

**UNIVERSIDAD DE GUAYAQUIL**

**FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS**

**CARRERA**

**DE**

**SOFTWARE**

**MATERIA:**

PROCESOS DE SOFTWARE

**DOCENTE:**

Msg. Miguel Botto Tobar

**PROYECTO PRIMER PARCIAL**

Generador de nube de palabras a partir de archivos de bases de datos de los usuarios de StackOverflow en español

**INTEGRANTES:**

* Lozano Portilla Anthony
* Delgado Olea Giampaolo

**CURSO: SOF-S-MA-3-2**

**Introducción:**

Una nube de palabras o Word cloud es una imagen formada por palabras( extraídas de un archivo de texto o base de datos) de diferentes tamaños y colores cuya disposición[[1]](#footnote-1) puede ser dada por la frecuencia de las palabras o por practicidad de la imagen según el algoritmo que se implemente para renderizarla.

El proyecto consiste en generar una imagen WordCloud a partir de la base de datos de etiquetas de la plataforma **https://es.stackoverflow.com/**  loa cuales contienen entre otros atributos , los nombres de las etiquetas y el conteo de preguntas en la que aparece. El objetivo es generar la nube de palabras con los nombres de las etiquetas y basándose en su frecuencia respectiva siendo la etiqueta más popular en preguntas de la plataforma, la que aparezca más grande en la imagen .

