuenta JoveJuristaCat, aparecerá un archivo titulado 1SD_jovejuristacat_TAG_Conjunto_inicial.csv

- 2: Recopilación de listas de seguidores.

Por cada lista de seguidores extraída, finalmente se recopilan todas en el archivo 2Conjunto_final.csv

3.2 Cómo son los archivos utilizados

- **.txt tipo 0**

Son las listas de cuentas que se van a analizar.

En cada línea hay una cuenta. Entonces, por ejemplo, el texto 0Fundaciones queda de la siguiente manera:

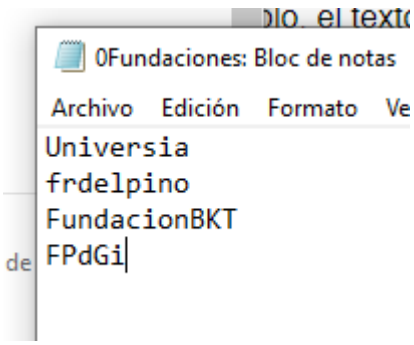

Figura 3.11 Ventana del archivo de texto .txt tipo 0

- **.txt tipo 1**

El archivo será un .csv que tiene

Poniendo como ejemplo la extracción de la cuenta @JoveJuristaCat, el archivo extraído queda de la siguiente manera:

1SD_jovejuristacat_TAG_Conjunto_inicial: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
Follower_1_Previo,Screen_Name,Cantidad,Friends_Count,Followers_Count
jovejuristacat,daniromeronet,1,243,59
jovejuristacat,exustion1,1,3193,790
jovejuristacat,_MRodriguezB_,1,109,352
jovejuristacat,nadiapeli01,1,50,10
jovejuristacat,JavierGArrocha,1,692,24
jovejuristacat,garmirbcn,1,3674,621
jovejuristacat,MrNoRelevant,1,891,716
jovejuristacat,JuanLuisCosta,1,2330,703
jovejuristacat,legaltages,1,262,69
jovejuristacat,essobreperros,1,35,1868
jovejuristacat,Jos07520204,1,662,25
jovejuristacat,MariaMedianero,1,319,349
jovejuristacat,m_ROSA_pb,1,467,95
jovejuristacat,jpalausa,1,424,23
jovejuristacat,Carles_GR,1,258,209
jovejuristacat,SSANSKRS,1,497,220
jovejuristacat,EsterPM15,1,227,117
jovejuristacat,toktokrly,1,521,243
jovejuristacat,ideadesarrollo,1,3317,100
jovejuristacat,angelicadarias,1,212,114
jovejuristacat,Miska_Lonette,1,436,364
jovejuristacat,ldonascimentof,1,368,193
jovejuristacat,Zencico,1,514,93
jovejuristacat,EconoiurisES,1,604,81
jovejuristacat,b_marn,1,115,63
jovejuristacat,BERNARDOGARCAS3,1,111,2
jovejuristacat,PEDROLUCO,1,895,128
```

Figura 3.12 Ventana del archivo de texto .txt tipo 1

Sin embargo, la selección de estos datos y la exclusión del resto tiene una explicación que se encuentra a continuación.

Cuando se extraen todas las variables y datos posibles, el resultado es el siguiente:

JoveJuristaCatAll followers: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
id,Name,Statuses_Count,Friends_Count,Screen_Name,Followers_Count,Location,Language,Created at,Time_zone,Geo_enable,Description
955450316,Cristina Tapial,485,501,CristinaTapiial,408,,,2012-11-18 13:01:53,,False,
300408285,Gisela,21,80,ggisela_ortiz,17,,,2011-05-17 18:23:18,,False,Abogada.                Defensora de los derechos .
245324728,Balms Abogados,5269,1981,balmsabogados,3720,España,,2011-01-31 12:31:35,,False,"Balms Abogados se fundó en 1989, avalad
1656332221,Nazaret,7994,1476,nazaret_gt,313,Ursi,,2013-08-08 22:15:29,,True,Derecho y Estudios Internacionales en la UC3M. La vid
1363777513,Antonio Salceda D.,3105,12989,ASalcedaD,16005,,,2013-04-19 06:53:02,,True,Abogado del @icam_es. Director de Salceda&Ab
937455210150154243,Pedro Trujillano Carci,174,128,PedroTrujillano,13,España,,2017-12-03 22:55:13,,False,Intelligence kills all.
1229838229177880576,Letrada Enfurecida,30,190,letradax,36,,,2020-02-18 18:41:09,,False,"Letrada progresista y protestona. Un buen
1553885881,Ana Zua²,4104,533,anazua14,245,Marbella,,2013-06-28 19:36:57,,False,Kill them with kindness❂À la folie
2885490377,ALMU :),626,1326,Almu_12,556,,,2014-11-20 10:55:13,,False,
280765970,#SonorenseSoy,8481,1825,Malobo_1980,569,"La Atravezada, Son.",,2011-04-12 00:15:27,,True,"LAP. Lic. en Derecho, Soci
628253875,Iván Ojeda Legaza,10327,2531,IvanLegaza,1035,Entre dos islas: GC TF ,,2012-07-06 10:19:36,,True,Graduado en Derecho
704144073419001858,Pachu,1186,453,pachu_3112,180,,,2016-02-29 03:20:04,,False,
706329276321239040,SARAH.,5400,258,scrl21,106,"Tunja, Colombia",,2016-03-06 04:03:17,,False,
2933035821,Jorge Viñuales,11,99,jvinaules,8,,,2014-12-20 03:57:25,,False,
1029500230373462016,@Magy,23,366,Magy93369297,18,Ixtlahuacán de los membrillos,,2018-08-14 22:49:16,,True,Linda y risueña
623736426,L a u r a,13458,320,lauratusuerte,1126,"Granada, España",,2012-07-01 11:19:56,,True, Derecho UGR
2280736346,Juanjo García Amorós,2542,867,JGarciaAmoros,672,España,,2014-01-07 15:17:53,,True,"Derecho y ADE -
Debate  -
Comunicación Política #ComPol"
708830915078672384,Giovanni Soto Córdova,552,5001,lic_gsc_1,228,"Baja California, México",,2016-03-13 01:43:54,,False,Prepárate p
948566185,Miguel Pasquau Liaño,18725,3555,miguelpasquau,23482,"Granada, España",,2012-11-14 21:49:25,,False,"Jurista (magistrado
https://t.co/l5WWLgkoSP"
903108124143480832,Ronisson P Silva,15116,3240,Ronissonps,255,"Roraima, Brasil",,2017-08-31 04:12:10,,True,Oppholdsvaer yakamoz ε
```

Figura 3.13 Ventana del archivo de texto .txt tipo 1 antes de seleccionar datos útiles

Este archivo es una extracción de datos de cada seguidor. Este archivo es .csv, es decir, que es como una tabla donde la separación de cada columna es una coma, y de cada fila es un "intro".

Los datos que pueden encontrarse en este archivo son los siguientes:

- Follower_1_previo

Es el nombre de usuario (Screen_Name) que tiene la cuenta de la que se han extraído la lista de seguidores. Esto no viene dado por Tweepy, sino que se ha añadido para más adelante hacer el análisis de relaciones.

- id

Es el número que Twitter le ha asignado al usuario. Es como una matrícula, y sirve para localizarlo. Puede ser útil para programas de extracción de datos.

- Name

Es el nombre que el propio usuario se ha puesto.

- Statuses_Count

Es el número de publicaciones y retweets que ha hecho la cuenta.

- Friends_Count

Es el número de amigos/seguidos, es decir, número de cuentas a las que sigue este usuario

- Screen_Name

Es el nombre de usuario con el que puede localizarse esta cuenta. Tiene la misma función que el id, pero es más fácil de recordar y escribir para el ser humano.

- Followers_Count

Es el número de seguidores que tiene esta cuenta.

- Location

Es el lugar donde el usuario ha escrito que se ubica.

- Language

Es el idioma con el que el usuario tiene su cuenta. Se supone que es su lengua materna.

- Created at

Es el momento en el que el usuario se creó la cuenta de Twitter.

- Time_zone

Huso horario del usuario.

- Geo_enable

Disponibilidad de ubicación geográfica

- Description

Es el resumen sobre sí mismo que cada usuario ha escrito.

De aquí se puede extraer información valiosa a través del análisis semántico.

Al hacer la extracción de datos e introducirlos en el software Rapidminer, puede observarse que hay muchos campos que se pueden dejar extraer porque están mayoritariamente en blanco, o dan errores.

De esta manera se consiguen archivos menos pesados y se eliminan datos que ni se van a necesitar, ni se pueden utilizar.

Para ello, a través del RapidMiner se analiza estadísticamente el número de elementos que hay en la muestra de la cuenta @JoveJuristaCat

Se importan los datos, se crea el esquema que hay a continuación y se miran las estadísticas. En este caso concreto, se han importado 10.000 filas.

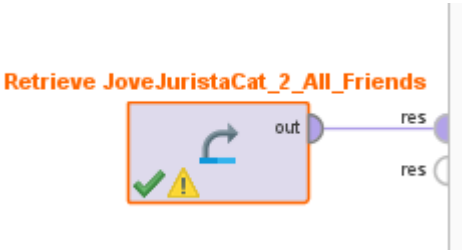

Figura 3.14 Esquema de bloques en Rapidminer

Los resultados son los siguientes:

