

Planteamiento

Práctica 1. Web Scraping

Introducción

Internet se ha convertido en una fuente inagotable de información. Aunque en algunas ocasiones es posible recuperar información de forma estructurada, la mayor parte del conocimiento en internet se encuentra integrado en la estructura y estilo de las diferentes páginas web. Es en estos casos donde la extracción de información puede convertirse en una tarea compleja.

El presente documento expone las tripas del proyecto analítico que, para su desarrollo, se ha hecho uso de las herramientas de software actualmente disponibles que permiten simplificar y **automatizar el proceso de extracción de los datos** relevantes para el mismo.

El documento incluye el propósito del proyecto, el valor que aporta, una descripción breve del conjunto de datos, información básica acerca de la fuente de los datos, principios éticos y legales seguidos, y el repositorio Git donde se encuentra almacenado el código junto al *dataset*.

1. Contexto

Los **fondos de capital riesgo** hacen un gran esfuerzo de investigación. Los datos que genera el mercado están dispersos y cada vez se vuelve más difícil capturar y procesar toda la información.

Nosotros creemos que los fondos de capital riesgo **deberían centrarse en buscar el próximo unicornio**. Nuestra misión es ayudar a estas organizaciones a recaudar fondos e invertir en *startups* de forma eficaz **proporcionándoles información valiosa en tiempo real sobre los movimientos del mercado**.

Nuestro MVP consiste en aplicar la técnica de *web scraping* para recopilar de forma automática datos sobre organizaciones listadas en Crunchbase.

Crunchbase es una plataforma que agrupa información sobre empresas: inversiones que han realizado y recibido, el listado de fundadores o individuos en posiciones relevantes, adquisiciones, noticias y tendencias de la industria. En definitiva, Crunchbase contiene información de empresas públicas y privadas a escala mundial.

2. Título

El conjunto de datos extraído contiene información sobre todo tipo de organizaciones, desde pymes hasta escuelas o fondos de capital riesgo. Por lo que, el título más descriptivo, de acuerdo al contenido del *dataset*, es:

Organizations—Crunchbase

3. Descripción del dataset

La información recogida de Crunchbase es información comercial, donde es posible consultar, entre otros campos,

- las industrias en las que una organización opera «Industries»,
- dónde se encuentra su sede «Headquarters Region»,
- año de fundación «Founded Date»,
- estado actual «Operating Status»,
- *email* de contacto «Contact Email»,
- teléfono de contacto «Phone Number»,
- sus fundadores «Founders»,
- si es un fondo en qué compañías y etapas invierte «Investment Stage»,
- etc.

Los tipos de organizaciones que se pueden observar en los datos son:

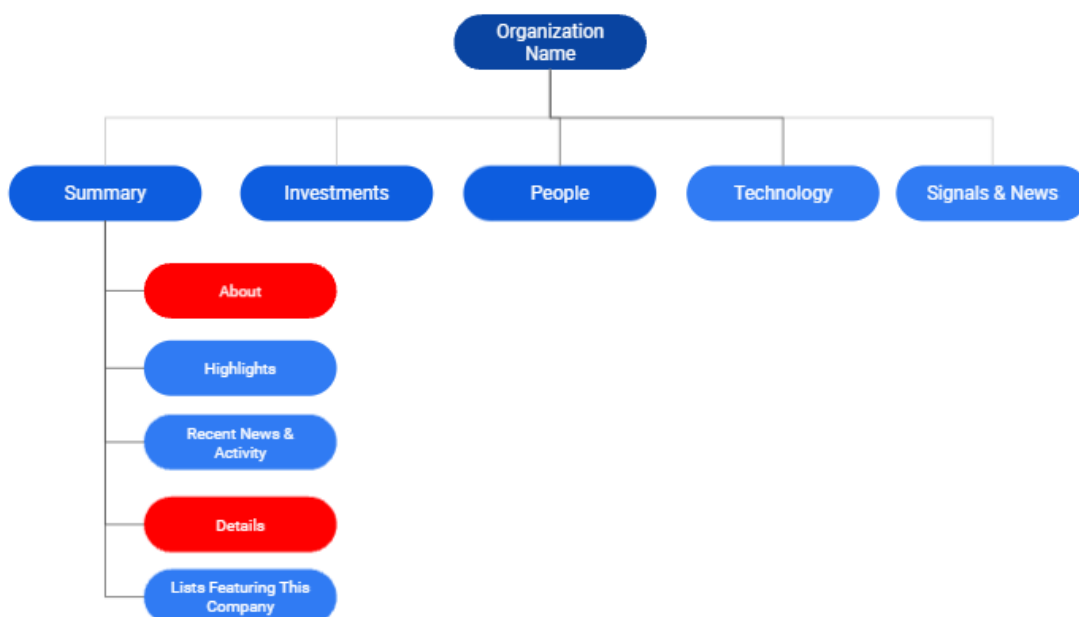
- Empresas
- Inversores
- Escuelas

El *dataset* se ha elaborado con el objetivo de ayudar a los fondos de capital riesgo a dedicar menos tiempo buscando información y más tiempo cerrando acuerdos.

4. Representación gráfica

La base de datos de Crunchbase alberga, además de perfiles organizacionales, perfiles de personas físicas, eventos, rondas de financiación y *mergers & acquisitions* (M&As). Pero, con el propósito de sacar un producto mínimo viable lo antes posible al mercado, se ha decidido circunscribirse a los datos organizacionales.

A alto nivel, la página de perfil de una organización tiene la siguiente estructura:



En “Summary” se encuentra la información que interesa recoger para el presente proyecto analítico, específicamente en la subsección “Details”.

La automatización de la extracción de la información, como comentaremos en el siguiente apartado, no ha sido sencilla. Se han identificado varias inconsistencias, siendo la más notable los campos visibles por organización.

El perfil de una entidad tiene una serie de campos informados, como se puede observar en la siguiente imagen,

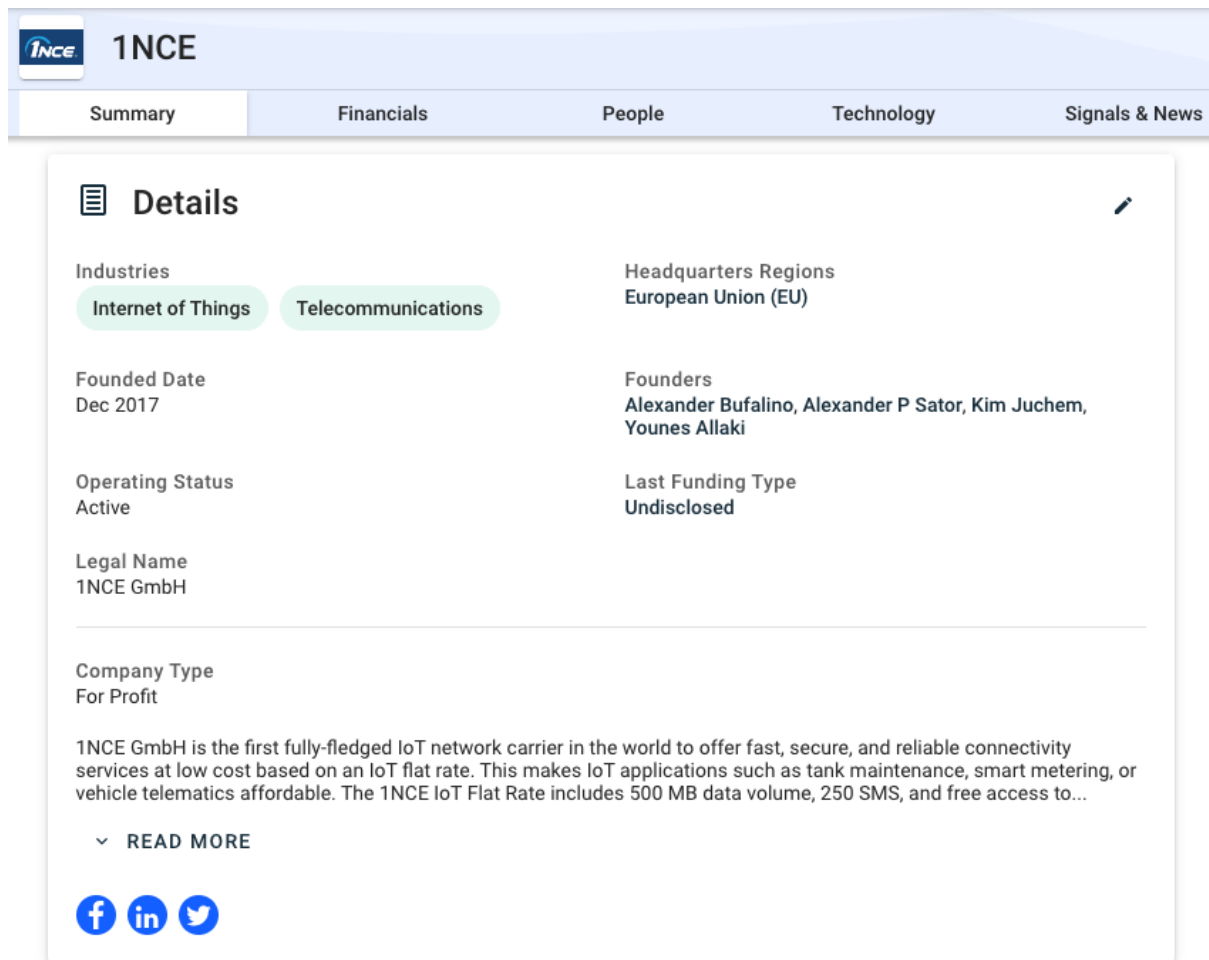
The image shows the Kahoot! company profile page on Crunchbase. The page has a navigation bar with tabs: Summary, Financials, People, Technology, and Signals & News. The 'Summary' tab is selected. Below the navigation bar, the 'Details' section is visible, containing several fields highlighted with red boxes:

- Industries:** EdTech, Education, Gaming, Mobile, Software
- Headquarters Regions:** Nordic Countries, Scandinavia
- Founded Date:** Dec 14, 2012
- Founders:** Alf Inge Wang, Asmund Furuseth, Jamie Brooker, Johan Brand, Morten Versvik
- Operating Status:** Active
- Last Funding Type:** Post-IPO Secondary
- Legal Name:** Kahoot! AS
- Stock Symbol:** OSE:KAHOOT-ME
- Company Type:** For Profit
- Contact Email:** hello@kahoot.com

Below the fields, there is a description of Kahoot! and a 'READ MORE' link. At the bottom, there are social media icons for Facebook, LinkedIn, and Twitter.

donde cada campo viene identificado por una etiqueta e informado con su correspondiente valor.

La complejidad de recoger datos de Crunchbase reside en que cada página tiene una serie de campos que no son necesariamente los mismos. Pueden coincidir, pero en la mayoría de las ocasiones difieren:



1NCE

Summary Financials People Technology Signals & News

Details

Industries
Internet of Things Telecommunications

Headquarters Regions
European Union (EU)

Founded Date
Dec 2017

Founders
Alexander Bufalino, Alexander P Sator, Kim Juchem, Younes Allaki

Operating Status
Active

Last Funding Type
Undisclosed

Legal Name
1NCE GmbH

Company Type
For Profit

1NCE GmbH is the first fully-fledged IoT network carrier in the world to offer fast, secure, and reliable connectivity services at low cost based on an IoT flat rate. This makes IoT applications such as tank maintenance, smart metering, or vehicle telematics affordable. The 1NCE IoT Flat Rate includes 500 MB data volume, 250 SMS, and free access to...

▼ READ MORE

f in t

Por ejemplo, el perfil de **1NCE** no dispone de un campo que indique el correo electrónico de contacto.

Dada esta inconsistencia, se ha considerado que en este caso **el modelo de datos debe ser dinámico**, es decir, que se genere a medida que se analizan perfiles, puesto que no se conoce de antemano qué variables se van a almacenar.

5. Contenido

El conjunto de datos extraído de Crunchbase contiene 1.000 observaciones, una muestra muy pequeña del conjunto de datos potencial.

El *script* desarrollado es capaz de replicar la base de datos de Crunchbase al completo, recogiendo información de las 1,6 M de organizaciones registradas. No obstante, se ha establecido un límite de registros por razones de practicabilidad.

En un primer momento, el desarrollo del proyecto se vio afectado por los diversos métodos diseñados para la prevención del *web scraping*. No obstante, se han tomado las medidas necesarias para resolver los obstáculos y extraer de forma exitosa los datos identificados como relevantes para el proyecto analítico.

El primer obstáculo apareció al inicio del proceso, con el lanzamiento de la primera sentencia del *script* que pretendía recoger datos del mapa del sitio web con el objetivo de navegarlo.

