



UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática



**TFG del Grado en Ingeniería
Informática**

**Aplicación Web para la
recopilación, tratamiento y
visualización de datos
públicos**



Presentado por Iván Arjona Alonso
en Universidad de Burgos — 2 de mayo
de 2018

Tutores: Dr. José Francisco Díez Pastor
y Dr. Jesús Manuel Maudes Raedo



UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería Informática



D. José Francisco Díez Pastor y D. Jesús Manuel Maudes Raedo, profesores del departamento de Ingeniería Civil, área de Lenguajes y Sistemas Informáticos.

Expone:

Que el alumno D. Iván Arjona Alonso, con DNI 71352655P, ha realizado el Trabajo final de Grado en Ingeniería Informática titulado “Aplicación Web para la recopilación, tratamiento y visualización de datos públicos”.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 2 de mayo de 2018

Vº. Bº. del Tutor:

Vº. Bº. del co-tutor:

José Francisco Díez Pastor

D. Jesús Manuel Maudes Raedo

Resumen

En este primer apartado se hace una **breve** presentación del tema que se aborda en el proyecto.

Descriptores

Palabras separadas por comas que identifiquen el contenido del proyecto Ej: servidor web, buscador de vuelos, android ...

Abstract

A **brief** presentation of the topic addressed in the project.

Keywords

keywords separated by commas.

Índice general

Índice general	III
Índice de figuras	IV
Índice de tablas	V
Introducción	1
Objetivos del proyecto	3
2.1. Objetivos funcionales	3
2.2. Objetivos técnicos	4
Conceptos teóricos	5
3.1. Bases de datos no relacionales	5
3.2. Web scraping	5
3.3. Mapas coropléticos	6
Técnicas y herramientas	7
4.1. Herramientas	7
Aspectos relevantes del desarrollo del proyecto	9
Trabajos relacionados	11
Conclusiones y Líneas de trabajo futuras	13
Bibliografía	15

Índice de figuras

3.1. Mapa coroplético del paro en las provincias españolas	6
----------------------------------------------------------------------	---

Índice de tablas

Introducción

La *minería de datos* es una metodología para descubrir relaciones ocultas en grandes cantidades de datos utilizando técnicas de inteligencia artificial.

Un tema que está siendo investigado actualmente es comprender las causas por las que se producen crímenes. Por ello, se han realizado estudios para investigar las tasas de criminalidad utilizando técnicas de minería de datos [1] [2] a partir de datos demográficos y sociológicos con el objetivo de determinar los lugares donde asignar más recursos para reducir estos índices de criminalidad.

El objetivo de este trabajo es llevar estos estudios al ámbito español. Integrando varias fuentes de datos públicas utilizando técnicas de web scraping (3.2). Para después visualizar estos datos de manera sencilla, permitir crear mapas temáticos (3.3) con esta información y aplicar técnicas de minería de datos para buscar relaciones entre los datos.

Fuentes de datos

A continuación se van a listar las fuentes de datos de organismos públicos que se han integrado.

- Servicio Público de Empleo Estatal (SEPE)
 - Paro registrado por municipio¹.
 - Contratos registrados por municipio².
 - Demandantes de empleo por municipio³.

¹<https://datos.gob.es/catalogo/e00142804-paro-registrado-por-municipios>

²<http://datos.gob.es/es/catalogo/e00142804-contratos-por-municipios>

³<http://datos.gob.es/es/catalogo/e00142804-demandantes-de-empleo-por-municipios>

- Estadísticas de la renta por municipio (Agencia tributaria)⁴.
- Instituto Nacional de Estadística (INE)
 - Estadísticas de población por sexo y edad⁵.
 - Relación de municipios y códigos por provincias⁶.

Material entregado

Material adjunto a la memoria:

- Aplicación web en Flask.
- Scripts para la integración de fuentes de datos.
- Scripts para la ejecución.
- Anexos.

Recursos disponibles en internet:

- [Página web del proyecto](#)⁷.
- [Repositorio del proyecto](#)⁸.