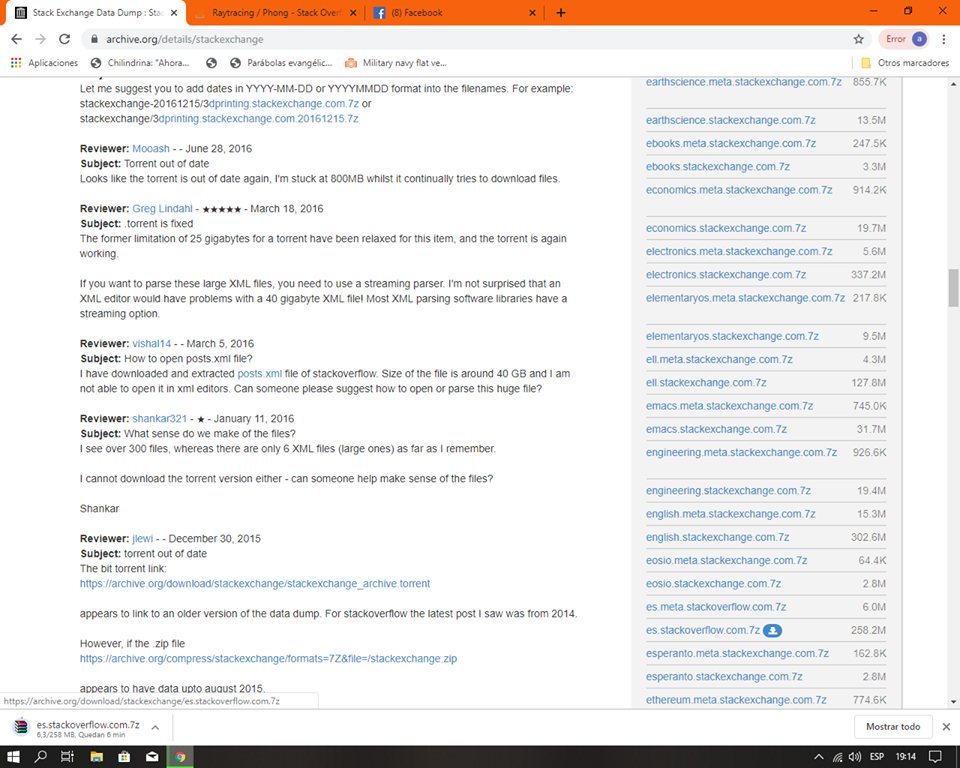
**Recursos:**

**Herramienta para búsqueda de archivos xml de base de datos:**

<https://archive.org/details/documentation-dump.7z>

**Dirección del archivo de la base de datos de usuarios de StackOverflow en español:**

<https://archive.org/details/stackexchange>



**Lenguaje en que se vá programar la solución :**

Python 3

**Entorno de desarrollo**:

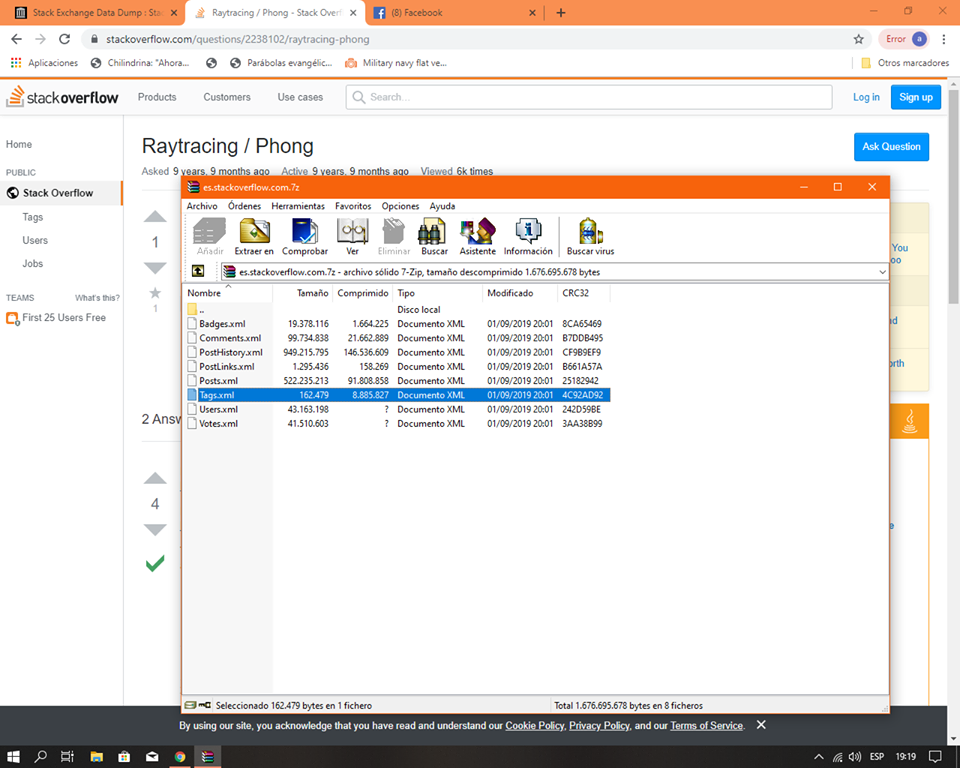
Pycharm

**Entorno de ejecución y compilación:**

Python 3.7 (venv) , entorno compartido con herramientas de desarrollo de visual c++

**Proceso:**

En primer lugar debemos descargar la base de datos **Tags.xml** la cual se encuentra dentro de la carpeta comprimida **es.StackOverflow.com** , en formato **xml** .



Antes de empezar a programar debemos convertir el archivo a formato **csv** debido a su practicidad para este proyecto, para esta tarea utilizamos una herramienta online: <http://www.convertcsv.com/xml-to-csv.htm> .

