

UOC - Tipología y Ciclo de vida de los datos

Victor María Cardoner Alvarez

Jose Oriol Bielsa Nogaledo

1. Contexto

Enisa (Empresa Nacional de Innovación, S.A.), entidad pública dependiente del Ministerio de Industria, Comercio y Turismo, concede, de manera anual, diversas líneas de **financiación para impulsar la actividad emprendedora**.

Los préstamos, que se conceden desde 1995, se organizan en diferentes líneas en función de la madurez de la empresa, por importes de entre 25.000€ y 1.5M€ y plazos hasta 9 años. Estos préstamos tienen además otras ventajas que los hacen muy interesantes para empresas en sus etapas iniciales, como la carencia hasta 7 años, la posibilidad de deducir los intereses del impuesto de sociedades, la remuneración en función de los resultados o que no requieren garantías ni avales.

El objetivo de este proyecto es analizar la distribución geográfica y sectorial de estos préstamos, las empresas a las que se les ha concedido (madurez, tamaño, sector, actividad en redes sociales), y la evolución de todos estos parámetros en el tiempo para determinar su impacto en la actividad de las empresas que los reciben.

2. Título del dataset

El título del dataset es "**Empresas financiadas por Enisa**"

3. Descripción del dataset

El dataset lo hemos construido sobre la base de la información proporcionada por Enisa y posteriormente lo hemos enriquecido con diferentes fuentes de datos mercantiles y redes sociales.

ENISA

La base de datos de Enisa contiene todo el histórico de préstamos concedidos desde 1995 y se pueden consultar en la página <https://www.enisa.es/es/comunidad-enisa/prestamos>. En la actualidad constan 7063 préstamos concedidos etiquetados con los siguientes campos: **Razón Social, Marca comercial, Importe, Fecha, Comunidad Autónoma y Provincia**. Este listado lo hemos obtenido haciendo scraping directo, y es el núcleo de datos sobre el que hemos trabajado.

Dada la limitación en el tiempo de ejecución, hemos limitado la extracción a los 120 préstamos más recientes. En cualquier caso, con un pequeño ajuste sobre el código Python se podría ampliar a voluntad hasta extraer la totalidad de datos, si la empresa u organización que utilizase este proceso dispone del tiempo y potencia necesaria para su ejecución.

INFOCIF y AXESOR

Seguidamente, hemos enriquecido los datos de Enisa con plataformas de datos mercantiles como Infocif.es y Axesor.es. El objetivo inicial es añadir el CIF de la empresa para una búsqueda más sencilla de otros datos como dirección, fecha de constitución, código de actividad CNAE y SIC.

Usando **Infocif** hemos recuperado unos 5000 **CIFs** del total de los préstamos. Con el CIF se puede recuperar información en Axesor (o Informa) usando Selenium y BeautifulSoup, aunque su fichero robots.txt desincentiva su uso vía scraping (son plataformas de pago con API propia pero con límite de consultas tanto en la versiones gratis como de pago). La idea es hacer una extracción de un número bajo de CIFs como muestra de la operativa, pero no extraer los 5000 datapoints para no incumplir las recomendaciones del fichero robots.

Usando Selenium, hemos generado la URL estática de cada empresa con la que, posteriormente, usando BeautifulSoup, hemos extraído 4 campos: **dirección completa, fecha de constitución de la empresa, CNAE y SIC**. Estos dos últimos campos se refieren al sector de actividad de la empresa.

TWITTER y LINKEDIN

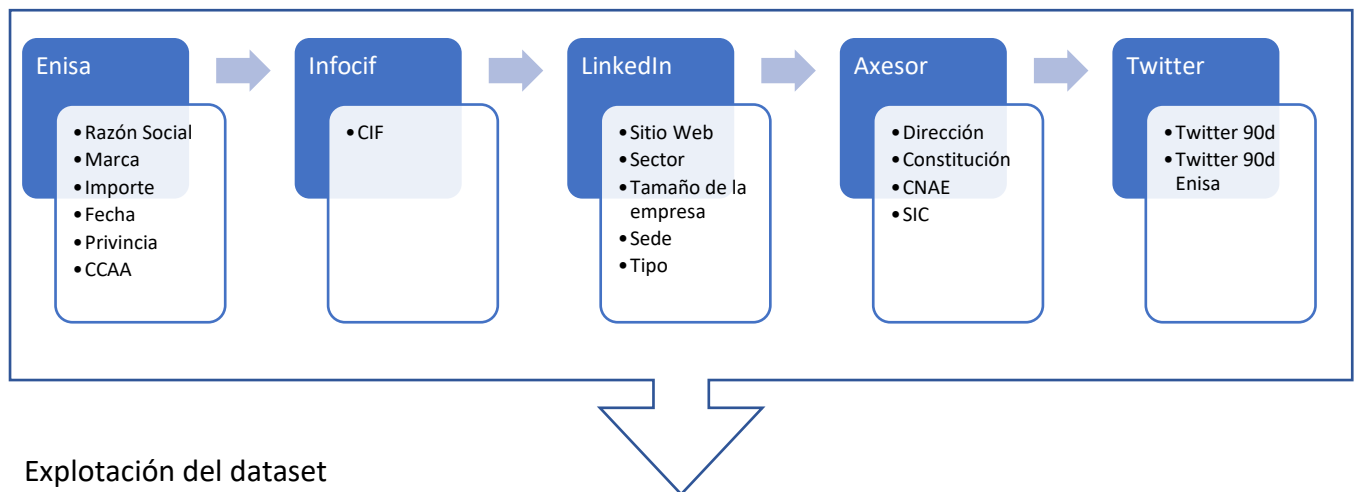
Para poder continuar con el estudio, el objetivo es analizar la actividad de LinkedIn y Twitter como indicadores de salud de la empresa.