⁴https://www.agenciatributaria.es/AEAT.internet/datosabiertos/catalogo/hacienda/Estadistica_de_los_declarantes_del_IRPF_por_municipios.shtml

⁵<http://www.ine.es/jaxi/Tabla.htm?path=/t20/e245/p05/a2011/10/&file=00000001.px&L=0>

⁶<http://www.ine.es/daco/daco42/codmun/codmunmapa.htm>

⁷<https://tfg-datos-publicos.nanoapp.io/>

⁸<https://github.com/IvanArjona/TFG-Datos-publicos>

Objetivos del proyecto

Este apartado explica de forma precisa y concisa cuales son los objetivos que se persiguen con la realización del proyecto. Se puede distinguir entre los objetivos marcados por los requisitos del software a construir y los objetivos de carácter técnico que plantea a la hora de llevar a la práctica el proyecto.

2.1. Objetivos funcionales

- Integrar varias fuentes de datos públicas en una única base de datos. Estos datos son datos de carácter sociológicos, económicos y demográfico a nivel municipal en España.
- Permitir añadir nuevos conjunto de datos de forma sencilla.
- Desarrollar un algoritmo para la carga de varias fuentes de datos en una base de datos de manera automatizada.
- Desarrollo de una aplicación web que permita la consulta de los datos de manera sencilla y visual.
- Facilitar la interpretación de los datos utilizando un mapas coropléticos interactivo.
- Desplegar la aplicación web en un servidor de forma que sea fácil de actualizar cada vez que se realice un cambio. Además de funcionar en un entorno local.

2.2. Objetivos técnicos

- Seguimiento de principios de desarrollo ágiles durante el desarrollo del proyecto.
- Utilizar *Flask* como framework para desarrollar la aplicación web utilizando la arquitectura *Modelo Vista Controlador*.
- Utilizar *git* como sistema de control de versiones, junto con *Github* y *Zenhub* para la organización del proyecto.
- Utilizar herramientas para mejorar la calidad del código como *codacy*.
- Manejar bases de datos no relacionales *NoSQL*.
- Elegir herramientas de *software libre* siempre que sea posible.
- Proporcionar una instalación sencilla para el desarrollador y una interfaz visual y fácil de utilizar para el usuario.

Conceptos teóricos

En aquellos proyectos que necesiten para su comprensión y desarrollo de unos conceptos teóricos de una determinada materia o de un determinado dominio de conocimiento, debe existir un apartado que sintetice dichos conceptos.

3.1. Bases de datos no relacionales

Las bases de datos no relacionales [9], también llamadas NoSQL (‘Not Only SQL’) son bases de datos optimizadas para ser utilizadas con modelos de datos sin esquema y potencialmente escalables.

A diferencia de las bases de datos relacionales, aquí no hay tables, esquemas ni relaciones, sino que los datos pueden almacenarse con cualquier esquema sin tener que seguir todos la misma estructura.

En este tipo de bases de datos tenemos colecciones en lugar de tablas y dentro de estas colecciones tenemos documentos. Estos documentos pueden contener cualquier información y se almacenan siguiendo un esquema json.

Una de las razones importantes por las que usar bases de datos NoSQL en lugar de relacionales es la alta velocidad de consulta y la gran capacidad de escalabilidad.

3.2. Web scraping

Web Scraping [10] es una técnica para extraer información de sitios web directamente de su código fuente sin utilizar APIs⁹ proporcionadas por el

⁹Interfaz de programación de aplicaciones

propio sitio.

Lo que hacemos con un *web scraper* es buscar información dentro de un documento web siguiendo ciertos patrones en la estructura de su código fuente y extraer esta información a nuestro entorno local.

Las técnicas de web scraping se centran en transformar datos sin estructura de una página web en datos estructurados para poder ser almacenados y analizados posteriormente. Por ejemplo, en una base de datos, hojas de cálculo o dataframes de pandas.

3.3. Mapas coropléticos

Un mapa coroplético [8] es un mapa topológico dividido en regiones en el que cada una de estas regiones se pinta de un color de acuerdo a una medida estadística.

