

SISTEMA DE ALERTAS PARA NOTICIAS

EL  MUNDO

EL PAÍS



The Washington Post

europa
press

Le Monde

Miembros del grupo

[Alvaro Hidalgo López](#)

[Eugenio Palacios París](#)

[Raúl Sánchez de Saboya](#)

[Luis Sánchez Asunción](#)

Contenido

1. Introducción.	2
1.1. Herramientas, librerías y lógica usadas.....	2
2. Conclusiones.	4
3. Ideas de futuro y posibles mejoras.	5
4. Referencias.....	6
5. Anexos.....	6

1. Introducción.

La propuesta elegida es:

- **Sistema de Alerta de Noticias Personalizado:** Implementar un sistema que recopile noticias basadas en intereses específicos del usuario y las envíe de forma periódica.

Se ha elegido esta propuesta porque para este grupo era una idea interesante poder recoger noticias de gente de interés o acontecimientos sin tener que buscar entre el resto de las noticias usando palabras clave. De esta manera, se ahorra tiempo y se evitan distracciones con noticias que no son relevantes.

1.1. Herramientas, librerías y lógica usadas.

Dado que se va a obtener información de páginas de noticias que normalmente son estáticas, y esto quiere decir que no se va a tener que hacer captchas o algo parecido.

Pero, en este caso, no se elige la librería **BeautifulSoup** para realizar consultas de información en sus HTML, sino que, simplemente se recurre a **Requests**.

Esta hace las consultas mucho más simples y es ideal para acceder a información mediante la **NEWSAPI [2]** y saca formato texto o **JSON**, siendo este último este caso.

Para ello se coge la **NEWSAPI_KEY** que se carga de las variables de entorno.

La biblioteca **Transformers** nos sirve para coger el **pipeline** para usar modelos, cuya finalidad es hacer resúmenes de textos de las noticias.

Por otro lado, **login** sirve para que no salten ciertos avisos sobre los límites de texto (todo esto se aplica en la función **dynamic_summarize**).

Después se obtienen las noticias de dos maneras:

1. A través de la NEWSAPI:

Se definen los campos que queremos en una variable de la siguiente manera:

Primero, se coge la url = "**https://newsapi.org/v2/everything**".

Después, se define lo siguiente:

```
params = {
    "q": topic,
    "language": language,
    "pageSize": page_size,
    "sortBy": "publishedAt",
}
```

También se define el **Header** con el que se hace la **request** usando la clave:

```
headers = {
    "Authorization": f"Bearer {NEWSAPI_KEY}"
}
```

Por último, se coge el **title**, **description** y **url** del artículo de la respuesta en JSON.

2. A través de una lista:

La única diferencia con la anterior es que se establecen dominios de diferentes páginas de noticias para consultas.

Tras esto definimos una función personalizada para determinar el tipo de artículo y así filtrarlos para que solo queden los de este tipo (**filter_by_article_type**).

Durante el proceso también definimos una función para pasarle los argumentos o parámetros del programa definiendo el tipo de noticia, que palabras clave debe tener, etc.

La librería Time sirve para ponerle un límite a la consulta de noticias dado que podemos sobrecargar las páginas, además de que pueden tomar medidas como bloquear la IP.

La librería Schedule sirve para ejecutar de forma periódica el programa a diario.

```
(venv) [luis@archluis Robotica]$ python main.py --source lista --lista "bbc.co.uk,washingtonpost.com,lemonde.fr,elpais.com,elmundo.es,europapress.es" --topic "Trump" AND "Musk" --language en --max-results 3 -o -t 19:45

Buscando noticias sobre "Trump" AND "Musk" en idioma 'en'...
Fuente: lista | Máx resultados: 3 | Tipo artículo: any
===== RESUMEN DE NOTICIAS =====
1. Título: Elon Musk changes his name to Kekius Maximus on X
   URL: https://www.bbc.co.uk/news/articles/cy53v2lqpx1o
   Resumen: Elon Musk changes his name to Kekius Maximus on X. The world's richest man sparks speculation after changing his name and using a picture of Pepe the Frog on his Twitter account. Musk: 'I'm not a billionaire. I'm just a guy who wants to be a better person. That's what I'm going to do. I don't want to change who I am. I just want to be more like Pepe the frog' He added: 'If you think I'm a bad

2. Título: Trump wants federal workers back in the office. It may be a tall task.
   URL: https://washingtonpost.com/politics/2024/12/26/trump-return-to-office-federal-workers-resistance/
   Resumen: President-elect Donald Trump warned federal employees last week that they must return to the office - or else "they're going to be dismissed." The threat was made in a letter to federal employees. Trump wants federal workers back in the office. He may have a tall task with the federal government in disarray after the November election. The president-elect is expected to be inaugurated on January 20. The letter was sent to federal workers last week. It was all so sent to employees in

3. Título: Trump threatens to try and regain control of Panama Canal
   URL: https://www.bbc.co.uk/news/articles/c9819w67jgo
   Resumen: Trump threatens to try and regain control of Panama Canal. It prompts a sharp rebuke from Panama's president, who says "every square metre" of the canal belongs to his country. Trump: 'I'm going to try to take back control of the Panama Canal, which is a very, very special place for me' He adds: "It's a very important place for the United States. It's a great place to have a business. And it's a wonderful place to live"

Iniciando el agente de noticias... Se ejecutará a diario a las 19:45.
Presiona Ctrl+C para detener.
```