En **LinkedIn** hemos hecho scraping usando Selenium, lanzando consultas directamente desde su buscador usando el campo Marca Comercial extraído de Enisa. De esta manera, hemos extraído la siguiente información de la empresa: **web, sector, tamaño, sede y tipo**.

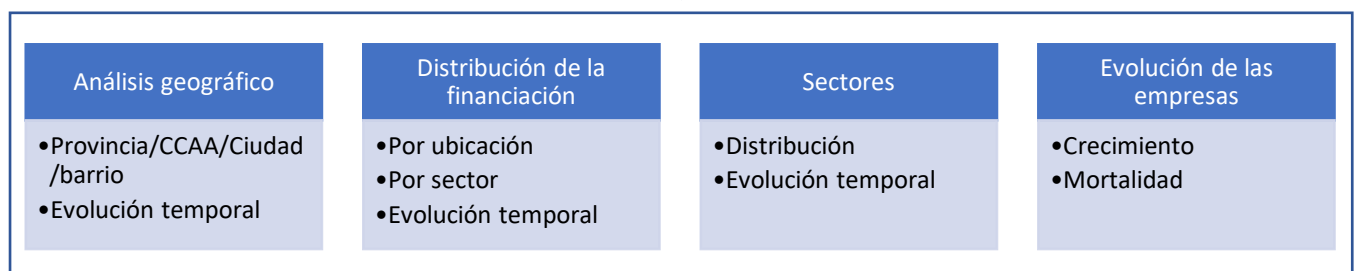
Por otra parte, para **Twitter** usamos la librería snsscrape - evitando el uso de la API. En esta plataforma hemos identificado dos vías para la explotación de datos: o bien fuerza bruta - relacionando la Marca Comercial con el handle de Twitter-, o bien buscando por el hashtag #clienteEnisa - usada por la propia Enisa para hacer difusión de las empresas a las que concede préstamos-. El objetivo final ha sido extraer un indicador básico de presencia o actividad en redes: número de **menciones en tweets en los últimos 90 días**.

4. Representación gráfica del dataset

Creación del dataset



Explotación del dataset



5. Contenido

A continuación, presentamos el contenido del dataset, así como el origen y el método usado para el scraping:

Fuente	Técnicas Usada	Campos	Descripción
ENISA https://www.enisa.es/	Scraping con BeautifulSoup Uso de user agent modificado	Razón Social Marca Importe Fecha Provincia CCAA.	Nombre con que la empresa está registrada legalmente en el registro mercantil Nombre comercial de la empresa Importe del préstamo concedido por ENISA Fecha de concesión del préstamo Provincia en la que está registrada la empresa Comunidad autónoma en la que está registrada la empresa
INFOCIF https://www.infocif.es/	Scraping con BeautifulSoup	CIF	El Código de Identificación Fiscal sirve para la identificación fiscal de las empresas, y es utilizado por las personas jurídicas, tanto empresas como fundaciones. Es un identificador único que facilita la búsqueda en los diferentes servicios de información mercantil
LINKEDIN https://www.linkedin.com/	Automatización con Selenium via chromedriver Scraping con BeautifulSoup	Sitio web Sector Tamaño de la empresa Sede Tipo	Dirección web de la empresa Sector empresarial Clasificación por volumen de trabajadores Sede empresarial Tipo de empresa
AXESOR https://www.axesor.es/	Automatización con Selenium via chromedriver Scraping con BeautifulSoup Uso de user agent aleatorio	Dirección Constitución CNAE SIC	Dirección postal de la empresa Fecha de constitución de la empresa Sistema de clasificación de las actividades de las empresas elaborado por el INE. Las siglas corresponden a Clasificación Nacional de Actividades Económicas Sistema Internacional de Clasificación de actividades empresariales elaborado por la Administración de Estados Unidos
TWITTER https://twitter.com/	Librería snsrape de Python	Twitter_90d Twitter_90d_enisa	Volumen de menciones - últimos 90 días Volumen de menciones con #clienteEnisa - últimos 90 días