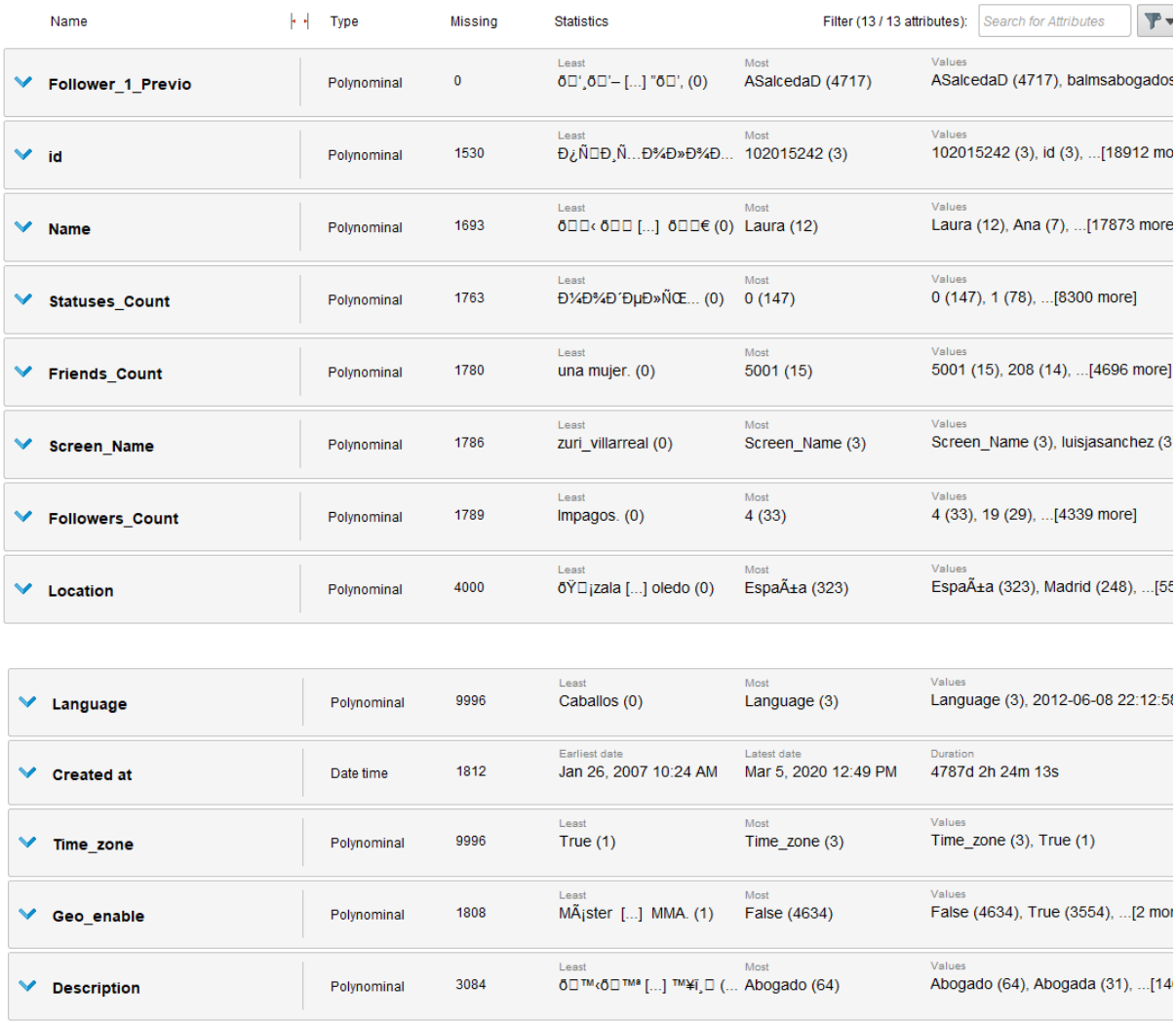

Figura 3.15 Resultados resumidos en Rapidminer

Estos datos pueden representarse en una tabla, teniendo únicamente en cuenta los "missings":

Follower_id	id	Name	Statuses_Count	Friends_Count	Screen_Name	Followers_Count
0	1530	1693	1763	1780	1786	1789
Location	Language	Created at	Time_zone	Geo_enable	Description	
4000	9996	1812	9996	1808	3084	

De estos datos, puede concluirse lo siguiente:

- Hay errores en la extracción de datos, dado que en la primera columna hay datos que no tienen id, por lo que hay un error en el documento. Esto se debe a que en la descripción, que es la última columna, de varios perfiles hay intros. Esto altera la naturaleza del documento, y puede llevar a error.
- Hay otros datos que los usuarios no suelen hacer público. En este caso destacan el idioma y la zona temporal. Al haber un número tan elevado de "missings" se podría ver la realidad infrarepresentada. Por esta misma razón, estos valores se dejan de extraer.
- El valor Geo_enable no nos proporciona ninguna información valiosa. Es un valor booleano. Por lo tanto, lo descartamos.
- La id podría ser de utilidad, pero el dato Screen_Name sirve igualmente para identificar cuentas. Entonces, se dejará de extraer este dato.
- El valor Created_at representa el momento en el que se creó la cuenta de Twitter. Podría tener alguna utilidad, como clasificar las cuentas más recientes en el caso que sean empresas de reciente creación. Sin embargo, en en este caso concreto, vamos a prescindir de este dato.
- Los datos Screen_Name, Friends_Count y Followers_Count son imprescindibles para analizar datos. Además de estar presente en todos los casos, y si no lo está es por un error relacionado con los intros. Entonces, se mantienen en la extracción 1 (followers).
- El dato Statuses_Count es el número de publicaciones. Para un análisis semántico, puede ser útil saber qué usuarios han estado más activos. Por lo tanto, se mantiene el dato Statuses_Count para la extracción 1 (followers)
- La descripción, es una de las herramientas más importantes de este estudio. Esto es porque gracias a ello se pueden realizar análisis semánticos y, con ello, descubrir más en profundidad está detrás de estas cuentas. Sin embargo, en este caso se obvia para evitar los errores de guardado, que imposibilitan el uso de estos datos cuando son guardados en formato .csv.
- Se necesita añadir una columna llamada Cantidad que tenga valor 1 en todas las lineas para, cuando se haga el minado de datos, simular transacciones con 1 producto.

Además de esto, se puede observar que en Rapidminer hay un problema que limita este estudio:

- Rapidminer no deja trabajar con más de 10.000 celdas con esta cuenta. Por lo tanto, para posibilitar este trabajo, se necesita eliminar el máximo número de cuentas posible que sean irrelevantes.

Entonces, se procederá a lo siguiente:

- Se mantiene la columna Follower_1_Previo
- Vamos a extraer únicamente los valores Screen_Name, Followers_Count, Friends_Count y Followers_Count
- Se añade la columna Cantidad, con un valor de 1 constante en todos los casos.

Los datos resultantes se pueden observar en esta imagen:

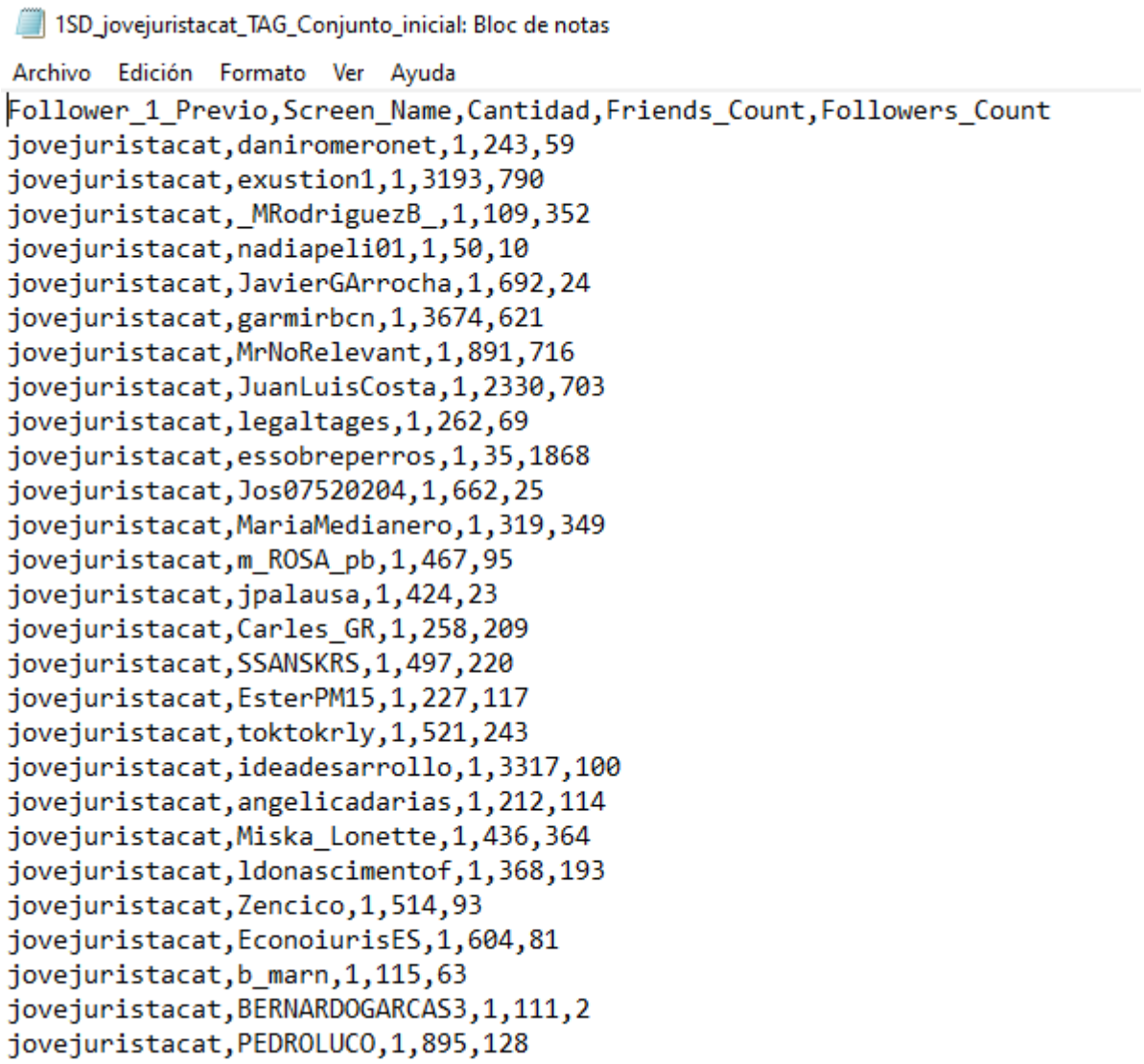

Figura 3.16 Ventana del archivo de texto .txt tipo 1 preparado para su uso

Y el programa creado para esta función es el siguiente:

```
import tweepy
import pandas as pd
```

```
nombre_lista = "Conjunto_inicial"
#Este nombre se sustituye por el de cada una de las listas de cuentas.
```