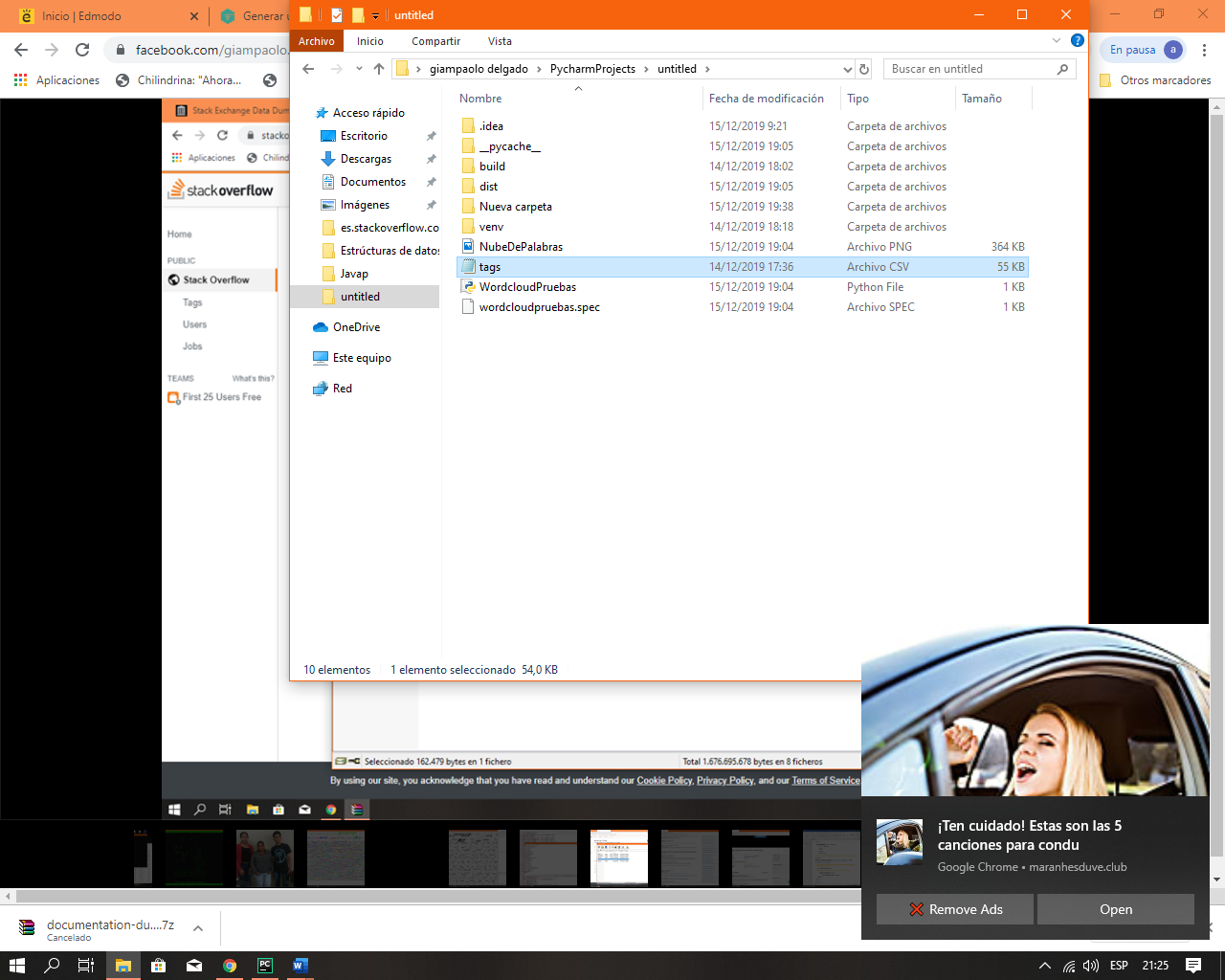


Ilustración 1Captura del archivo csv resultante ya agregada a la carpeta del proyecto

Una vez ya tenemos la información de la base de datos, nos adentramos en la programación, debido a que la solución que pretendemos implementar requiere un algoritmo de renderizado gráfico a demás de manipular texto y generar datos estadísticos , para esto utilizaremos una serie módulos externos de los cuales estos son los principales:

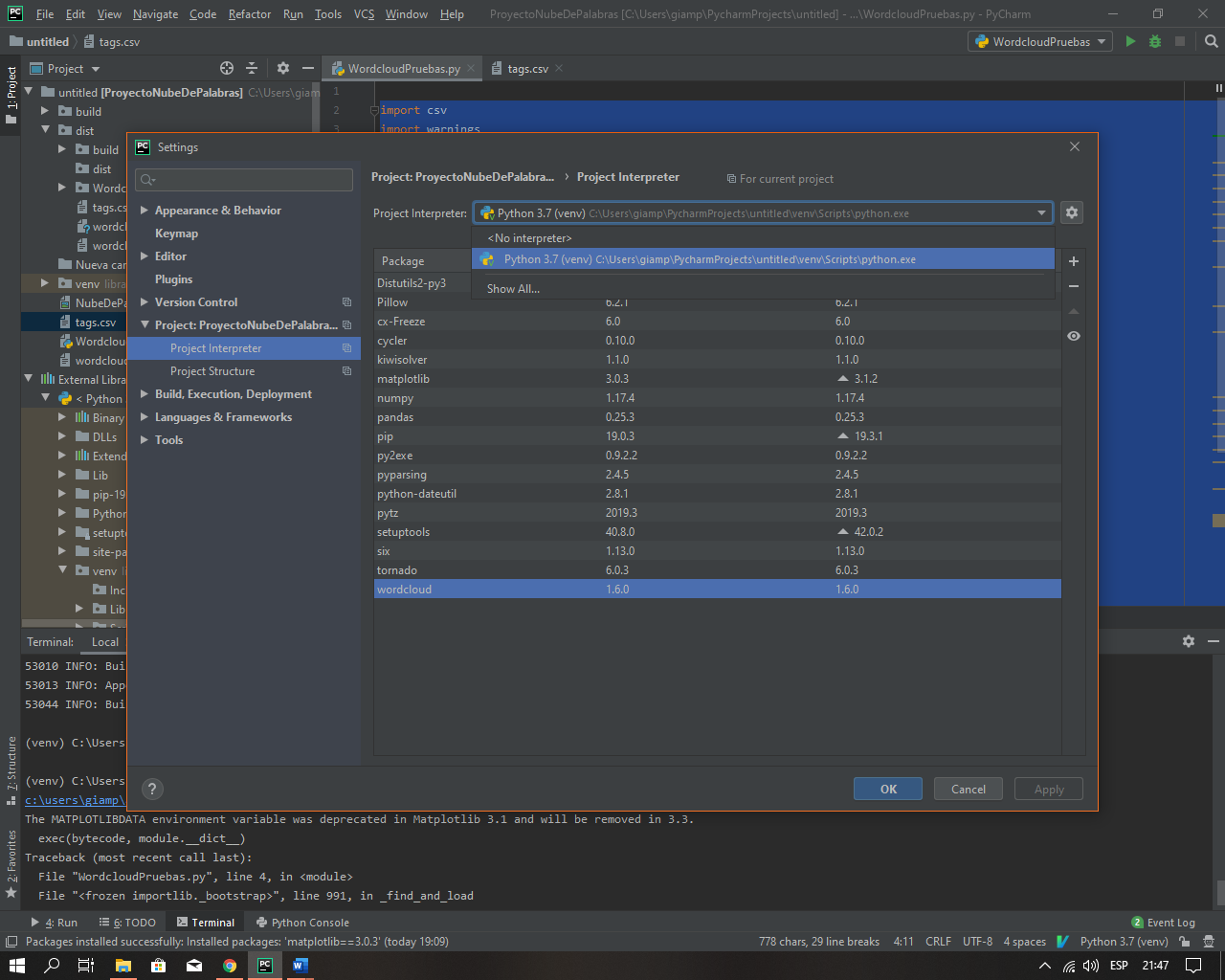
Numpy

Pillow

Matplotlib(genera una ventana con una serie de características donde se va a mostrar la imagen generada)

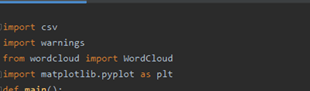
Wordcloud(se encarga de realizar los cálculos estadísticos y generar la imagen )