De forma predeterminada, las bibliotecas utilizadas para realizar peticiones HTTP de forma automática establecen su propio *user agent* basándose en el nombre de la librería, lo que facilitó la tarea a Crunchbase que, por defecto, bloqueó el acceso del robot a su página web.

```
<p>
Access to this page has been denied because we believe you are using automation tools to browse the
website.
</p>
<p>
This may happen as a result of the following:
</p>
<ul>
<li>
Javascript is disabled or blocked by an extension (ad blockers for example)
</li>
<li>
Your browser does not support cookies
</li>
</ul>
<p>
Please make sure that Javascript and cookies are enabled on your browser and that you are not blocking
them from loading.
</p>
<p>
Reference ID: #8d4be225-29f0-11ec-87ba-724e71484749
</p>
</div>
</div>
<div class="page-footer-wrapper">
<div class="page-footer">
```

La respuesta a este primer obstáculo fue sencilla. Inmediatamente se modificó el *user agent* y otras cabeceras HTTP para ocultar el hecho de que las peticiones realizadas provienen de un *script*.

```
In [4]: headers = {
    "Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,\n
    /*;q=0.8",
    "Accept-Encoding": "gzip, deflate, sdch, br",
    "Accept-Language": "en-US,en;q=0.8",
    "Cache-Control": "no-cache",
    "dnt": "1",
    "Pragma": "no-cache",
    "Upgrade-Insecure-Requests": "1",
    "User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_3) AppleWebKit/5\
    37.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/537.36"
}
```

Este cambio fue suficiente para esquivar el bloqueo de peticiones (por un tiempo). Una vez finalizada la fase de análisis, donde se identificó la estructura de la web, y se diseñó el proceso que automatizaría la extracción de la información, se lanzó el *script*.

Crunchbase apenas tardó 1 segundo en bloquear de nuevo el acceso, esta vez poniendo la IP en la lista negra.

Entonces, se procedió a implementar la segunda medida de prevención: **simular el comportamiento humano estableciendo un tiempo de espera** fijo. Y funcionó. Pero, ya habiendo sido bloqueados dos veces, se fue un paso más allá, y se estableció un tiempo de espera aleatorio de entre 5 y 20 segundos para evitar que el algoritmo de Crunchbase observara un patrón con facilidad.

```

In [21]: response = requests.get(url, stream=True, headers=headers)
root = etree.fromstring(response.content)
#rows = []

for sitemap in root:
    children = sitemap.getchildren()
    # filtrar tipo de perfil
    #organization = re.match(r".+organizations-[0-9]+", children[0].text)
    organization = re.match(r".+organizations-9", children[0].text)
    if organization:
        print(children[0].text)
        r = requests.get(children[0].text, stream=True, headers=headers)
        g=gzip.GzipFile(fileobj=BytesIO(r.content))
        content=g.read()
        rootorg = etree.fromstring(content)
        count = 0
        for company in rootorg:
            if count > 500:
                childrenorg = company.getchildren()
                time.sleep(random.randint(5,20))
                rlink = requests.get(childrenorg[0].text, headers=headers)
                soup = BS(rlink.content, 'html.parser')
                # indicar por pantalla si el acceso ha sido denegado, y terminar el proceso
                if soup.head.title.text == "Access to this page has been denied.":
                    print("Access denied")
                    break
                elif len(rows) >= 1000:
                    print("Reached limit")
                    break
                # obtener la información de la empresa y añadirla a lista de registros
                row = parse2row(soup, log=True)
                rows.append(row)
            else:
                count = count + 1
        else:
            continue
        break

```

Anécdota: La extracción del conjunto de datos se ha realizado con los datos móviles de uno de los integrantes del grupo, dado que su red privada estaba vetada de realizar cualquier tipo de petición al sitio.

En lo que a campos se refiere, es importante indicar que estos han sido identificados a posteriori: «Name, Profile, Industries, Headquarters Region, Founded Date, Operating Status, Company Type, Also Known As, Contact Email, Phone Number, Founders, Legal Name...»