Se crea una función para llamar a la API con las claves que ha proporcionado Twitter

```
def lookup_user_list(user_id_list, api):
    full_users = []
    users_count = len(user_id_list)
    try:
        for i in range((users_count // 100) + 1):
            print (i)
            full_users.extend(api.lookup_users(user_ids=user_id_list[i * 100:min((i + 1) * 100, users_count)]))
        return full_users
    except tweepy.TweepError:
        print('Something went wrong, quitting...')

consumer_key = "exZu6K0OS05W9M5H7sYHXdKO1"
consumer_secret = "GMU6ZAZ7oESyivYlcg5Xb1WSKsBTq9aLBHXSMjrskmmInMEhti"
access_token = "3207578585-mQKA8QpDB1tRFyCgN0lqAnHDuQNtEJblvPDu1te"
access_token_secret = "nDW6hUTlZhIhBJyNwmpy2Uf74Ck37WeOY8GXLopfGDw6k"

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth, wait_on_rate_limit=True, wait_on_rate_limit_notify=True)
```

```
n=0
pre_lista = open("0"+nombre_lista+".txt", "r")
longitud_lista=len(pre_lista.readlines())   # longitud de la lista
pre_lista.close()
pre_lista = open("0"+nombre_lista+".txt", "r")
print(longitud_lista)

while n<longitud_lista:
    pre_usuario=pre_lista.readline()
    usuario=pre_usuario[:len(pre_usuario)-1]
    print(usuario)
    n=n+1
    #programa para extraer el csv 1, lista followers
    ids = []
    for page in tweepy.Cursor(api.followers_ids, screen_name=usuario).pages():
        ids.extend(page)
    results = lookup_user_list(ids, api)
    all_users = [{'Follower_1_Previo': usuario,
                  'Screen_Name': user.screen_name,
                  'Cantidad': '1',
#                 'id': user.id,
#                 'Name': user.name,
#                 'Statuses_Count': user.statuses_count,
                   'Friends_Count': user.friends_count,
                   'Followers_Count': user.followers_count,
#                 'Location': user.location,
#                 'Language': user.lang,
#                 'Created_at': user.created_at,
#                 'Time_zone': user.time_zone,
#                 'Geo_enable': user.geo_enabled,
#                  'Description': descripcion,
                  } for user in results]

    df = pd.DataFrame(all_users)

    #aquí se guardan los seguidores extraídos a un archivo para cada cuenta, con su etiqueta "TAG"
    df.to_csv('1SD_'+usuario+'_TAG_'+nombre_lista+'.csv', index=False, encoding='utf-8')

    #aquí se agregan al mismo archivo todas las cuentas extraídas
    df.to_csv('2SD_Conjunto_final.csv', mode='a', index=False, header=False, encoding='utf-8')

    print('archivo 1 guardado')
```

4 Minado de datos y creación del mapa de calor

4.1 Market Basket Case

Un análisis de Market Basket Case consiste en crear modelos para prever qué productos están los usuarios más dispuestos a comprar según compras anteriores. [12]

Por ejemplo, Amazon utiliza esta técnica cuando añades al carrito de la compra un producto y te sugiere otros.

Por ejemplo, cuando seleccionamos el libro Harry Potter y la Piedra Filosofal, nos aparece la sugerencia de comprar los 3 libros de la trilogía.

Figura 4.1 Interfaz de Amazon

Figura 4.2 Interfaz de Amazon

Y una vez has añadido al carrito el libro, te sugiere otros productos que no son libros de la saga.

Figura 4.3 Interfaz de Amazon

En este caso, a través de históricos de compra, Amazon ha hecho su Basket Case Analysis y ha concluído que, si compras el libro de Harry Potter, es posible que quieras comprarte la barita, la camiseta, la figura y/o un libro de una bruja.

De esta manera, suponiendo que un 5% de los compradores de libros de Harry Potter compra la barita de 12€, y suponiendo 1000 compradores al mes:

Libros de Harry Potter vendidos	Porcentaje de conversión	Precio del complemento	Aumento de la facturación	Margen bruto Amazon (30%)
1.000	5%	12,04€	602€	180,60€
10.000	5%	12,04€	6.020€	1.806€

En este caso, Amazon puede aumentar su margen bruto en 1.806€ cada 10.000 libros.

Otro ejemplo que muestra el poder de esta técnica es en la compra de lector de libros electrónico Kindle Paperwhite.

Figura 4.4 Interfaz de Amazon

Cuando hacemos click a Añadir a la cesta, nos aparecen las siguientes sugerencias:

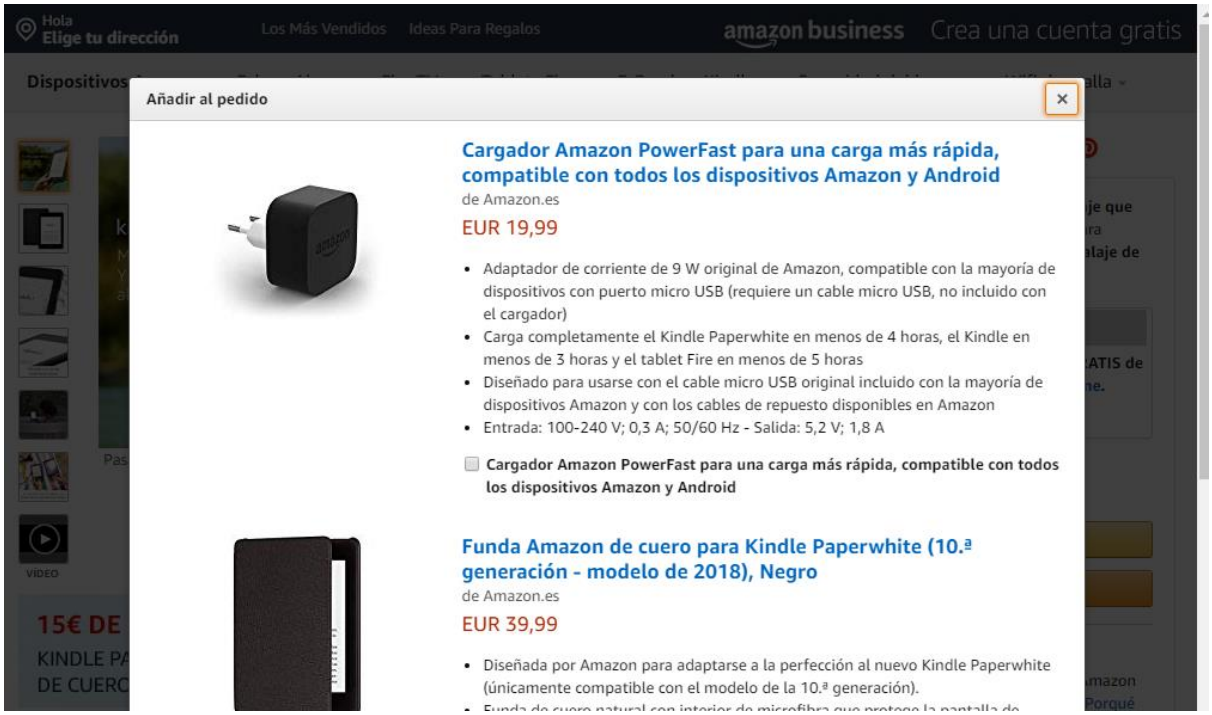

Figura 4.5 Interfaz de Amazon

En el caso de Amazon, utilizando métricas y analizando qué productos se han vendido más, han creado sus propias gamas de productos. En este caso, el cargador Amazon Powerfast y la funda Amazon de cuero, hacen que los márgenes de beneficio aumenten mucho más.

Suponiendo que un 10% de los compradores del Paperwhite compra la funda, y un 5% compra el cargador, y suponiendo también que al ser producción propia tienen un margen bruto por producto del 80%, se puede observar en la siguiente tabla los potenciales márgenes para la empresa:

Ereaders vendidos	Porcentaje de conversión	Precio del complemento	Aumento de la facturación	Margen bruto Amazon (80%)
1.000	10%	39,99€	3.999€	3.199,20€
1.000	5%	19,99€	999,50€	799,6€

Margen bruto de Amazon aumentado: 3.998,80€

Como puede observarse, este tipo de prácticas puede aumentar unos 4.000€ el margen por cada 1.000 Ereaders vendidos. A escala mundial y de forma automatizada, esto implica beneficios multimillonarios.

En el caso de Twitter, al igual que otras redes sociales, la propia plataforma te sugiere cuentas a seguir según lo que has seguido previamente, basándose en la afinidad de otros usuarios.

Lo que ocurre con Twitter es que estos análisis los hace automáticamente, y no ofrece datos sobre ello. Para analizar una cuenta en concreto, o un conjunto similar de cuentas, no es suficiente con ver sugerencias, sino que hay que crear un Basket Case propio.

En este caso, se hará una extrapolación entre los análisis de ventas de productos a cuentas seguidas. El paralelismo es el siguiente:

Amazon	Twitter
Producto comprado	Cuenta seguida
Conversión (%)	Afinidad (%)

Entonces, se interpretará la afinidad de un conjunto de cuentas para seguir a otras según la confianza.

4.2 Funcionamiento del algoritmo del Market Basket Case

Los pasos a seguir del software programado para el Market Basket Case son los siguientes:

- Lectura de los datos de 2Conjunto_final
- Tokenización de los datos, es decir, construir una tabla con valores 1 y 0 según se produzca o no el seguimiento de la cuenta.
- Se aplica el algoritmo Apriori para obtener las relaciones entre transacciones.
- Se aplica la función "association_rules" para obtener la información necesaria que será interpretada
- A partir de estas reglas y datos estadísticos obtenidos, se crean las matrices y se guardan cada una en un .csv
- Se importan los .csv a Excel, se ordenan, se les añade el número de seguidores y se les añade un formato condicional que es un mapa de calor.

El código programado y explicado en Jupyter Notebook (Python) ha sido el siguiente:

Minado de datos: Market Basket Case

```
import pandas as pd
import numpy as np
import seaborn
import matplotlib as plt
import matplotlib
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules

import seaborn as sns1
n=1
n_str=str(n)
```

```
df = pd.read_csv('2SD_Conjunto_final.csv')
df.head()
```

	Follower_1_Previo	Screen_Name	Cantidad	Friends_Count	Followers_Count
0	Universia	brendarichterar	1	627	295
1	Universia	alvarochaavez	1	266	289
2	Universia	FrkSer	1	38	5
3	Universia	IbersonAndree	1	80	17
4	Universia	brrrrendino	1	784	1056

Se tokenizan las "transacciones", es decir, se transforma cada acción de seguir en una tabla de unos y ceros.

```
basket = (df.groupby(['Screen_Name', 'Follower_1_Previo'])['Cantidad']
          .sum().unstack().reset_index().fillna(0)
          .set_index('Screen_Name'))
```

```
basket
```

Follower_1_Previo Screen_Name	AEGEEMalaga	AEGEE_Zaragoza	BESTMadridUPM	BESTUPC	BESTValenciaUPV	EFEspana	ELSA_Spain_
0001sa1000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
001Cris	0.0	0.0	0.0	0.0	0.0	0.0	0.0
001_carmen	0.0	0.0	0.0	0.0	0.0	0.0	0.0
001_miemi_	0.0	0.0	0.0	0.0	0.0	0.0	0.0
002jac	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...
zzamanmurad	0.0	0.0	0.0	0.0	0.0	0.0	0.0
zzocalo	0.0	0.0	0.0	0.0	0.0	0.0	0.0
zzste	0.0	0.0	0.0	0.0	0.0	0.0	0.0
zzzclara	0.0	0.0	0.0	0.0	0.0	0.0	0.0
zzzjulian	0.0	0.0	0.0	0.0	0.0	0.0	0.0

136861 rows × 26 columns

Esto sirve para que, en caso de haber un error como poner dos veces la misma cuenta, se transforme en 1

```
def encode_units(x):
    if x <= 0:
        return 0
    if x >= 1:
        return 1

basket_sets = basket.applymap(encode_units)
```