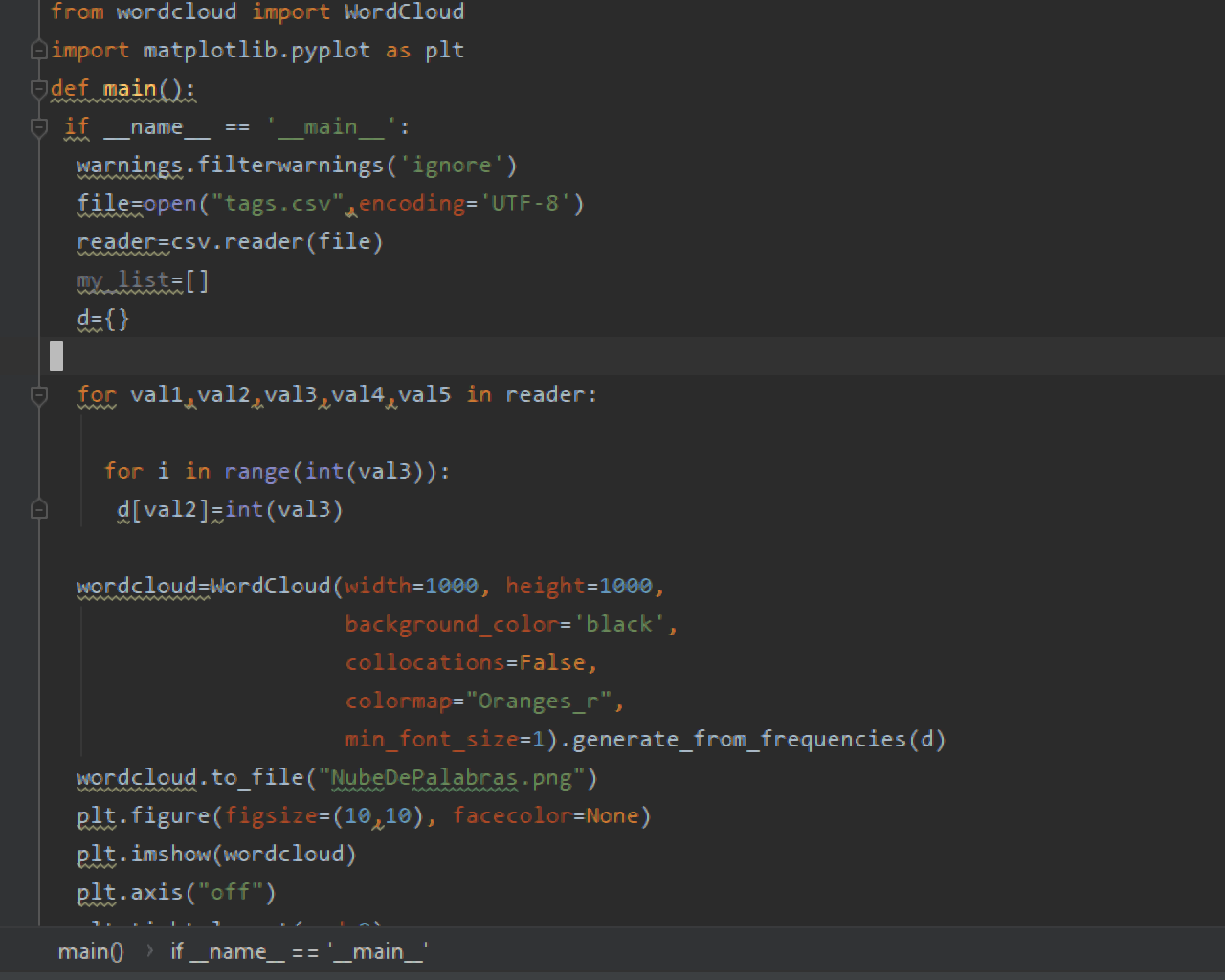
Una vez en pycharm creamos el proyecto configuramos el intérprete:



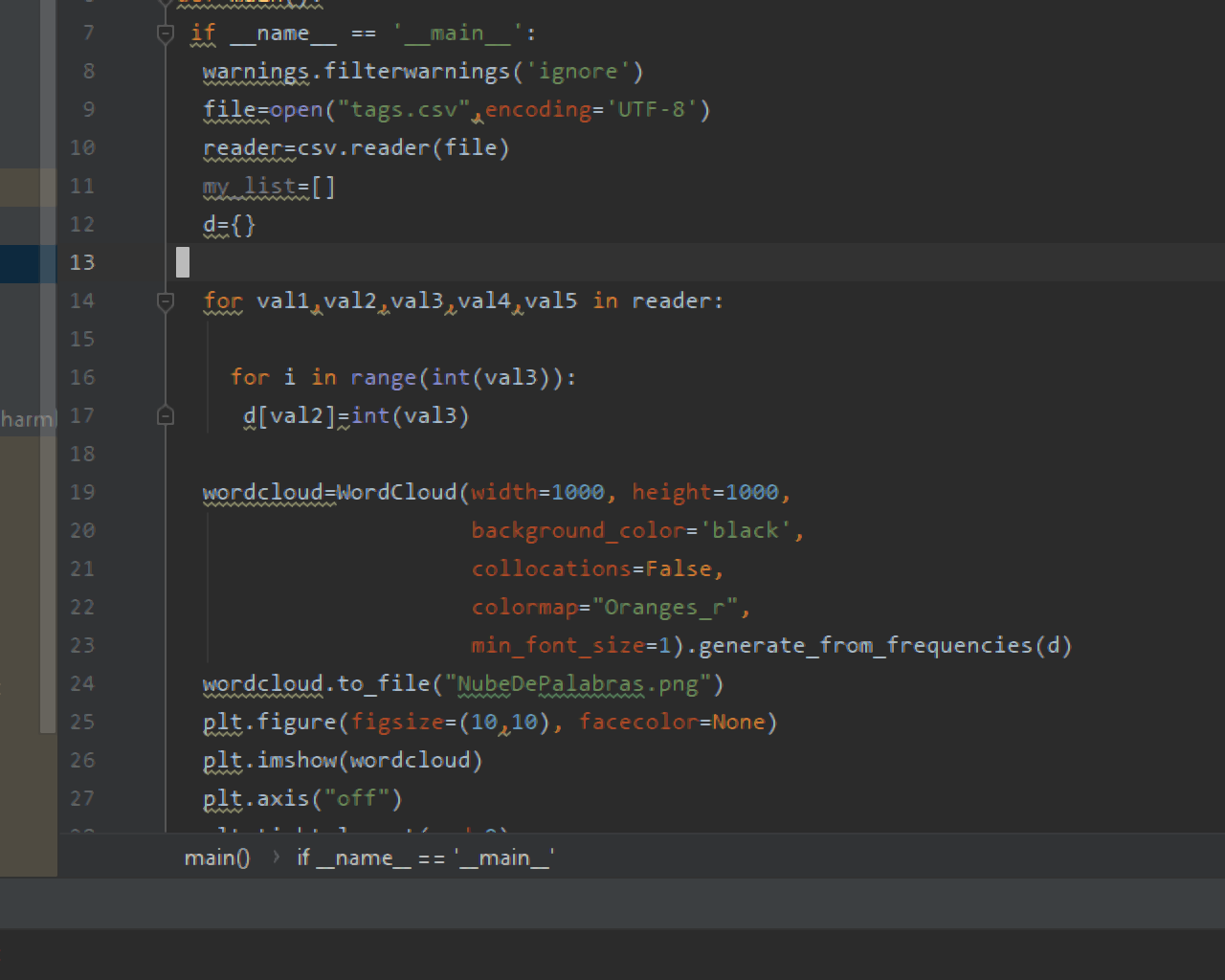
Como ya hemos mencionado estamos utilizando un módulo específicamente creado para generar wordclouds llamado: **wordcloud 1.6.0 creado por Andreas Mueller.** En este módulo podemos generar nubes de palabras de diferentes formas usando sus métodos y atributos, pero usaremos solo en método genrate\_from\_frequencies(diccionario de frecuencias de palabras)

Para averiguar sobre todos sus métodos y atributos podemos ingresar a: <https://amueller.github.io/word_cloud/generated/wordcloud.WordCloud.html>