	A	B	C	D	E	F	G	H	I	J	K	L
1	Name	Profile	Industries	Headquarter	Founded Dat	Operating St	Company Ty	Also Known	Contact Ema	Phone Numt	Founders	Legal N
2	E-Law Soluti	Organization	Information	Asia-Pacific (2013	Active	For Profit					
3	E Lawsuit Lo	Organization	Finance, Fin	Greater New Jul	13, 2000	Active	For Profit	150 Essex St	sales@elaws	(800)972-5560		
4	eLawTalk.cor	Organization	Consulting, L	Asia-Pacific (Jan 15, 2013	Active	For Profit	eLawTalk	beausensei@	808-321-1594		
5	Elaxer	Organization	Mobile Apps	Asia-Pacific (APAC)		Active	For Profit		hello@elaxe	91-98913-55	Varun Razora	
6	Elaxy	Organization	Software	European Union (EU)		Active	For Profit					Elaxy gr
7	eLayaway	Organization	E-Commerce	East Coast, S	Oct 1, 2005	Active	For Profit		info@elayaw	850-583-5019		
8	Elaydin Tech	Organization	iOS, Softwar	Greater New York Area, E		Active	For Profit		elaydin.tech	+1 212-982-4576		
9	ELayer	Organization	Web Hosting	Great Lakes,	1999	Active	For Profit		sales@elaye	1-877-ELAYER.COM		
10	ELayers Inter	Organization	Mobile Apps	Asia-Pacific (2006	Active	For Profit		contact@ela	+91 814-000-9888		eLayers
11	Elavøs Cosm	Organization	Food and Be	European Union (EU)		Active	For Profit		contact@elays	-cosmetique.com		
12	Elazig Cimen	Organization	Building Material, Food ar		1950	Active	For Profit					
13	El Azteca Tac	Organization	Restaurants	East Coast, Southern US		Active	For Profit			239-574-0056		
14	Elba	Organization	Consumer Goods, Manufacturing, Reta			Active	For Profit		arktika-mw@	(044) 425 91 88		
15	Elba Assicura	Organization	Insurance, Pi	European Ur	jun-08	Active	For Profit		info.elba@el	02 92885700		Elba As
16	elbaC Cable	Organization	Electronics, f	European Ur	Jan 1, 2007	Active	For Profit		info@elbac.f	33 232620092		
17	Elba Comput	Organization	Computer, Information Technology, So			Active	For Profit		info@elbasystem.it			
18	El Badr Plast	Organization	Industrial, Manufacturing		2003	Active	For Profit	El Badr Plast	info@elbadr	100-461-9968		
19	ELBA FLORES	Organization	Information	Latin Americ	sept-14	Active	For Profit	Uniaqua Tec	eflores@cru	5,26E+11		
20	Elbagate	Organization	Electronics, Lighting			Active	For Profit		sales@elbag	44-20 7254 9991		
21	Elbait	Organization	Blockchain, C	Asia-Pacific (Dec 1, 2017	Active	For Profit		marketing@elbait	com.au		Elbait
22	Elba Laborat	Organization	Manufacturi	Greater Detr	1982	Active	For Profit			248-288-6098		Elba Lal
23	Elba Liquefac	Organization	Energy, Foo	East Coast, S	2013	Active	For Profit					
24	El-Barbary In	Organization	Automotive, Industrial Au		2008	Active	For Profit	Barbary Inve	info@big.co	9,714E+10		

Dada la complejidad de la base de datos de Crunchbase, y las inconsistencias observadas en los distintos perfiles organizacionales, se ha desarrollado una función que, sin conocer de antemano el modelo de datos, fuera capaz de identificarlos y añadirlos a medida que se los encuentra, generando así un **modelo de datos dinámico**, que tiene como mínimo 31 variables, siendo «Name» y «Profile» los únicos campos fijos.

```
In [3]: def parse2row(soup, log=False):
        row = {}
        print("----") if log else None
        try:
            name = soup.select('h1[class="profile-name"]')[0].text.strip()
            row["Name"] = name
            print("Name: " + name) if log else None

            profile = soup.select('div[class="profile-type"] > span')[0].text
            row["Profile"] = profile
            print("Profile: " + profile) if log else None
        except IndexError:
            pass

        fields = soup.select('page-centered-layout .main-content profile-section .section-content fields-card > ul li')
        for li in fields:
            value = ''
            try:
                label = list(li.select('label-with-info'))[0].stripped_strings)[0]
                values = list(li.select('field-formatter'))[0].stripped_strings)

                try:
                    while True:
                        values.remove(',')
                except ValueError:
                    pass
                value = ', '.join(values)

                row[label] = value
                print(label + ': ' + value) if log else None
            except IndexError:
                pass

        return row
```

El modelo de datos deducido está limitado a las 1.000 observaciones extraídas. No obstante, el modelo final puede contener un mayor número de variables que por limitación de recursos no se ha podido obtener en este proyecto.