Para ilustrarlo mejor pondremos el ejemplo del siguiente mapa en la figura 3.1. En él se compara el paro de todas las provincias de España pintando con colores más cálidos las provincias con mayor porcentaje de paro y con colores más fríos las provincias con menor porcentaje de paro.

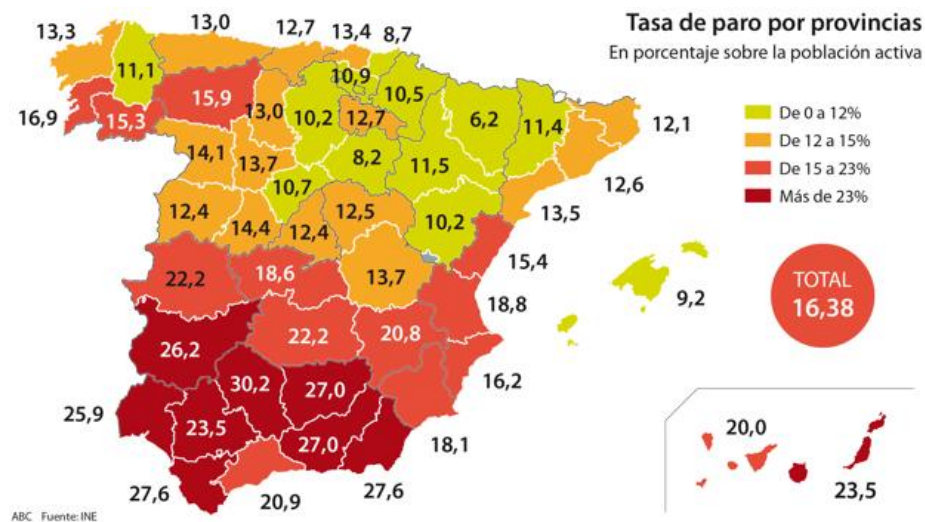


Figura 3.1: Mapa coroplético del paro en las provincias españolas [7]

Técnicas y herramientas

4.1. Herramientas

MongoDB

- Herramientas consideradas: [Riak](#), [Cassandra](#), [MongoDB](#), [LevelDB](#).
- Herramienta elegida: [MongoDB](#).

MongoDB [4] es un sistema de bases de datos NoSQL. En esta herramienta los datos se guardan en forma de documentos con un esquema similar a json. Con este sistema se consigue una consulta de datos más rápida. Además se ha usado PyMongo [3] como herramienta para integrar MongoDB en python.

Tanto Riak como Cassandra son también bases de datos NoSQL, se descartaron porque no ofrecen soportes para equipo con sistema operativo windows.

LevelDB es una base de datos NoSQL de pares clave-valor, esta herramienta se descartó porque no se cree conveniente utilizar pares clave-valor para un proyecto como este y no hay tantos ejemplos en la documentación como en las otras herramientas.

DigitalOcean

- Herramientas consideradas: [Heroku](#), [PythonAnywhere](#), [DigitalOcean](#), [Amazon Web Services](#).
- Herramienta elegida: [DigitalOcean](#).

Digital Ocean [6] es un proveedor de servidores privados, por ello podemos hacer lo que queramos con el servidor sin limitaciones más allá de la capacidad de procesamiento y memoria ram. Se ha elegido este servicio porque nos da total libertad y se puede probar gratuitamente con [GitHub Education](#).

Una alternativa que se consideró y de hecho, se probó es Heroku, en este caso se instala el entorno necesario de forma automática. El problema es que la base de datos en la capa gratuita sólo puede pesar 500MB como máximo y no es suficiente para este proyecto.

PythonAnywhere es un hosting para aplicaciones web en python, el problema con este proveedor es que no ofrece bases de datos locales, por lo que habría que utilizar una remota. La única gratuita que se ha encontrado es [mLab](#), la misma que usa Heroku, por lo que volvemos al mismo problema del límite de tamaño.

Por último, se consideró utilizar una instancia de [Amazon AWS EC2](#). Es muy similar a DigitalOcean, se eligió el primero porque es más sencillo de utilizar.