Se crea la Basket Case a través del algoritmo Apriori. Se limita a 2 elementos para poderse visualizar en una matriz más adelante, y se pone un soporte mínimo exageradamente bajo para que se representen todas las relaciones.

```
frequent_itemsets = apriori(basket_sets,min_support=0.00000001, use_colnames=True, max_len=2)
```

```
frequent_itemsets
```

	support	itemsets
0	0.000321	(AEGEEMalaga)
1	0.007270	(AEGEE_Zaragoza)
2	0.010880	(BESTMadridUPM)
3	0.006262	(BESTUPC)
4	0.004808	(BESTValenciaUPV)
...	...	...
323	0.000095	(teamlabs, jovejuristacat)
324	0.000066	(wearetrivu, jovejuristacat)
325	0.001812	(laakademia_org, teamlabs)
326	0.000621	(laakademia_org, wearetrivu)
327	0.002616	(wearetrivu, teamlabs)

328 rows × 2 columns

Ahora, con estas relaciones, se crean las reglas de asociación para obtenerse datos visualizables.

```
rules = association_rules(frequent_itemsets, metric="lift",  min_threshold=0)
df_rules = pd.DataFrame(rules)
rules.head()
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(AEGEE_Zaragoza)	(AEGEEMalaga)	0.007270	0.000321	0.000051	0.007035	21.882755	0.000049	1.006761
1	(AEGEEMalaga)	(AEGEE_Zaragoza)	0.000321	0.007270	0.000051	0.159091	21.882755	0.000049	1.180544
2	(BESTMadridUPM)	(AEGEEMalaga)	0.010880	0.000321	0.000007	0.000672	2.088971	0.000004	1.000350
3	(AEGEEMalaga)	(BESTMadridUPM)	0.000321	0.010880	0.000007	0.022727	2.088971	0.000004	1.012123
4	(AEGEEMalaga)	(BESTUPC)	0.000321	0.006262	0.000007	0.022727	3.629495	0.000005	1.016848

```
len(rules)
```

```
604
```

```
print(basket.shape)
print(rules.shape)
```

```
(136861, 26)
(604, 9)
```

Se seleccionan los datos que estarán en las matrices: Confidence, Support y Lift.

```
#selección confidence

nuevo_rules = rules
pre_matriz=nuevo_rules

#confidence en porcentaje
pre_matriz_confidence=nuevo_rules.iloc[:, [0,1,5]] # Columnas 1, 2 y confidence
pre_matriz_confidence.confidence=pre_matriz.confidence.multiply(100)
print(pre_matriz_confidence)

#support multiplicado por 1000
pre_matriz_support=nuevo_rules.iloc[:, [0,1,4]] # Columnas 1, 2 y confidence
len_basket=len(basket)
len_basket_str=str(len_basket)

pre_matriz_support.support=pre_matriz.support.multiply(1000)

print(pre_matriz_support)

pre_matriz_lift=nuevo_rules.iloc[:, [0,1,6]] # Columnas 1, 2 y confidence
print(pre_matriz_lift)
```

```
          antecedents       consequents  confidence
0    (AEGEE_Zaragoza)     (AEGEEMalaga)    0.703518
1       (AEGEEMalaga)  (AEGEE_Zaragoza)   15.909091
2     (BESTMadridUPM)     (AEGEEMalaga)    0.067159
3       (AEGEEMalaga)   (BESTMadridUPM)    2.272727
4       (AEGEEMalaga)         (BESTUPC)    2.272727
..                ...               ...         ...
599         (teamlabs)  (laakademia_org)    2.706833
600  (laakademia_org)      (wearetrivu)    1.101750
601      (wearetrivu)  (laakademia_org)    2.131394
602      (wearetrivu)        (teamlabs)    8.976931
603        (teamlabs)      (wearetrivu)    3.907444
```

Ahora se crean las matrices y se guarda cada una en un .csv

```
n_matriz=2
n_matriz_str=str(n_matriz)

matriz_confidence_nan=pre_matriz_confidence.pivot_table(columns='antecedents', index='consequents', va
matriz_confidence = matriz_confidence_nan.fillna(0)
matriz_confidence.to_csv('matriz_confidence'+n_matriz_str+'.csv', index=False, encoding='utf-8')

matriz_support_nan=pre_matriz_support.pivot_table(columns='antecedents', index='consequents', values='
matriz_support = matriz_support_nan.fillna(0)
matriz_support.to_csv('matriz_support'+n_matriz_str+'.csv', index=False, encoding='utf-8')

matriz_lift_nan=pre_matriz_lift.pivot_table(columns='antecedents', index='consequents', values='lift').
matriz_lift = matriz_lift_nan.fillna(0)
matriz_lift.to_csv('matriz_lift'+n_matriz_str+'.csv', index=False, encoding='utf-8')
```