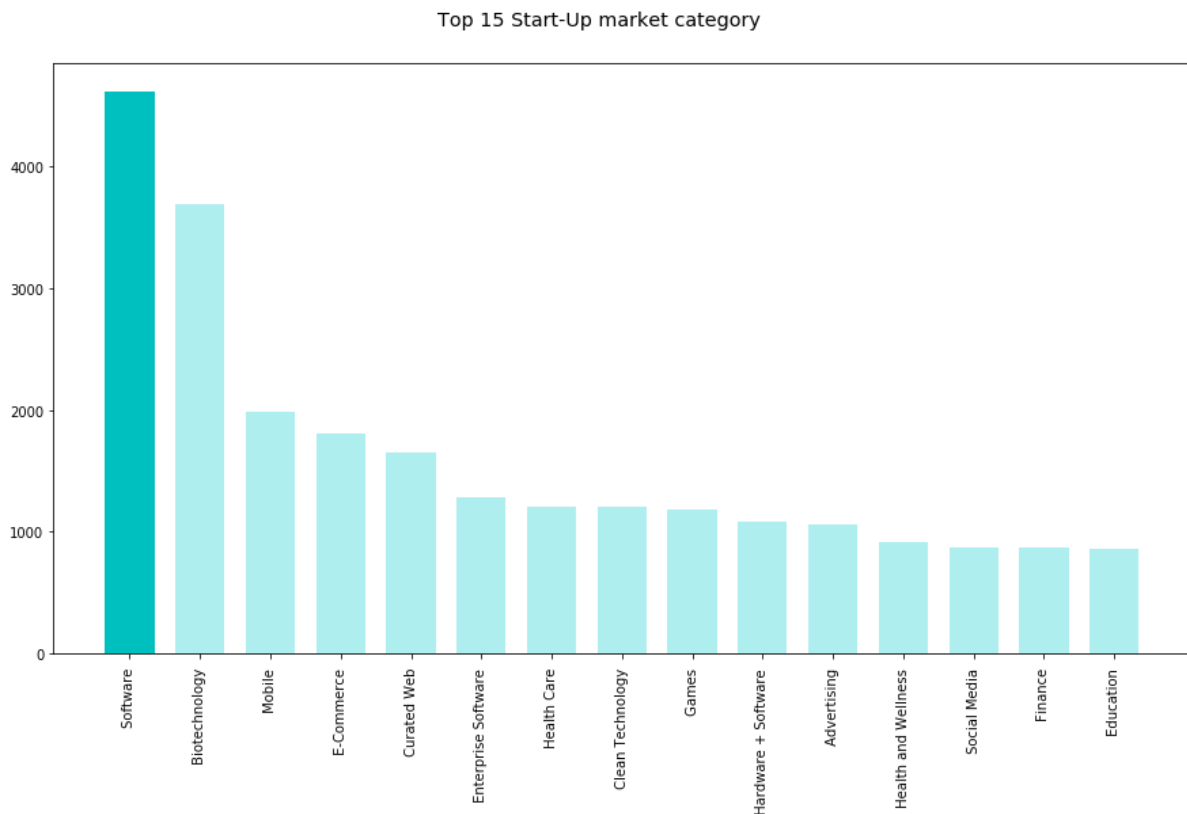
En cuanto al periodo de tiempo de los datos, sería ideal que el conjunto se actualizara en tiempo real. Al fin y al cabo, los fondos de capital riesgo necesitan la última información disponible para la toma de decisiones, siendo esta un factor determinante en el éxito de sus inversiones.