Consideraciones adicionales:

- (1) Al usar Selenium, es necesaria la instalación de chromedriver (<https://chromedriver.chromium.org/downloads>). El programa de scraping requerirá que se introduzca la ruta de instalación por línea de comandos para poder ejecutarlo correctamente
- (2) LinkedIn permite consultar todos sus perfiles públicos, a partir de un perfil creado. Por tanto, para realizar este scraping, es necesario introducir credenciales válidas de usuario de LinkedIn. Esto también se requerirá por línea de comandos en la ejecución.

(Estas dos cuestiones, ante un hipotético despliegue en proceso batch, se podrían ajustar para no requerir de interacción por línea de comandos. Se podría usar algún tipo de fichero de configuración, o bien directamente parametrizarlo hardcode en el propio código.)

- (3) La librería snsrape para Twitter se debe instalar en modo dev, para permitir explotar sin limitaciones las herramientas que ofrece. Para esto, se debe hacer la instalación de la siguiente manera:

```
pip install git+https://github.com/JustAnotherArchivist/snsrape.git
```

6. Agradecimientos

Nuestro principal agradecimiento es para ENISA, una entidad de financiación dependiente del Ministerio de Industria, Comercio y Turismo. Esta empresa tiene por objetivo colaborar activamente en la financiación de proyectos empresariales que supongan una innovación en su ámbito.

Como parte de su actividad, esta entidad recoge los datos relativos a los préstamos concedidos de forma ordenada y pública, para su consulta y explotación sin restricciones -tal y como se puede observar en su *robots.txt*-.

Dado que el punto de partida de nuestra práctica es el estudio de las empresas con préstamos concedidos por ENISA, esto no sería posible sin los datos abiertos que esta comparte.

Adicionalmente, a nivel técnico también queremos agradecer a los creadores y desarrolladores de la librería snsrape, que permite un scraping fácil y sin limitaciones de Twitter, mejorando las capacidades de la API oficial de la red social.

7. Inspiración

La inspiración principal que nos ha movido a contruir este dataset es el informe que, por octavo año consecutivo, Enisa ha encargado al catedrático de Economía Financiera de la UCM, José Martí Pellón. El informe es un ejercicio de seguimiento y valoración del impacto económico y social de los préstamos de Enisa, que respalda la tesis de que el préstamo participativo es un instrumento financiero que ofrece ventajas para pymes y emprendedores.

Al obtener información sobre qué líneas de financiación tienen mayor impacto económico y social, en qué sectores, en qué fase del desarrollo de la empresa, etc., se dispone de elementos de valoración para afianzar y respaldar la concesión de estos préstamos y ayudan a definir y desarrollar nuevas iniciativas.

El estudio concluye que estos préstamos contribuyen de manera positiva en las empresas que los reciben y fomentan una mayor creación de empleo, un crecimiento más rápido, una mayor facturación y un menor mortalidad comparadas con un grupo de control de empresas similares sin esta financiación.

Hemos identificado multitud de áreas en las que este dataset puede ser usado para obtener conocimiento sobre el ecosistema emprendedor nacional.

- **Ubicación geográfica de los principales polos de emprendimiento a nivel nacional:** analizando la ubicación de las empresas receptoras de estos préstamos, podemos ubicarlas a nivel de comunidad autónoma, de provincia e incluso de dirección exacta en las ciudades. Extrapolando estos datos, podemos identificar las regiones, ciudades y barrios con mayor concentración de iniciativas emprendedoras. Añadiendo la variable temporal, podemos identificar también regiones pujantes y las que pierden peso en actividad emprendedora.
- **Distribución de la financiación:** analizando los importes concedidos, podemos determinar qué regiones/provincias/ciudades y qué sectores de actividad reciben más fondos. Añadiendo la variable temporal, se puede analizar la evolución de estos flujos de financiación en el tiempo.
- **Sectores:** analizando el CNAE, el SIC y el sector descrito en LinkedIn, podemos identificar los sectores con mayor empuje emprendedor.
- **Evolución temporal:** el dataset y su rutina de generación, están pensados para hacer un seguimiento periódico de las empresas. Actualizando el dataset de manera mensual o trimestral, se puede hacer un seguimiento de la evolución de estas empresas a nivel de empleados (campo de LinkedIn), actividad en RRSS e incluso mortalidad.