```
matriz_confidence
```

antecedents	consequents	(AEGEEMalaga)	(aiesecspain)	(Universia)	(MondragonTA)	(FPdGi)	(ESNSpain)	(teamlabs)
0	(AEGEE_Zaragoza)	15.909091	0.504081	0.014049	0.086003	0.007221	0.873362	0.010915
1	(aiesecspain)	0.000000	0.000000	0.231807	1.462051	0.635425	2.125182	1.506221
2	(Universia)	0.000000	0.792127	0.000000	0.967534	0.852047	1.164483	1.102379
3	(MondragonTA)	0.000000	1.632261	0.316100	0.000000	2.375623	0.378457	22.538747
4	(FPdGi)	0.000000	2.112338	0.828885	7.073748	0.000000	1.921397	7.432875
5	(ESNSpain)	0.000000	1.752280	0.280978	0.279510	0.476569	0.000000	0.305610
6	(teamlabs)	2.272727	3.312530	0.709469	44.399054	4.917323	0.815138	0.000000
7	(jovejuristacat)	2.272727	0.096015	0.063220	0.086003	0.028883	0.058224	0.141890
8	(frdelpino)	2.272727	2.880461	0.913178	4.966674	4.411871	1.630277	6.472386

Los datos de salida del análisis relacional pueden verse en la siguiente imagen:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(aegeebarcelona)	(AEGEE_Zaragoza)	0.003103	0.010110	0.001683	0.542484	53.660721	0.001652	2.163618
1	(AEGEE_Zaragoza)	(aegeebarcelona)	0.010110	0.003103	0.001683	0.166499	53.660721	0.001652	1.196037
2	(BESTMadridUPM)	(BESTUPC)	0.015149	0.008761	0.002058	0.135877	15.509461	0.001926	1.147104
3	(BESTUPC)	(BESTMadridUPM)	0.008761	0.015149	0.002058	0.234954	15.509461	0.001926	1.287309
4	(BESTMadridUPM)	(BESTValenciaUPV)	0.015149	0.006672	0.001795	0.118474	17.756680	0.001694	1.126828
5	(BESTValenciaUPV)	(BESTMadridUPM)	0.006672	0.015149	0.001795	0.268997	17.756680	0.001694	1.347260
6	(BESTValenciaUPV)	(BESTUPC)	0.006672	0.008761	0.001683	0.252280	28.796085	0.001625	1.325682
7	(BESTUPC)	(BESTValenciaUPV)	0.008761	0.006672	0.001683	0.192130	28.796085	0.001625	1.229564
8	(_celera)	(frdelpino)	0.020919	0.117187	0.006104	0.291808	2.490102	0.003653	1.246573
9	(frdelpino)	(_celera)	0.117187	0.020919	0.006104	0.052090	2.490102	0.003653	1.032884

Figura 4.6 Datos de salida del análisis relacional

Los elementos que se encuentran en cada columna significan lo siguiente:

- Antecedente: cuenta anterior (aegeebarcelona en la 1ª línea)
- Consecuente: cuenta posterior.(AEGEE_Zaragoza en la 1ª línea)

Soporte: es la proporción del conjunto de datos donde la función no es cero. En este caso es, del total de transacciones, las que incluyen la cuenta. El soporte se calcula dividiendo la cantidad de veces que aparece la cuenta entre el número total de transacciones.

El soporte nos ayuda a visualizar qué cantidad de seguidores tienen las cuentas y relaciones en comparación con el total.

$$support(I) = \frac{Number\ of\ transactions\ containing\ I}{Total\ number\ of\ transactions}$$

En el caso de la primera interacción, teniendo en cuenta que:

- La cuenta de AEGEE_Zaragoza tiene 996 seguidores
- La cuenta de aegeebarcelona tiene 306 seguidores
- El número total de transacciones es de 98620

El soporte de AEGEE_Zaragoza es:

996/98620 = 0,01011 = 1,011%

El soporte de aegeebarcelona es:

306/98620 = 0,0031 = 0,31%

El soporte de las transacciones es de 0,001683. Esto significa que el número de transacciones es:

0,001683 · 98620 = 166

Por ello, Aegee Zaragoza y Aegee Barcelona tienen 166 seguidores en común.

Confianza: es la probabilidad de que los seguidores de la cuenta antecedente sigan a la consecuente. En este caso, de que los seguidores de Aegee Barcelona sigan a Aegee Zaragoza.

Este valor se obtiene dividiendo el número de seguidores en común entre el número de seguidores totales de la cuenta antecedente.

El cálculo es

$$confidence(I1 \rightarrow I2) = \frac{Number\ of\ transactions\ containing\ items\ I1\ and\ I2}{Total\ number\ of\ transactions\ containing\ I1}$$

En este caso, la probabilidad de que un seguidor de Aegee Barcelona siga a la cuenta de Aegee Zaragoza es de 0,542484 , es decir, el 54,25%.

Sin embargo, en la siguiente “transacción” puede observarse que la probabilidad de que un seguidor de Aegee Zaragoza siga a Aegee Barcelona es del 0,1665 , un 16,65%.

Interpretando estos datos, se crea la hipótesis de que la cuenta de Aegee Zaragoza tiene más seguidores que no forman parte de la organización y que, por tanto, no interactúan con otras “antenas” de otras ciudades. También se supone que los seguidores de la cuenta de Barcelona son más de la mitad de la misma asociación Aegee.

El parámetro lift es la relación entre la confianza y el soporte de la cuenta consecuente.

Este parámetro indica el desempeño del modelo, por lo que a mayor lift, mejor se estará comportando.

$$lift(I1 \rightarrow I2) = \frac{confidence(I1 \rightarrow I2)}{support(I2)}$$

La información extraída se visualiza a través de una tabla de relaciones, que también se mostrará como un mapa de calor.

Un mapa de calor ("*heatmap*" en inglés) es la manera de mostrar una tabla de datos asignando un color diferente según el número que contenga cada celda. De esta manera se pueden interpretar los datos visualmente de una forma más completa y sencilla.

Los mapas de calor del primer análisis relacional de las cuentas son los siguientes:

Confianza
Número absoluto de cuentas
Lift

- Heatmap de confianza

El heatmap de la confianza de que un seguidor de una cuenta siga a otra es el siguiente:

Título: Heatmap de confianza (%) entre seguir la cuenta antecedente y la consecuente

\ antecedentes consecuentes	_celera	E_Zaragoza	aegeebarcelona	AEGEEMalaga	aiesecspain	MadridUPM	BESTUPC	ValenciaUPV	EFEspana	ELSA_Spain_	ESNSpain	factoriatalent0	FPdGi	frdelpino	FundacionBKT	IAESTE_Spain	jovejuristacat	laakademia_org	MondragonTA	ShapersMadrid	teamlabs	ThePowerMBA	Universia	UniversiaES	wearetrivu
-> Nº seguidores	2138	995	306	44	4166	1489	857	658	6303	391	3435	1282	13849	11970	13627	275	453	7715	4651	1203	9162	3249	14236	46723	3988
_celera	0	0.1	0.65	2.27	0.86	0.94	0.7	0.76	0.14	0.26	0.35	3.43	1.487	5.2	2.018	0.36	0.88	0.51	2.15	6.73	2.24	1.631	0.436	0.48	4.44
AEGEE_Zaragoza	0.05	0	53.3	15.9	0.5	0.4	0.58	0.76	0.16	1.28	0.87	0.31	0.007	0.06	0.022	0.73	0.22	0.03	0.09	0.17	0.01	0	0.014	0.22	0.1
aegeebarcelona	0.09	16.4	0	15.9	0.14	0.13	0.23	0.15	0.05	0.77	0.61	0.08	0.036	0.02	0.044	0.36	0.22	0.01	0.02	0.08	0.02	0	0.014	0.03	0.05
AEGEEMalaga	0.05	0.7	2.29	0	0	0.07	0.12	0	0.02	0.26	0	0.08	0	0.01	0.007	0.36	0.22	0	0	0.08	0.01	0	0	0.01	0.03
aiesecspain	1.68	2.11	1.96	0	0	1.14	0.82	0.76	0.83	2.56	2.13	3.35	0.635	1	0.69	2.55	0.88	0.44	1.46	2.08	1.51	0.369	0.232	0.65	2.51
BESTMadridUPM	0.65	0.6	0.65	2.27	0.41	0	23.3	26.6	0.24	0.26	0.61	2.26	0.195	0.43	0.264	2.55	0.22	0	0.09	0.91	0.22	0.185	0.126	0.52	0.55
BESTUPC	0.28	0.5	0.65	2.27	0.17	13.4	0	25.1	0.1	0.77	0.29	0.23	0.217	0.08	0.11	1.09	0.22	0.03	0.13	0.33	0.14	0	0.042	0.12	0.35
BESTValenciaUPV	0.23	0.5	0.33	0	0.12	11.8	19.3	0	0.11	0	0.32	0	0.101	0.05	0.154	0.36	0	0.06	0.13	0.25	0.08	0	0.014	0.16	0.13
EFEspana	0.42	1.01	0.98	2.27	1.25	1.01	0.7	1.06	0	0.51	0.7	0.86	0.477	0.49	0.338	1.09	0.44	0.32	0.39	0.33	0.36	0.4	0.197	0.82	0.8
ELSA_Spain_	0.05	0.5	0.98	2.27	0.24	0.07	0.35	0	0.03	0	0.41	0.47	0.029	0.16	0.095	0.36	2.65	0	0.04	0.33	0.05	0.062	0.056	0.05	0.23
ESNSpain	0.56	3.02	6.86	0	1.75	1.41	1.17	1.67	0.38	3.58	0	0.94	0.477	0.47	0.389	2.55	0.44	0.1	0.28	0.42	0.31	0.062	0.281	0.74	0.48
factoriatalent0	2.06	0.4	0.33	2.27	1.03	1.95	0.35	0	0.17	1.53	0.35	0	0.477	0.82	0.66	0.73	0.44	0.34	0.69	1.5	0.83	0.616	0.176	0.45	2.66
FPdGi	9.64	0.1	1.63	0	2.11	1.81	3.5	2.13	1.05	1.02	1.92	5.15	0	5.1	5.078	1.82	0.88	3.21	7.07	6.65	7.43	1.908	0.829	1.67	6.27
frdelpino	29.1	0.7	0.65	2.27	2.88	3.49	1.05	0.91	0.94	4.86	1.63	7.64	4.412	0	7.757	2.18	3.75	1.58	4.97	12.3	6.47	3.355	0.913	1.86	13.4
FundacionBKT	12.9	0.3	1.96	2.27	2.26	2.42	1.75	3.19	0.73	3.32	1.54	7.02	4.997	8.83	0	0.73	2.21	2.33	8.08	6.9	8.5	4.094	2.185	1.88	8.4
IAESTE_Spain	0.05	0.2	0.33	2.27	0.17	0.47	0.35	0.15	0.05	0.26	0.2	0.16	0.036	0.05	0.015	0	0.22	0	0	0.08	0.02	0	0.014	0.04	0.08
jovejuristacat	0.19	0.1	0.33	2.27	0.1	0.07	0.12	0	0.03	3.07	0.06	0.16	0.029	0.14	0.073	0.36	0	0.05	0.09	0.08	0.14	0.092	0.063	0.04	0.23
laakademia_org	1.82	0.2	0.33	0	0.82	0	0.23	0.76	0.4	0	0.23	2.03	1.791	1.02	1.321	0	0.88	0	2.58	1.08	2.71	0.831	0.288	0.5	2.13
MondragonTA	4.68	0.4	0.33	0	1.63	0.27	0.7	0.91	0.29	0.51	0.38	2.5	2.376	1.93	2.759	0	0.88	1.56	0	4.07	22.5	1.231	0.316	0.53	3.74
ShapersMadrid	3.79	0.2	0.33	2.27	0.6	0.74	0.47	0.46	0.06	1.02	0.15	1.4	0.578	1.24	0.609	0.36	0.22	0.17	1.05	0	0.97	0.246	0.042	0.14	1.83
teamlabs	9.59	0.1	0.65	2.27	3.31	1.34	1.52	1.06	0.52	1.28	0.82	5.93	4.917	4.95	5.717	0.73	2.87	3.21	44.4	7.4	0	3.509	0.709	1.22	8.98
ThePowerMBA	2.48	0	0	0	0.29	0.4	0	0	0.21	0.51	0.06	1.56	0.448	0.91	0.976	0	0.66	0.35	0.86	0.67	1.24	0	0.253	0.22	2.23
Universia	2.9	0.2	0.65	0	0.79	1.21	0.7	0.3	0.44	2.05	1.16	1.95	0.852	1.09	2.282	0.73	1.99	0.53	0.97	0.5	1.1	1.108	0	1.67	1.45
UniversiaES	10.6	10.4	4.58	9.09	7.25	16.2	6.53	11.2	6.08	6.14	10	16.3	5.625	7.24	6.443	6.91	3.97	3.02	5.33	5.49	6.24	3.17	5.472	0	8.4
wearetrivu	8.28	0.4	0.65	2.27	2.4	1.48	1.63	0.76	0.51	2.3	0.55	8.27	1.805	4.48	2.458	1.09	1.99	1.1	3.2	6.07	3.91	2.739	0.407	0.72	0

Figura 4.7: Heatmap de confianza entre seguidores de cuentas en %