Ilustración 1: Ejecución del programa por lista.

```
luis@archluis:~/Documentos/Robotica
(venv) [luis@archluis Robotica]$ python main.py -o -s newsapi -l es -q "Cambio climático" OR "Valencia" -m 8

Buscando noticias sobre "Cambio climático" OR "Valencia" en idioma 'es'...
Fuente: newsapi | Máx resultados: 8 | Tipo artículo: any
===== RESUMEN DE NOTICIAS =====
1. Título: Endrick se entrena con normalidad
   URL: https://www.marca.com/futbol/real-madrid/2025/01/01/endrick-entrena-normalidad.html
   Resumen: Endrick se entrena con normalidad. El brasileño se había retirado con molestias en la sesión de ayer. Leer: Endrick seEntrena with normalidad, el brasileño se ha entrenado con normalidades. Endrick: Endricks se enterna connormalidad, the brasileña se haEntrenado with normalidade, the Brasileño le entrenamos con normalizaci

2. Título: El Extra de Navidad de la ONCE reparte 800.000 euros entre Silla y Vilamarxant, municipios afectados por la dana
   URL: https://www.abc.es/gaspa/comunidad-valenciana/extra-navidad-once-reparte-800000-euros-silla-20250101170255-nt.html
   Resumen: El Extra de Navidad de la ONCE repartido 800.000 euros entre Silla y Vilamarxant, municipios afectados by the dana. Tras el sorteo celebrado en la mañana de this week, las localidades valencianas han recuperado mucho de sus premios mayores. The Extra of the ONCE establece una lista de premios de 500.000 euros per personas.

3. Título: Dónde se juega el Sudamericano Sub 20 2025
   URL: https://www.lanacion.com.ar/deportes/futbol/donde-se-juega-el-sudamericano-sub-20-2025-nid01012025/
   Resumen: El Sudamericano Sub 20 2025 se desarrollará en enero y febrero en la misma sede del Preolímpico clasificatorio a Paris 2024. Dónde se juega el Sudamericano Sub20 2025. El torneo se desarrollarán in enero and febrero en the same sede de Paris. El Sudamero Sub 20 2024 se jugará en Paris.

4. Título: Un fallo informático paraliza los trenes de alta velocidad entre Madrid y Asturias
   URL: https://www.lavozdeasturias.es/noticia/asturias/2025/01/01/fallo-informatico-paraliza-trenes-alta-velocidad-madrid-asturias/80031735744370574225133.htm
   Resumen: Un fallo informático paraliza los trenes de alta velocidad entre Madrid y Asturias. La compañía Talgo detalla que la incidencia ha sido debido a un fallo de comunicación entre el sistema de control y los cargadores de baterías oficiales de los Trenes. El fallo happened a las 20:30 en Madrid

5. Título: El árbitro del Valencia-Madrid, uno de los señalados por Real Madrid TV
   URL: https://www.mundodeportivo.com/futbol/real-madrid/20250101/1002379395/arbitro-valencia-madrid-senalados-real-madrid-tv.html
   Resumen: La RFEF ha hecho oficial la designación arbitral para el encuentro de Valencia y Real Madrid en Mestalla (21.00 horas, Movistar LaLiga) correspondiente a la jornada 12 que se aplazó por the DANA. El colegiado elegido por el C... El árbitro del Valencia-Madrid, uno of los señalados por Real Madrid TV.

6. Título: Investigan a una policía que prometió dar "una paliza" a quienes robaban en Valencia tras la dana
```

Ilustración 2: Ejecución con newsapi.

2. Conclusiones.

Esta propuesta, en comparación con otras, no tiene que tratar bibliotecas como **Selenium** para pasar sistemas **antibot** como captchas, ya que no se tratan de webs dinámicas.

Sin embargo, sí que se han añadido cosas que pueden ayudar a obtener noticias un poco más a **gusto del consumidor**, como, por ejemplo, el **idioma** en que se buscan o de que **tipo** queremos la noticia.

La ventaja principal, es que al utilizar la API **no estamos infringiendo términos de estas páginas**, pero no se profundiza mucho más al no utilizar estas otras librerías.

A diferencia de otras propuestas como la de **Steam**, que necesitaría **Selenium**, lo que quiere decir que se necesitaría saber que **navegador** se va a usar y que **driver** coger, también se tiene que añadir la **lógica** para tratar sistemas antibot, además que esto también **inflige sus normas**, lo cual puede derivar en bloqueos de IP o cosas más **graves**.

3. Ideas de futuro y posibles mejoras.

1. **Añadir notificación por Telegram o email.** El objetivo es que cada día te mande un resumen con las noticias como si fuera una Newsletter, pero con agentes de manera automática de medios variados.
2. **Hacer GUI de forma que sea más fácil y accesible el uso a cualquier usuario,** además de permitir la lectura en la propia aplicación donde puedas elegir cosas como fuente y tamaño de la letra.
3. Guardar **historico** y permitir **búsqueda posterior**.
4. Implementar un **clasificador** entrenado más preciso.

4. Referencias.

- [1] Lo visto en la asignatura: <https://github.com/etsisi/Robotica/tree/main/Slides>
- [2] Documentación de Newsapi: <https://newsapi.org/docs>
- [3] Modelo de resumen: <https://github.com/huggingface/transformers>
- [4] Librería de schedule: <https://pypi.org/project/schedule/>

5. Anexos.

Código del proyecto: [Repositorio GitHub con el código del proyecto.](#)