6. Agradecimientos

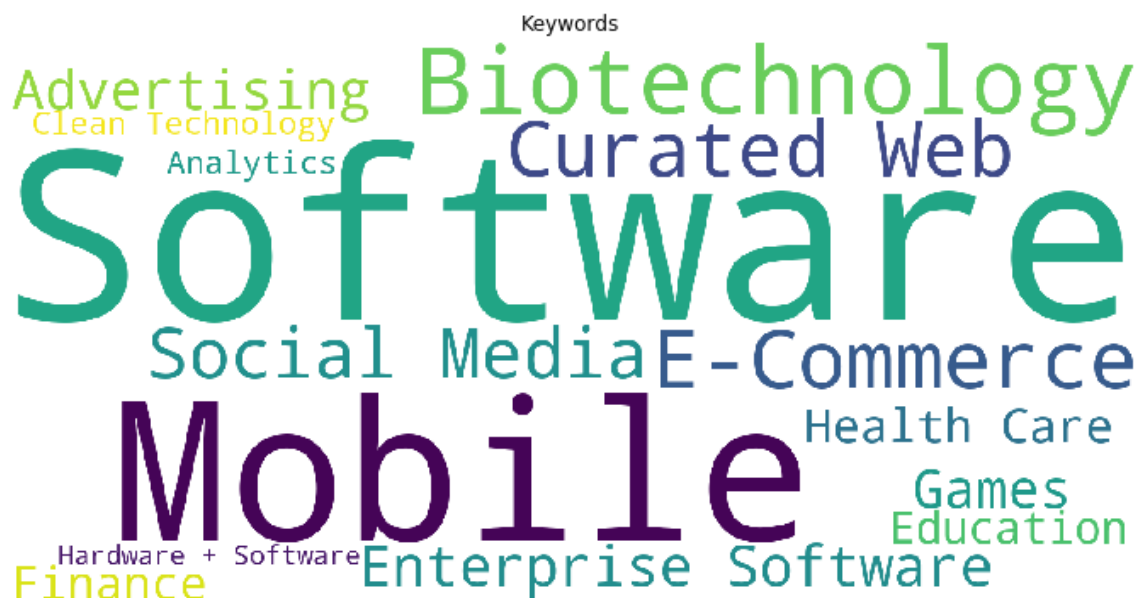
Este proyecto no presume de ser el primero que captura información de Crunchbase. Se han realizado proyectos similares anteriormente, entre los que cabe destacar “[Startup Investments EDA](#)”, publicado en Kaggle por Suraj Kumar.

Suraj, en su análisis, por un lado muestra la relación que tiene el lugar y el momento de fundación de las *startups* y el impacto que han tenido estos factores sobre el capital levantado. Y, por otro lado, hace un breve repaso de las tendencias del mercado, comparando las industrias entre sí y relacionándolo siempre de vuelta al capital levantado.

Otro análisis relacionado es el que realiza *Lastnight*, en su notebook “[Explanation of Startup Investment](#)”. En él, muestra las 15 industrias más activas en cuanto a *startups* fundadas,



y crea una nube de tags que recoge las palabras clave más frecuentadas en este ámbito,



entre otros.

Estos son ejemplos de proyectos similares al que se presenta en este documento.

Antes de realizar el ejercicio de automatización, se ha incorporado una fase previa al *web scraping* con el fin de evaluar los siguientes aspectos:

- 1) el archivo *robots.txt*,
- 2) el mapa del sitio web y
- 3) el propietario del mismo.

Archivo *robots.txt*

Este archivo contiene las restricciones a tener en cuenta cuando se pretende rastrear una página web, facilitando al desarrollador actuar de acuerdo a los principios éticos y legales.

El [archivo robots.txt de Crunchbase](#) define restricciones que excluye el acceso de todos los robots a los siguientes directorios: login, register, account, reset-password, subscriptions, contribute, add-new, edit, buy, account-setup, verify, admin, v4 y home.

Y, al mismo tiempo, otorga permiso de acceso completo a todos los robots al directorio situado en la siguiente ruta: /v4/md/applications/crunchbase, que hace de interfaz con su [API oficial](#). Sin embargo, y como se observó en el apartado anterior, no está permitida la aplicación del *web scraping*.

Mapa del sitio web

Examinar el mapa del sitio web (*sitemap*, del inglés) ayudará a localizar el contenido actualizado sin necesidad de rastrear cada una de las páginas que lo componen.

El [mapa del sitio web de Crunchbase](#) indica las categorías en las que está dividida su web:

- acquisitions
- events,
- funding_rounds,
- hubs,
- organizations and
- people.

Los perfiles alrededor de los que gira el proyecto son las *startups* y los fondos de capital riesgo. Esta información está contenida en la categoría *organizations*.