Eje X: Cuenta antecedente. Eje Y: Cuenta consecuente

- Heatmap de número absoluto de cuentas

El heatmap del soporte es muy confuso porque al ser porcentajes tan bajos, apenas se puede distinguir entre relaciones. Por ello, se ha multiplicado la matriz por el número total de transacciones y se ha obtenido el número absoluto de relaciones.

Título: heatmap de lista de consecuentes, número de cuentas en común

\ antecedentes	_celera	E_Zaragoza	aegeebarcelona	AEGEEMalaga	aiesecspain	MadridUPM	BESTUPC	ValenciaUPV	EFEspana	ELSA_Spain_	ESNSpain	factoriatalent0	FPdGi	frdelpino	FundacionBKT	IAESTE_Spain	jovejuristacat	laakademia_org	MondragonTA	ShapersMadrid	teamlabs	ThePowerMBA	Universia	UniversiaES	wearetrivu
consecuentes ->	2138	995	306	44	4166	1489	857	658	6303	391	3435	1282	13849	#####	13627	275	453	7715	4651	1203	9162	3249	14236	#####	3988
_celera	1	2	1	1	7	7	3	1	3	1	7	2	5	6	2	0	1	0	0	1	2	0	2	19	3
AEGEE_Zaragoza	0	1	2	1	36	14	6	5	9	1	12	44	206	622	275	1	4	39	100	81	205	53	62	226	177
aegeebarcelona	1	0	163	7	21	6	5	5	10	5	30	4	1	7	3	2	1	2	4	2	1	0	2	103	4
AEGEEMalaga	2	163	0	7	6	2	2	1	3	3	21	1	5	2	6	1	1	1	1	1	2	0	2	14	2
aiesecspain	1	7	7	0	0	1	1	0	1	1	0	1	0	1	1	1	1	0	0	1	1	0	0	4	1
BESTMadridUPM	36	21	6	0	0	17	7	5	52	10	73	43	88	120	94	7	4	34	68	25	138	12	33	302	100
BESTUPC	14	6	2	1	17	0	200	175	15	1	21	29	27	52	36	7	1	0	4	11	20	6	18	241	22
BESTValenciaUPV	6	5	2	1	7	200	0	165	6	3	10	3	30	9	15	3	1	2	6	4	13	0	6	56	14
EFEspana	12	30	21	0	73	21	10	11	24	14	0	12	66	56	53	7	2	8	13	5	28	2	40	344	19
ELSA_Spain_	5	5	1	0	5	175	165	0	7	0	11	0	14	6	21	1	0	5	6	3	7	0	2	74	5
ESNSpain	9	10	3	1	52	15	6	7	0	2	24	11	66	59	46	3	2	25	18	4	33	13	28	383	32
factoriatalent0	44	4	1	1	43	29	3	0	11	6	12	0	66	98	90	2	2	26	32	18	76	20	25	209	106
FPdGi	206	1	5	0	88	27	30	14	66	4	66	66	0	611	692	5	4	248	329	80	681	62	118	779	250
frdelpino	622	7	2	1	120	52	9	6	59	19	56	98	611	0	1057	6	17	122	231	148	593	109	130	867	536
FundacionBKT	275	3	6	1	94	36	15	21	46	13	53	90	692	1057	0	2	10	180	376	83	779	133	311	878	335
IAESTE_Spain	1	5	3	1	10	1	3	0	2	0	14	6	4	19	13	1	12	0	2	4	5	2	8	24	9
jovejuristacat	4	1	1	1	4	1	1	0	2	12	2	2	4	17	10	1	0	4	4	1	13	3	9	18	9
laakademia_org	39	2	1	0	34	0	2	5	25	0	8	26	248	122	180	0	4	0	120	13	248	27	41	233	85
MondragonTA	100	4	1	0	68	4	6	6	18	2	13	32	329	231	376	0	4	120	0	49	2065	40	45	248	149
ShapersMadrid	81	2	1	1	25	11	4	3	4	4	5	18	80	148	83	1	1	13	49	0	89	8	6	66	73
teamlabs	205	1	2	1	138	20	13	7	33	5	28	76	681	593	779	2	13	248	2065	89	0	114	101	572	358
ThePowerMBA	53	0	0	0	12	6	0	0	13	2	2	20	62	109	133	0	3	27	40	8	114	0	36	103	89
Universia	62	2	2	0	33	18	6	2	28	8	40	25	118	130	311	2	9	41	45	6	101	36	0	779	58
UniversiaES	226	103	14	4	302	241	56	74	383	24	344	209	779	867	878	19	18	233	248	66	572	103	779	0	335
wearetrivu	177	4	2	1	100	22	14	5	32	9	19	106	250	536	335	3	9	85	149	73	358	89	58	335	0

Figura 4.8: Heatmap de número absoluto de cuentas en común

Eje X: Cuenta antecedente. Eje Y: Cuenta consecuente

Y en el caso del lift, el mapa de calor es el siguiente:

												Lift												
\ antecedentes	EFEspana	E_Zaragoza	aegeebarcelona	AEGEEMalaga	aiesecspain	MadridUPM	BESTUPC	ValenciaUPV	ELSA_Spain_	IAESTE_Spain	ESNSpain	factoriatalent0	FPdGi	frdelpino	FundacionBKT	jovejuristacat	laakademia_org	_celera	MondragonTA	teamlabs	ThePowerMBA	Universia	UniversiaES	ShapersMadrid	wearetrivu
consecuentes -> I	6303	995	306	44	4166	1489	857	658	391	275	3435	1282	13849	#####	13627	453	7715	2138	4651	9162	3249	14236	46723	1203	3988
_celera	0.09	0.06	0.42	1.45	0.55	0.6	0.45	0.48	0.16	0.23	0.22	2.19	0.947	3.31	1.285	0.56	0.32	0	1.37	1.42	1.04	0.277	0.308	4.29	2.83
AEGEE_Zaragoza	0.22	0	72.9	21.8	0.69	0.55	0.8	1.04	1.75	1	1.2	0.43	0.01	0.08	0.03	0.3	0.04	0.06	0.12	0.01	0	0.019	0.302	0.23	0.14
aegeebarcelona	0.21	72.9	0	70.8	0.64	0.6	1.04	0.68	3.41	1.62	2.72	0.35	0.161	0.07	0.196	0.98	0.06	0.42	0.1	0.1	0	0.063	0.133	0.37	0.22
AEGEEMalaga	0.49	21.8	70.8	0	0	2.08	3.61	0	7.91	11.3	0	2.41	0	0.26	0.227	6.83	0	1.45	0	0.34	0	0	0.265	2.57	0.78
aiesecspain	0.27	0.69	0.64	0	0	0.37	0.27	0.25	0.84	0.83	0.69	1.1	0.208	0.33	0.225	0.29	0.14	0.55	0.48	0.49	0.12	0.076	0.211	0.68	0.82
BESTMadridUPM	0.22	0.55	0.6	2.08	0.37	0	21.3	24.3	0.23	2.33	0.56	2.07	0.178	0.4	0.242	0.2	0	0.6	0.08	0.2	0.17	0.116	0.472	0.84	0.5
BESTUPC	0.15	0.8	1.04	3.61	0.27	21.3	0	39.8	1.22	1.73	0.46	0.37	0.344	0.12	0.175	0.35	0.04	0.45	0.2	0.23	0	0.067	0.19	0.53	0.56
BESTValenciaUPV	0.23	1.04	0.68	0	0.25	24.3	39.8	0	0	0.75	0.66	0	0.209	0.1	0.319	0	0.13	0.48	0.27	0.16	0	0.029	0.328	0.52	0.26
EFEspana	0	0.22	0.21	0.49	0.27	0.22	0.15	0.23	0.11	0.24	0.15	0.19	0.103	0.11	0.073	0.1	0.07	0.09	0.08	0.08	0.09	0.042	0.177	0.07	0.17
ELSA_Spain_	0.11	1.75	3.41	7.91	0.84	0.23	1.22	0	0	1.27	1.42	1.63	0.101	0.55	0.332	9.22	0	0.16	0.15	0.19	0.21	0.196	0.179	1.16	0.79
ESNSpain	0.15	1.2	2.72	0	0.69	0.56	0.46	0.66	1.42	1.01	0	0.37	0.189	0.19	0.154	0.17	0.04	0.22	0.11	0.12	0.02	0.111	0.292	0.16	0.19
factoriatalent0	0.19	0.43	0.35	2.41	1.1	2.07	0.37	0	1.63	0.77	0.37	0	0.506	0.87	0.701	0.47	0.36	2.19	0.73	0.88	0.65	0.186	0.475	1.59	2.82
FPdGi	0.1	0.01	0.16	0	0.21	0.18	0.34	0.21	0.1	0.18	0.19	0.51	0	0.5	0.499	0.09	0.32	0.95	0.7	0.73	0.19	0.081	0.164	0.65	0.62
frdelpino	0.11	0.08	0.07	0.26	0.33	0.4	0.12	0.1	0.55	0.25	0.19	0.87	0.502	0	0.882	0.43	0.18	3.31	0.56	0.74	0.38	0.104	0.211	1.4	1.53
FundacionBKT	0.07	0.03	0.2	0.23	0.23	0.24	0.17	0.32	0.33	0.07	0.15	0.7	0.499	0.88	0	0.22	0.23	1.29	0.81	0.85	0.41	0.218	0.188	0.69	0.84
IAESTE_Spain	0.24	1	1.62	11.3	0.83	2.33	1.73	0.75	1.27	0	1.01	0.77	0.179	0.25	0.073	1.09	0	0.23	0	0.11	0	0.07	0.201	0.41	0.37
jovejuristacat	0.1	0.3	0.98	6.83	0.29	0.2	0.35	0	9.22	1.09	0.17	0.47	0.087	0.43	0.221	0	0.16	0.56	0.26	0.43	0.28	0.19	0.116	0.25	0.68
laakademia_org	0.07	0.04	0.06	0	0.14	0	0.04	0.13	0	0	0.04	0.36	0.316	0.18	0.233	0.16	0	0.32	0.46	0.48	0.15	0.051	0.088	0.19	0.38
MondragonTA	0.08	0.12	0.1	0	0.48	0.08	0.2	0.27	0.15	0	0.11	0.73	0.695	0.56	0.808	0.26	0.46	1.37	0	6.6	0.36	0.093	0.155	1.19	1.09
ShapersMadrid	0.07	0.23	0.37	2.57	0.68	0.84	0.53	0.52	1.16	0.41	0.16	1.59	0.654	1.4	0.689	0.25	0.19	4.29	1.19	1.1	0.28	0.048	0.16	0	2.07
teamlabs	0.08	0.01	0.1	0.34	0.49	0.2	0.23	0.16	0.19	0.11	0.12	0.88	0.731	0.74	0.849	0.43	0.48	1.42	6.6	0	0.52	0.105	0.182	1.1	1.33
ThePowerMBA	0.09	0	0	0	0.12	0.17	0	0	0.21	0	0.02	0.65	0.188	0.38	0.409	0.28	0.15	1.04	0.36	0.52	0	0.106	0.092	0.28	0.94
Universia	0.04	0.02	0.06	0	0.08	0.12	0.07	0.03	0.2	0.07	0.11	0.19	0.081	0.1	0.218	0.19	0.05	0.28	0.09	0.11	0.11	0	0.159	0.05	0.14
UniversiaES	0.18	0.3	0.13	0.26	0.21	0.47	0.19	0.33	0.18	0.2	0.29	0.48	0.164	0.21	0.188	0.12	0.09	0.31	0.16	0.18	0.09	0.159	0	0.16	0.24
wearetrivu	0.17	0.14	0.22	0.78	0.82	0.5	0.56	0.26	0.79	0.37	0.19	2.82	0.616	1.53	0.839	0.68	0.38	2.83	1.09	1.33	0.94	0.139	0.245	2.07	0

Figura 4.9: Heatmap de número absoluto de cuentas en común

Eje X: Cuenta antecedente. Eje Y: Cuenta consecuente

5 Análisis de datos y propuestas a la toma de decisiones

5.1 Análisis específico de los seguidores de las cuentas

En todos los casos, excepto las cuentas con un número similar de seguidores, tienen diferencias importantes en el porcentaje de seguidores en común según si se mira las mismas dos cuentas como antecedente – consecuente, y como consecuente - antecedente. Esto se debe a la diferencia de seguidores, es decir, si una cuenta tiene 1000 seguidores, la otra 100 y tienen 50 seguidores en común, será el 5% para la primera y el 50% para la segunda.

Sin embargo, también se plantearán una serie de hipótesis para entender estas diferencias entre seguidores.

Las cuentas con un valor más verde, con una confianza muy alta, de entre el 22% y el 45% son de instituciones y sus respectivo proyectos o entidades dependientes de esta:

- Teamlabs y MondragonTA

Antecedente	Consecuente	Confianza
teamlabs	MondragonTA (Mondragón Team Academy)	23%
MondragonTA (Mondragón Team Academy)	teamlabs	44%

Teamlabs es una institución, y Mondragon Team Academy es una red/comunidad de los alumnos, alumni, coaches y colaboradores de programas de Teamlabs.

Mondragon Team Academy es conocido por los alumnos de Teamlabs y otros estudios de la universidad de Mondragón, pero los seguidores de Teamlabs no conocen o conocen mucho menos esta red.
Así, puede concluirse que Mondragon Team Academy tiene un público relacionado con los miembros de los estudios de Teamlabs.

- Celera (programa educativo) con Fundación Rafael del Pino

Antecedente	Consecuente	Confianza
_celera	frdelpino (Fundación Rafael del Pino)	29%
frdelpino (Fundación Rafael del Pino)	_celera	5%

Los seguidores de la Fundación Rafael del Pino que están interesados en el programa lo han seguido, aumentando así su proporción.

Sin embargo, al tener Celera una comunidad de seguidores 6 veces más pequeña que la de la fundación Rafael del Pino, estos seguidores son solo el 5% de los de la fundación.

De aquí se puede concluir que, con un porcentaje tan pequeño, la Fundación tiene otros programas educativos con mucha más relevancia para su comunidad de seguidores que Celera.