Nanobox

- Herramientas consideradas: [Heroku](#), [Nanobox](#).
- Herramienta elegida: [Nanobox](#).

Nanobox [5] es una herramienta que nos permite desplegar nuestra aplicación sin centrarnos en la infraestructura del servidor.

Para ello enlazamos nuestra cuenta de un proveedor en la nube (en este caso DigitalOcean) y nanobox se encargará de instalar el sistema operativo, configurarlo, instalar nuestra aplicación y sus dependencias y ejecutarla.

Heroku es un servicio muy similar, con la diferencia de que no podemos utilizar servidores en la nube externos, se descartó por lo ya explicado en el punto anterior.

Aspectos relevantes del desarrollo del proyecto

Este apartado pretende recoger los aspectos más interesantes del desarrollo del proyecto, comentados por los autores del mismo. Debe incluir desde la exposición del ciclo de vida utilizado, hasta los detalles de mayor relevancia de las fases de análisis, diseño e implementación. Se busca que no sea una mera operación de copiar y pegar diagramas y extractos del código fuente, sino que realmente se justifiquen los caminos de solución que se han tomado, especialmente aquellos que no sean triviales. Puede ser el lugar más adecuado para documentar los aspectos más interesantes del diseño y de la implementación, con un mayor hincapié en aspectos tales como el tipo de arquitectura elegido, los índices de las tablas de la base de datos, normalización y desnormalización, distribución en ficheros³, reglas de negocio dentro de las bases de datos (EDVHV GH GDWRV DFWLYDV), aspectos de desarrollo relacionados con el WWW... Este apartado, debe convertirse en el resumen de la experiencia práctica del proyecto, y por sí mismo justifica que la memoria se convierta en un documento útil, fuente de referencia para los autores, los tutores y futuros alumnos.

Trabajos relacionados

Este apartado sería parecido a un estado del arte de una tesis o tesina. En un trabajo final grado no parece obligada su presencia, aunque se puede dejar a juicio del tutor el incluir un pequeño resumen comentado de los trabajos y proyectos ya realizados en el campo del proyecto en curso.

Conclusiones y Líneas de trabajo futuras

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

Bibliografía

- [1] Francisco A. Rodrigues Luiz G.A. Alves, Haroldo V. Ribeiro. Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, 505:435–443.
- [2] Mohammad Reza Ebrahimi Mohammad Reza Keyvanpour, Mostafa Javideh. Detecting and investigating crime by means of data mining: a general crime matching framework. *Procedia Computer Science*, 3:872–880.
- [3] MongoDB. Pymongo 3.6.1 documentation. <https://api.mongodb.com/python/3.6.1/>. [Internet; descargado 22-abril-2018].
- [4] MongoDB. What is mongodb? <https://www.mongodb.com/what-is-mongodb>. [Internet; descargado 22-abril-2018].
- [5] Nanobox. Nanobox documentation. <https://docs.nanobox.io/>. [Internet; descargado 22-abril-2018].
- [6] Digital Ocean. Product documentation. <https://www.digitalocean.com/community/tags/product-documentation/tutorials>. [Internet; descargado 22-abril-2018].
- [7] Roberto Pérez. Cómo funcionan las mongodb injection. <http://www.abc.es/economia/abci-mapa-paro-espana-peores-y-mejores-provincias-para-encontrar-trabajo-201710-noticia.html>, octubre 2017. [Internet; descargado 22-abril-2018].
- [8] Wikipedia. Choropleth map — wikipedia, la enciclopedia libre. https://en.wikipedia.org/w/index.php?title=Choropleth_

- [map&oldid=832756109](#), marzo 2018. [Internet; descargado 23-abril-2018].
- [9] Wikipedia. Nosql — wikipedia, la enciclopedia libre. <https://en.wikipedia.org/w/index.php?title=NoSQL&oldid=836191884>, abril 2018. [Internet; descargado 23-abril-2018].
- [10] Wikipedia. Web scraping — wikipedia, la enciclopedia libre. https://en.wikipedia.org/w/index.php?title=Web_scraping&oldid=836897411, abril 2018. [Internet; descargado 23-abril-2018].