Para analizar este último punto y, en ausencia de un grupo de control, nos podríamos basar en **estudios que miden la mortalidad de las startups en sus etapas iniciales**, para comparar estas métricas con las observadas entre los receptores de los préstamos Enisa.






Un estudio ya clásico encabezado por investigadores de las Universidades de Berkeley y Stanford que analizó más de 3200 start-ups, el Startup Genome Project, cifraba la tasa de fracaso de estas empresas en la cota del 90%, y en el caso del 10% que sí tenían éxito también sufrían experiencias de crisis graves que las expusieron a la desaparición. Estos datos han venido siendo confirmados, con leves diferencias estadísticas, por análisis posteriores de Bloomberg, Forbes o Kaufman Foundation.

Analizando la evolución del dataset, podríamos identificar patrones de decrecimiento (o desaparición) de la actividad en RRSS, o ausencia de datos en las bases de datos mercantiles o LinkedIn, que fueran indicadores de la desaparición de la empresa. Una tasa inferior al 90% de fracaso que arrojan los estudios podría indicar el impacto positivo de los préstamos en la evolución de las empresas.

En resumen, este dataset pretende ser una **herramienta de trabajo y control para identificar las tendencias y problemas** comentados. Entendemos que con la información que proporciona, se podrían generar diversas métricas, indicadores, visualizaciones e informes que permitirían tener una visión detallada del ecosistema emprendedor.

8. Licencia

Sabemos que existen distintos tipos de licencias Creative Commons (CC), caracterizadas por distintos atributos estandarizados:

- **BY**  – Implica citar el propietario del trabajo original
- **NC**  – No permite el uso comercial del material
- **ND**  – No permite el uso del material para adaptaciones o trabajos derivados
- **SA**  – Implica que cualquier adaptación de este material se debe compartir con esta misma licencia
- **CC0**  - Licencia abierta universal, sin restricciones ni limitaciones

La combinación de estos elementos da lugar a las distintas licencias CC disponibles: CC BY, CC BY-SA, CC BY-NC, CC BY-NC-SA, CC BY-ND, etc.

En el caso que nos ocupa, vamos a optar por usar **CC BY-NC-SA 4.0 License**.



Como comentábamos anteriormente, esta licencia implica el **reconocimiento del autor** de los datos, **no permitiendo el uso comercial** de estos, y forzando que trabajos derivados de este sean **licenciados bajo estas mismas condiciones**. Creemos que esto encaja con la filosofía con la que hemos generado este dataset, teniendo en cuenta adicionalmente que nos encontramos en un ámbito académico.

9. Código

El detalle del código se encuentra en la carpeta **scr** repositorio GitHub de la entrega:
<https://github.com/oriolbielsa/uoc-prestamos-enisa/tree/main/src>

10. Dataset

El dataset final se encuentra en la carpeta **csv** repositorio GitHub de la entrega:
<https://github.com/oriolbielsa/uoc-prestamos-enisa/tree/main/csv>

Adicionalmente, también se ha subido a Zenodo, dónde se encuentra registrado con el siguiente DOI:

DOI 10.5281/zenodo.4677114

También se puede acceder desde el siguiente enlace:
<https://zenodo.org/record/4677114#.YHG6VOj7TIU>

Referencias

Pellón, J.M. (Diciembre 2020). Valoración del impacto económico y social de los préstamos otorgados por ENISA entre 2005 y 2016. <https://www.enisa.es/es/sala-de-prensa/estudios-informes/impacto-economico-y-social-de-los-prestamos-enisa-262>

Startup Genome (2020). The Global Startup Ecosystem Report - GSER 2020.
<https://startupgenome.com/report/gser2020>

Fairlie, R., Desai, S. State report on early stage entrepreneurship in the United States: 2020. https://indicators.kauffman.org/wp-content/uploads/sites/2/2021/03/2020_Early-Stage-Entrepreneurship-State-Report.pdf

Beck, Martin (2020). How to Scrape Tweets From Tweeter.
<https://towardsdatascience.com/how-to-scrape-tweets-from-twitter-59287e20f0f1>

GitHub (2021) – Documentation snsscape.
<https://github.com/JustAnotherArchivist/snsscape>

Beck, Martin (2020). How to Scrape Tweets With snsrape.

<https://betterprogramming.pub/how-to-scrape-tweets-with-snsrape-90124ed006af>

Mitchel, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly Media, Inc.

Participación

Contribuciones	Firma
Investigación previa	Víctor María Cardoner Álvarez José Oriol Bielsa Nogaledo
Redacción de las respuestas	Víctor María Cardoner Álvarez José Oriol Bielsa Nogaledo
Desarrollo del código	Víctor María Cardoner Álvarez José Oriol Bielsa Nogaledo