Otro dato que se puede concluir es que el 71% de la comunidad de Twitter de Celera no sigue a la fundación que lo hace posible. Por ello, se podría concluir que no les ha despertado el interés, y que la Fundación Rafael del Pino puede aprovechar este programa para ganar más seguidores.

- Cuentas de asociaciones

Hay otro grupo de relaciones con un alto porcentaje, entre el 23 y el 12%, que son las cuentas de organizaciones de estudiantes que tienen comités en diferentes ciudades.

Antecedente	Consecuente	Confianza	Relaciones
AEGEE_Zaragoza	AEGEE_Malaga	0.7%	7
AEGEE_Malaga	AEGEE_Zaragoza	16%	7
aegeebarcelona	AEGEE_Zaragoza	53%	163
AEGEE_Zaragoza	aegeebarcelona	16%	163
AEGEE_Malaga	aegeebarcelona	15,91%	7
aegeebarcelona	AEGEE_Malaga	2,3%	7

La antenna de AEGEE de Málaga tiene solo 44 seguidores en Twitter. Eso produce una sobrerrepresentación de las otras cuentas de Barcelona y Zaragoza, con unos porcentajes de confianza muy altos o muy bajos.

Respecto a las interacciones entre Barcelona y Zaragoza, con comunidades de Twitter más grandes (306 y 995 seguidores respectivamente), se puede observar como comparten un 53% y un 16% respectivamente.

En este caso, puede plantearse la hipótesis de que sucede porque Zaragoza tiene una comunidad más amplia en Twitter, dirigiéndose a otros grupos como universitarios que no están en la asociación y socios de AEGEE de otras ciudades.

Antecedente	Consecuente	Confianza	Relaciones
BESTMadridUPM	BESTUPC	13%	200
BESTUPC	BESTMadridUPM	23%	200
BESTUPC	BESTValenciaUPV	19%	165
BESTValenciaUPV	BESTUPC	25%	165
BESTValenciaUPV	BESTMadridUPM	26%	175
BESTMadridUPM	BESTValenciaUPV	11%	175

En estos casos, las comunidades en Twitter son más grandes, que puede deberse a que los estudiantes tecnológicos utilizan más Twitter que los estudiantes en general.

También puede observarse como las relaciones son también similares, con una diferencia de solo 35 seguidores en común.

La cuenta de Best Madrid UPM con Best UPC, tiene 13% de seguidores en común, mientras que al revés es de un 23%. Esto sucede porque la cuenta de Best UPC tiene muy pocos seguidores comparada con la cuenta de Best Madrid, que es más activa. Entonces, de la comunidad de Best Madrid, con que solo 20 sigan a Best Barcelona, ya tiene un porcentaje relativo muy elevado.

- Relación entre asociaciones

Las asociaciones que vamos a analizar son las siguientes:

- AEGEE y sus 3 cuentas
- BEST y sus 3 cuentas
- AIESEC
- ESN
- ELSA

\ antecedentes consecuentes	_celera	E_Zaragoza	aegeebarcelona	AEGEEMalaga	aiesecspain	MadridUPM	BESTUPC	ValenciaUPV	EFEspana	ELSA_Spain_	ESNSpain
-> Nº seguidores	2138	995	306	44	4166	1489	857	658	6303	391	3435
_celera	0	0.1	0.65	2.27	0.86	0.94	0.7	0.76	0.14	0.26	0.35
AEGEE_Zaragoza	0.05	0	53.3	15.9	0.5	0.4	0.58	0.76	0.16	1.28	0.87
aegeebarcelona	0.09	16.4	0	15.9	0.14	0.13	0.23	0.15	0.05	0.77	0.61
AEGEEMalaga	0.05	0.7	2.29	0	0	0.07	0.12	0	0.02	0.26	0
aiesecspain	1.68	2.11	1.96	0	0	1.14	0.82	0.76	0.83	2.56	2.13
BESTMadridUPM	0.65	0.6	0.65	2.27	0.41	0	23.3	26.6	0.24	0.26	0.61
BESTUPC	0.28	0.5	0.65	2.27	0.17	13.4	0	25.1	0.1	0.77	0.29
BESTValenciaUPV	0.23	0.5	0.33	0	0.12	11.8	19.3	0	0.11	0	0.32
EFEspana	0.42	1.01	0.98	2.27	1.25	1.01	0.7	1.06	0	0.51	0.7
ELSA_Spain_	0.05	0.5	0.98	2.27	0.24	0.07	0.35	0	0.03	0	0.41

Figura 5.1: Heatmap seleccionado de confianza en porcentaje (%)

Eje X: Cuenta antecedente. Eje Y: Cuenta consecuente

A pesar de que haya amistades entre personas de distintas asociaciones, la relación entre seguidores de cuentas de asociaciones es muy baja, de menos del 1,2% excepto AIESEC, que tiene del 2,1% para abajo. (Nota: AEGEE Málaga, al ser un número tan bajo, está sobrerrepresentado. El 2,27% es solo un seguidor en común).

Universia

La cuenta internacional de Universia tiene aproximadamente un tercio de seguidores que la cuenta en España.

Al tener una cantidad tan grande de seguidores, su porcentaje con el resto de cuentas es muy pequeño. Sin embargo, si miramos a UniversiaES como consecuente, puede observarse un porcentaje relativamente alto (4-16%) en las asociaciones de estudiantes, en fundaciones (4-16%), en programas educativos y en comunidades.

Con esto puede concluirse que Universia se ha dado a conocer de forma efectiva a sectores muy amplios de jóvenes talentos. De esta forma se concluye que Universia tiene una reputación y difusión real para cualquier iniciativa de desarrollo de talento juvenil.

- Trivu

Los seguidores de Trivu tienen una alta tasa de compartición de seguidores con las tres fundaciones, con entre el 6, 8 y el 13%, que implican 250, 335 y 536 seguidores en común.

La comunidad de Trivu no solo se compone de jóvenes, sino también empresas y profesionales de Recursos Humanos y talento.

Sin embargo, con estos datos se confirma que esta comunidad es real, y que una colaboración con esta empresa para aparecer en sus redes sociales implicará llegar a un público interesado en fundaciones. Los proyectos como la conferencia Unleash y sus eventos dejan claro que han estrechado relaciones y mostrando afinidad real entre comunidades.

También se comprueba que el público de Teamlabs, más emprendedor e innovador, muestra también afinidad con Trivu, y su interconexión.

De Trivu también sorprende que la única organización de estudiantes con una cantidad considerable de seguidores en comú es AIESEC, con un 2,5% y 100 seguidores en común, que sigue siendo un número reducido.

- Fundaciones

\ antecedentes consecuentes	FPdGi	frdelpino	FundacionBKT
-> Nº seguidores	13849	11970	13627
AEGEE_Zaragoza	0.007	0.058	0.022
aegeebarcelona	0.036	0.017	0.044
AEGEEMalaga	0	0.008	0.007
aiesecspain	0.635	1.003	0.69
BESTMadridUPM	0.195	0.434	0.264
BESTUPC	0.217	0.075	0.11
BESTValenciaUPV	0.101	0.05	0.154
ELSA_Spain_	0.029	0.159	0.095
ESNSpain	0.477	0.468	0.389
IAESTE_Spain	0.036	0.05	0.015
EFEspana	0.477	0.493	0.338
factoriatalent0	0.477	0.819	0.66
FPdGi	0	5.104	5.078
frdelpino	4.412	0	7.757
FundacionBKT	4.997	8.83	0
_celera	1.487	5.196	2.018
jovejuristacat	0.029	0.142	0.073
laakademia_org	1.791	1.019	1.321
MondragonTA	2.376	1.93	2.759
ShapersMadrid	0.578	1.236	0.609
teamlabs	4.917	4.954	5.717
ThePowerMBA	0.448	0.911	0.976
Universia	0.852	1.086	2.282
UniversiaES	5.625	7.243	6.443
wearetrivu	1.805	4.478	2.458

Figura 5.2: Heatmap seleccionado de confianza en porcentaje (%)

Eje X: Cuenta antecedente. Eje Y: Cuenta consecuente

Las fundaciones entre ellas tienen un número de seguidores y afinidad en común muy elevadas, entre el 4 y el 7%.

Respecto a su relación con asociaciones y organizaciones de estudiantes, su interrelación es muy baja, del 0 al 0,4% a excepción de AIESEC, que llega al 1% con la Fundación Rafael del Pino. En el caso de AIESEC, el número absoluto va del 88 a 120, y el caso de BEST de 6 a 52.

AEGEE es la asociación que, a nivel de Twitter, no tienen prácticamente afinidad, de 7 a 0. De aquí puede deducirse que, al haber una afinidad de Twitter prácticamente nula, en caso de crear colaboraciones con asociaciones, esta estaría en los últimos lugares.

También podría interpretarse como un potencial de crecimiento, al ser las fundaciones desconocidas para los seguidores de AEGEE, se podría probar a hacer colaboraciones y probarlo.

Entonces, para mayor efectividad, se pueden recomendar campañas de marketing pagadas para las cuentas que sigan a AIESEC y a BEST.

Respecto a los programas formativos para jóvenes, en el caso de Jóvenes Juristas puede observarse una afinidad muy prácticamente nula, mientras que con La Akademia y Celera tienen sinergia relativamente elevada, con entre 122 y 280 cuentas en común, etnre el 1 y el 1,8% de las fundaciones.

También se observa una alta afinidad con Teamlabs, MondragonTA, Trivu y Universia, por lo que hay potencial de colaboración.

Sin embargo, se observa poco interés mutuo en ThePowerMBA, Shapers e IAESTE.

- Programas formativos

Los programas formativos tienen una altísima afinidad con las fundaciones. Factoría de Talento, entre el 5 y el 7,6%, y Celera entre el 9 y el 29%. El caso de Jóvenes Juristas, al ser una cuenta pequeña es prácticamente nulo, y en La Akademia es de entre el 1,5 y el 3,2%.

En el caso de Jóvenes Juristas, el interés mutuo más considerable es con ELSA, asociación de estudiantes de derecho, con un 2,6%, con Fundación Rafael del Pino, con un 3,7% y Fundación Bankinter, con un 2,2%.

- ThePowerMBA

ThePowerMBA es una academia online de cursos sobre negocios, cuya base es un MBA online. Se dirige sobretodo a personas emprendedoras, y eso se nota en su alta afinidad con Teamlabs, un 3,5%. También se muestra un interés en la educación alternativa que producen las fundaciones, con un 4%, un 3,3% y un 1,9%. También un número considerable sigue a Universia, un 3,2%, como todos los otros conjuntos de seguidores.

- Comunidades

La comunidad de Global Shapers Madrid tiene un afinidad altísima con las fundaciones, de entre un 6,6% y un 12,3%, con el programa Celera con un 6,7%, con Mondragon TA y Teamlabs con un 7,4% y un 4%.

La comunidad Trivu, que forma parte de una consultora, tiene un comportamiento muy similar al de Global Shapers. Por tanto, podría deducirse que son comunidades con inquietudes similares.