Propietario del mismo

El [propietario de la página web](#) se encuentra registrado en whois:

- Nombre: Jager McConnell
- Organización: Crunchbase Inc.
- Dirección: 410 Townsend St Suite 450
- Ciudad: San Francisco
- Estado: California
- Código Postal: 94107
- País: Estados Unidos
- Teléfono de contacto: (+1) 415-849-00-02
- Correo electrónico: accounts@crunchbase.com

Y recoge los casos de uso en su [página de términos y condiciones](#), donde indica que no está permitido aplicar *scraping* a su web:

You represent, warrant, and agree that you will not contribute or submit any User Submission (defined below) or other materials or otherwise use the Service or interact with the Service in a manner that:

[...]

(h) "Crawls," "scrapes," or "spiders" any page, data, or portion of or relating to the Service or Content (through use of manual or automated means);

(i) Copies or stores any significant portion of the Content;

[...]

A violation of any of the foregoing is grounds for account suspension or termination of your right to use or access the Service.

Los términos y condiciones son un conjunto de términos legales definidos por el propietario de una página web, que establecen los términos y condiciones que rigen las actividades de los visitantes de la página web y la relación entre los visitantes y el propietario del sitio web.

Por lo que, con el objetivo de evitar ser bloqueados, se ajustó la descarga utilizando tasas más conservadoras en la descarga de información.

7. Inspiración

Crunchbase es una de las principales bases de datos sobre startups e inversión del mundo. La mayor parte de las operaciones que tienen lugar cada día se registran en ella. Esto proporciona una visión privilegiada para entender lo que ocurre en el mundo de la tecnología y la inversión.

A partir de los datos extraídos de la plataforma, se pueden responder preguntas como:

- ¿Cuántos años han de madurar las empresas de media para ser vendibles? ¿Y para salir a bolsa?
- ¿A qué valoración se han vendido y cuánto han levantado de financiación hasta el momento?

Los resultados obtenidos de este análisis se podrían desglosar por industria, e incluso por regiones, de tal manera que los fondos de capital riesgo puedan estimar el capital a levantar de los fondos de fondos.

El conjunto de datos también permite analizar la competencia, fondos de capital riesgo.

Este análisis es parecido al que presentan los dos proyectos citados. Ambos estudian cómo ha evolucionado la financiación de las *startups* a lo largo de los años, qué industrias son las más atractivas de cara a emprender, cuántas *startups* siguen operando a día de hoy, y cuántas han cerrado, etc. Todas estas son preguntas similares a las que se pretende resolver con el conjunto de datos extraído. No obstante, el enfoque varía.

El proyecto analítico que se tiene entre manos es un encargo de un fondo de capital riesgo. La perspectiva que se emplea en el análisis de datos es la de un inversor interesado en

conocer las tendencias del mercado para diferenciarse del resto y buscar oportunidades donde nadie está buscando, obteniendo así mayores probabilidades de éxito.

El fondo no solo quiere obtener información de *startups*, sino también de perfiles como el suyo que le permita estar siempre un paso por delante, e incluso contemplar levantar fondos apoyándose en fondos de fondos que inviertan en empresas similares.

El **abanico de análisis** que se puede realizar con el conjunto de datos obtenidos en este proyecto es muy amplio, **al igual que la perspectiva** desde la que se parte. Y este **es un factor diferenciador** con los ejemplos citados en el apartado anterior.

8. Licencia

La mayoría de datos publicados en repositorios como Kaggle tienen la licencia desconocida, e indican en la descripción cómo se ha obtenido el conjunto de datos resultante.

Por lo tanto, el *dataset* resultante de este ejercicio se publicaría bajo “Unknown Licence”, dado que se ha obtenido de una forma no permitida por el propietario del sitio web.

9. Código

El proyecto se ha desarrollado en Jupyter Notebook, permitiendo la inclusión de texto, así como la ejecución de código a través del navegador.

Enlace al repositorio Git: https://github.com/makhfib/cb_organizations.

10. Dataset

DOI [10.5281/zenodo.5572871](https://zenodo.org/record/5572871)

DOI: <https://zenodo.org/record/5572871>

Dedicación

Contribuciones	Firma
Investigación previa	MM, NM
Redacción de las respuestas	MM, NM
Desarrollo del código	MM, NM