Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

WEB SCRAPING PARA LA DESCARGA DE INFORMACIÓN PROCEDENTE DE GOOGLE TRENDS PARA DNP+

Entidad

Departamento Nacional de Planeación

- Dirección General.
- Dirección de Desarrollo Digital.

Sector	Lenguaje	Fuente de datos
Planeación.	Python.	Google Trends

Presentación

DNP+ es una herramienta que se utiliza para pronosticar el PIB con anticipación a la publicación de la cifra oficial del DANE. Para lograr dicho pronóstico, hace uso de Google Trends y la información que permite conocer sobre las búsquedas de las personas en el territorio de análisis. El proceso de recolección de datos ha sido manual desde que se creó el proyecto de predicción por lo que se realizó un algoritmo que automatizara mensualmente la descarga de cada indicador por palabra y generara un archivo con el número de columnas equivalente a cada palabra utilizada para cada sector de proyección y cada fila como el valor en cada mes/año desde el 2004 a la fecha de ejecución.

Objetivo general

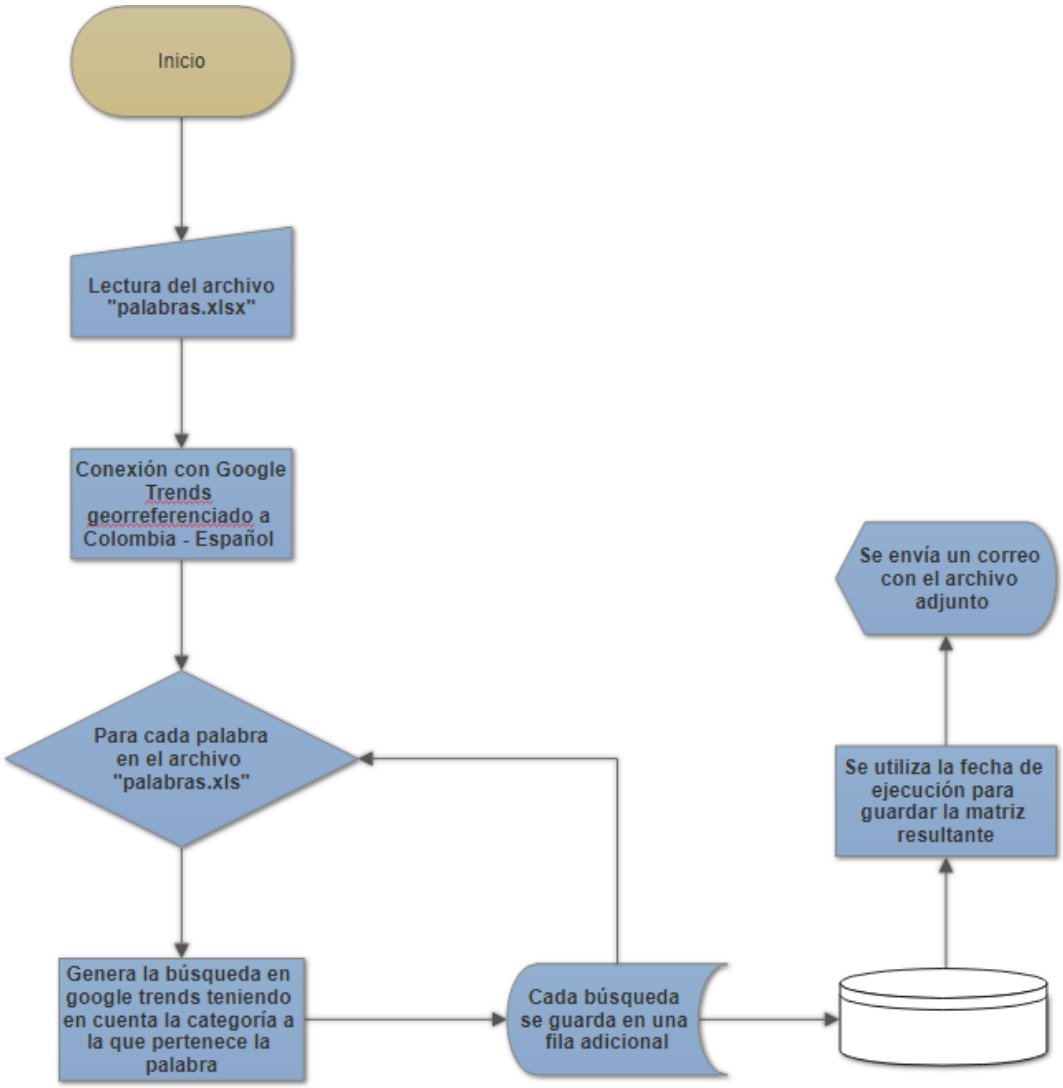
Facilitar la recolección de indicadores por término para cada sección de proyección.

Objetivos específicos:

- Automatizar la ejecución del algoritmo para que se realice una vez al mes
- Crear un archivo por cada ejecución del algoritmo y que automáticamente lo adjunte y envíe por correo electrónico

Metodología

Para la descarga automatizada se utilizó Python como lenguaje de programación que realiza la lectura de un archivo de palabras, seguida de una conexión con Google Trends que permite obtenerla categoría de pertenencia a las palabras y almacenar las búsquedas en un archivo. Las siguientes instrucciones que sigue el algoritmo se presentan a través de un diagrama de flujo a continuación:



Resultados

Se genera mensualmente, de forma automática, un archivo de texto plano que contiene los resultados de las búsquedas desde 2004 hasta el último mes y se envía por correo electrónico a los usuarios interesados.

Conclusiones

1. Este ejercicio aporta a la automatización y calibración de las regresiones que pronostican el comportamiento del PIB, por lo que el aporte de este algoritmo facilita uno de los procesos que más tiempo tomaban al momento de recalcular las proyecciones.
2. En el ejercicio de recolección y generación de datos para este algoritmo, la UCD identificó que los indicadores de popularidad cambian dependiendo de la ventana de tiempo que se esté consultando y esto afecta que los valores históricos sean los mismos en cada archivo; si el término es buscado atípicamente en un momento dado, los valores del índice de popularidad cambian drásticamente.

Socialización

Los archivos de salida de este proyecto son utilizados como insumo para el grupo asesor de la dirección general a cargo del proyecto DNP+.