También sorprende que de los seguidores en común de Trivu con el de las asociaciones sea tan reducido, a excepción de AIESEC, que comparten un 2,5% que son 100 personas.

- Education First

Siendo Education First una referencia internacional de cursos y programas en el extranjero, sorprende que del resto de cuentas analizadas (Excepto AEGEE Málaga, cuyo 2,27% es una sola cuenta), esté siempre por debajo del 1,1% mirando EF como consecuente, y Universia).

Eso puede deberse a que su marketing y público no es tan similar como el de las organizaciones de estudiantes, fundaciones y comunidades.

5.2 Conclusiones y propuestas concretas

1. Los programas formativos, sean o no de una fundación, tienen una alta afinidad con otras similares estas. Entonces, a la hora de hacer publicidad pagada, se puede recomendar hacer un esfuerzo especial en seguidores de cuentas de fundaciones.
2. Universia es un medio con afinidad considerable prácticamente de todas las cuentas analizadas. Entonces, colaborar en este medio con iniciativas educativas, publicidad pagada y redacción de artículos es una buena estrategia.
3. Las comunidades de talento en una cantidad importante conocen a las fundaciones con proyectos educativos para jóvenes. Entonces, colaborar con estas siempre tendrá cierta garantía de éxito.
4. Los seguidores de las asociaciones y organizaciones de estudiantes tienen poco conocimiento de las fundaciones.
5. La comunidad de Teamlabs tiene una gran afinidad con las fundaciones. Entonces, las colaboraciones con esta institución o publicidad pagada a los seguidores de esta, serán especialmente efectivas.

Sumario del trabajo

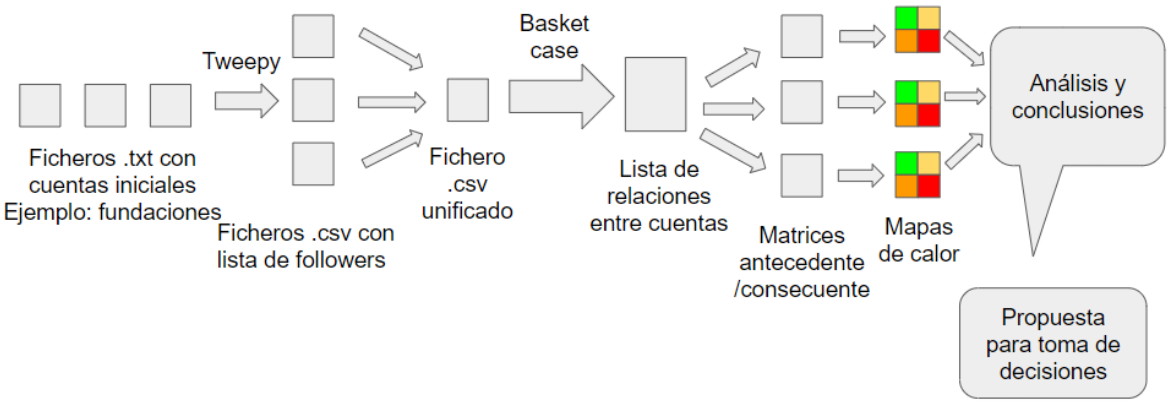

Figura 6.1 Diagrama de procesos de este trabajo

1. Para este trabajo, en primer lugar, se han mapeado una serie de organizaciones cuyas cuentas se van a analizar.
2. Después, tras programar en Python un software se ha extraído una serie de datos de cada seguidor de cada cuenta de la lista.
3. Tras ello, se han filtrado las columnas de Follower_1_Previo y los nombres para, a través de un Market Basket Case, obtener una lista de relaciones y otros datos como la confianza entre seguidores y el soporte.
4. Se crea una matriz donde aparecen las confianzas entre cuentas y se guarda en formato .csv
5. Se importan las matrices en Excel, se ordenan las columnas y filas por orden alfabético y se le añade una fila con el número de seguidores de cada cuenta para interpretarlos con más precisión. Las matrices son de confianza y la de soporte del precedente, que se transforma en número de seguidores en común para ser más facilmente interpretable.
6. Se crean los mapas de calor en base a los datos de las matrices
7. Se analizan los casos y se llegan a conclusiones.

Conclusiones de uso práctico del estudio

1. Los programas formativos, sean o no de una fundación, tienen una alta afinidad con otras similares estas. Entonces, a la hora de hacer publicidad pagada, se puede recomendar hacer un esfuerzo especial en seguidores de cuentas de fundaciones.
2. Universia es un medio con afinidad considerable prácticamente de todas las cuentas analizadas. Entonces, colaborar en este medio con iniciativas educativas, publicidad pagada y redacción de artículos es una buena estrategia.
3. Las comunidades de talento en una cantidad importante conocen a las fundaciones con proyectos educativos para jóvenes. Entonces, colaborar con estas siempre tendrá cierta garantía de éxito.
4. Los seguidores de las asociaciones y organizaciones de estudiantes tienen poco conocimiento de las fundaciones.
5. La comunidad de Teamlabs tiene una gran afinidad con las fundaciones. Entonces, las colaboraciones con esta institución o publicidad pagada a los seguidores de esta, serán especialmente efectivas.

Posibles evoluciones de este proyecto:

- Aplicar el Machine Learning al aprendizaje. Programar un software que vaya almacenando las cuentas aprendidas, y descartando aquellas cuyo máxima afinidad con el resto sea menor del 1%, y por tanto, no pueda extraerse información útil más que no existe esa afinidad.
- Extraer los datos de Instagram. Al ser una red social más actual y donde hay un público más joven, podrían incluso haber más cuentas, y extraerse mayor información. Pueden utilizarse técnicas de Web Scraping, o una API no oficial.
- Analizar la lista de seguidores de una cuenta, en base a las cuentas que siguen estos seguidores. Esta era la idea inicial del proyecto, pero hubo complicaciones técnicas de la API de Twitter, que impidió una extracción tan masiva de datos.

Ampliaciones de este tipo a otras situaciones o problemáticas

Este tipo de análisis además, por ejemplo, puede ser muy útil para:

- Analizar compras, y relaciones entre artículos adquiridos para hacer sugerencias.

- Analizar influencers y relacionarlos con cuentas de empresa. De esta manera, una empresa u organización que invierta miles de euros en promoción con influencers será más eficaz cuando lo haga con los que tienen mayor afinidad. Por ejemplo: marcas de ropa, cursos, eventos, ordenadores, etc.
- Relacionar compras para colocar estratégicamente los productos. Por ejemplo, si en un supermercado el 50% de los que compran huevos también compran patatas, puede considerarse poner los huevos y patatas cerca para facilitarle la compra al cliente y que sea más probable que compre ambas cosas.
- Optimización de rutas de tren, autobús y otros transportes. Teniendo un registro de entrada y de salida del billete de cada pasajero, se pueden optimizar las frecuencias y las rutas según los pasajeros que caben en el autobús o tren para que la mayor cantidad posible tarde menos.

Valoración del aprendizaje y competencias adquiridas

El hecho de haber aprendido sobre técnicas de Business Intelligence, sumado al conocimiento adquirido en la Ingeniería Eléctrica, aumenta el valor que se le puede ofrecer a las empresas.

Para mí ha sido una sorpresa ver que las relaciones entre cuentas similares fuera tan pequeña. Antes de desarrollar este trabajo me esperaba una afinidad, por ejemplo, del 20%, pero a la práctica estas cantidades han sido puntuales.

También me ha sorprendido ver que los seguidores de algunas asociaciones de estudiantes tienen poca cantidad relativa de seguidores en común con fundaciones, cuando en un principio ofrecen desarrollo profesional al mismo público. Esto me ha reforzado la idea de que, por muy obvia que parezca una hipótesis, respaldarla con datos puede evitar el derroche de miles de euros mal invertidos.

Presupuesto

El presupuesto de este trabajo es el siguiente:

	Horas trabajadas	Precio por hora	Precio
Planteamiento inicial del proyecto	75	25€	1.250€
Mapeo de organizaciones	50	25€	2.500€
Programación en Python de la extracción de datos, con arquitectura de datos	125	25€	2.500€
Programación en Python del Basket Case y transformación en mapa de calor	100	25€	2.500€
Análisis de datos y propuesta de toma de decisiones	100	25€	2.500€
Conclusiones y bibliografía	50	25€	1.250€
	TOTAL HORAS		TOTAL BRUTO
	500 horas		12.500€
		Total IVA 21%	TOTAL NETO
		2.625€	15.125€

Bibliografía

[1] Academia de consultores, Consulta: 07/2019, https://academiadeconsultores.com/consultoria-estrategica/

[2] Oracle, Consulta: 07/2019, https://www.oracle.com/es/big-data/guide/what-is-big-data.html

[3] Universia, Consulta: 08/2019, https://noticias.universia.edu.uy/cultura/noticia/2018/07/05/1160593/casos-exito-como-utiliza-amazon-big-data.html

[4] Bernard Marr, Forbes, Consulta: 09/2019, https://www.forbes.com/sites/bernardmarr/2018/04/30/27-incredible-examples-of-ai-and-machine-learning-in-practice/#502f8d2f7502

[5] Unesco, Consulta: 11/2019, http://www.unesco.org/new/es/social-and-human-sciences/themes/youth/about-youth/

[6] P. Sempere, El País, Consulta: 10/2019, https://cincodias-elpais-com.cdn.ampproject.org/c/s/cincodias.elpais.com/cincodias/2019/08/20/fortunas/1566324163_444952.amp.html

[7] Fundació Princesa de Girona, Consulta: 2/2020, https://es.fpdgi.org/quienes-somos/presentacion/

[8] Esther Hochsztain y Andrómaca Tasistro, Aplicación de técnicas de Business Intelligence para el almacenamiento y análisis de datos en la dirección de proyectos informáticos, Facultad de Ciencias de Económicas y de administración (Universidad de la República Uruguay), Consulta: 2/2020, http://fcea.edu.uy/Jornadas_Academicas/2013/file/ADMINISTRACION/Aplicacion%20de%20tecnicas%20de%20Business%20Intelligence.pdf

[9] Project Management Institute España, Consulta: 1/2020, https://pmi-mad.org/socios/articulos-direccion-proyectos/317-resolucion-de-la-paradoja-de-cobb

[10] Marketing4ecommerce, Consulta: 1/2020, https://marketing4ecommerce.net/cuales-redes-sociales-con-mas-usuarios-mundo-2019-top/

[11] The Social Media Family, Consulta: 1/2020,
https://thesocialmediafamily.com/informe-redes-sociales/

[12] NewGenApps, Consulta: 3/2020,
https://www.newgenapps.com/blog/what-is-market-basket-analysis-predicting-customer-purchases-big-data