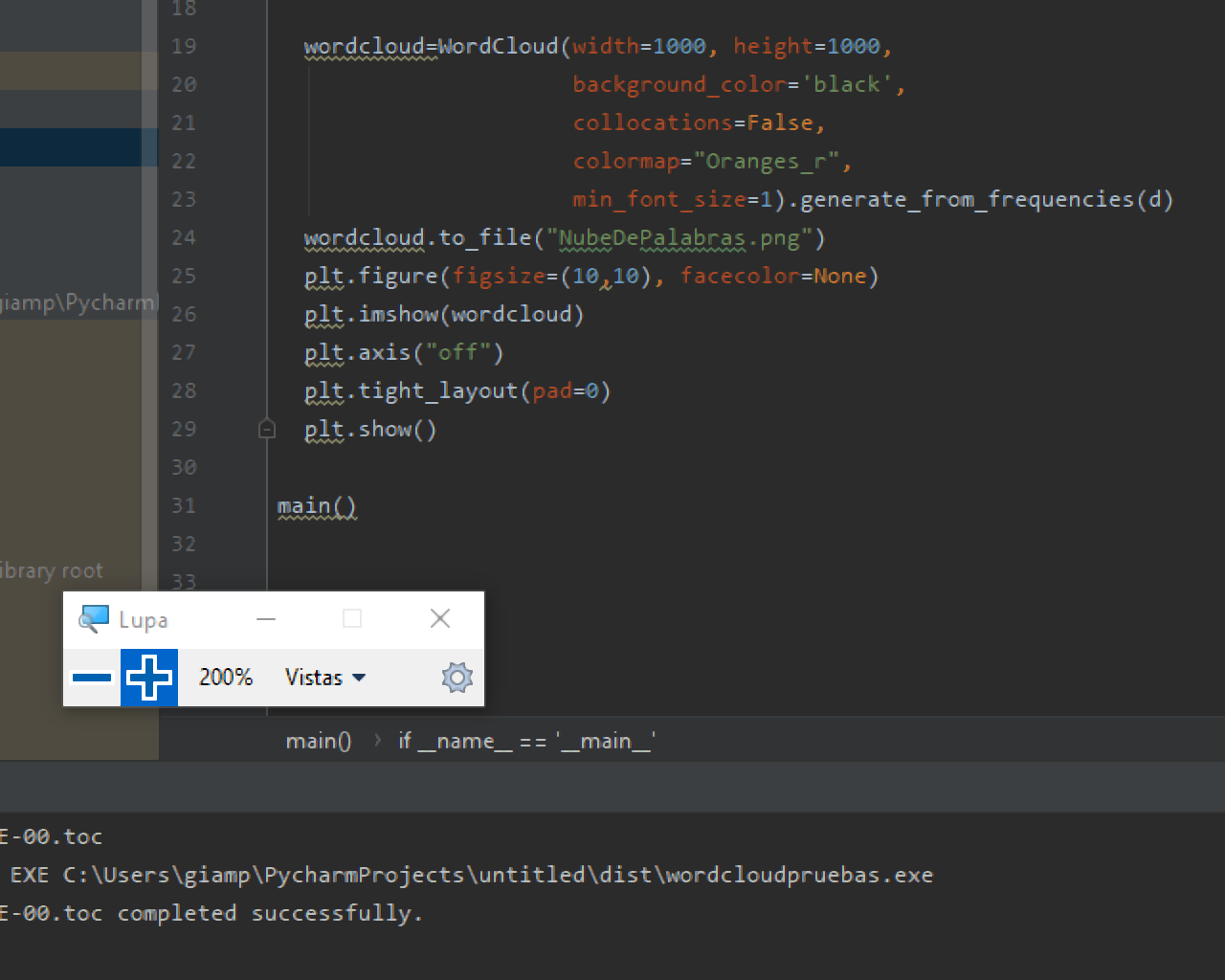
Para crear el diccionario iteramos la columna 2 y 3 del archivo Tags.csv las cuales contiene el nombre de la etiqueta y su frecuencia respectivamente.



Generamos la nube de palabras mediante el método **wordcloud.generate\_from\_frequencies(d)** usando como parámetro el diccionario de palabras y frecuencias y el método **wordcloud.to\_file(“NubeDePalabras.png”)** para generar un archivo con el nombre especificado como parámetro **.**



Finalmente creamos la ventana con la librería **matplotlib.pyplot** y empaquetamos el objeto wordcloud previamente generado para mostrarlo por pantalla.



**Al ejecutar nuestro script obtenemos la siguiente imagen como resultado:**

**Imagen que contiene texto, libro, colgante

Descripción generada automáticamente**

Ilustración 2 Imagen resultante generada de las etiquetas más populares de la plataforma stack overflow en español.

1. Wordcloud es un concepto más no un algoritmo, por lo tanto puede tener diferentes implementaciones , pero la característica más común es que el tamaño de las palabras es directamente proporcional a la frecuencia de estas en el texto o base de datos de dónde extraemos las palabras. [↑](#footnote-ref-1)