

YouTubeCrawlerTool:

Aplicación web para habilitar el estudio del movimiento antivacuna

Javier Sánchez Mendoza
Grado de ingeniería Informática
Health IT

Carlos Luis Sánchez Bocanegra
José Antonio Morán Moreno

Junio de 2018



Universitat
Oberta
de Catalunya

Estudios
de Informática,
Multimedia
y Telecomunicación

- Motivación
- Objetivos
- Planificación y proceso de desarrollo
- Metodología: fases y técnicas de desarrollo
 - Contexto de uso
 - Requisitos
 - Diseño del producto
 - Pruebas
- Demostración
- Mejoras
- Conclusiones

Índice

- El movimiento antivacuna se encuentra en expansión, sobre todo en Europa y Estados Unidos.
- Causante de la reaparición de enfermedades que se consideraban erradicadas.
- En el peor de los casos, ha causado victimas mortales.
- La desinformación es el principal motivo del auge del movimiento.

The screenshot shows a news article from the website buenavibra.es. The article is titled "Más de 5.000 casos de sarampión en Italia por los movimientos antivacunas" (More than 5,000 cases of measles in Italy due to anti-vaccination movements) and is categorized under "SALUD". It includes a sub-headline "SANIDAD No estaba vacunado" (Healthcare Not vaccinated) and a main headline "Muere el niño de 6 años de Olot infectado de difteria" (A 6-year-old boy from Olot dies from diphtheria). The article features an editorial titled "Asegurar la vacunación" (Ensuring vaccination) with the subtitle "El plan de UE para reforzar las campañas de inmunización y combatir la desinformación sobre las vacunas merece todo el apoyo" (The EU plan to reinforce vaccination campaigns and combat misinformation about vaccines deserves full support).

<https://buenavibra.es/movida-sana/salud/ italia-sarampion-movimientos-antivacunas> (16/03/2018)
<http://www.elmundo.es/cataluna/2015/06/27/558e5fb2e2704ea41e8b4576.html> (07/03/2018)
https://elpais.com/elpais/2018/04/26/opinion/1524767736_924183.html (09/06/2018)

Motivación

Desarrollar una aplicación web que permita la obtención de información en YouTube que posibilite el estudio del movimiento antivacuna.

- Investigar la API de YouTube y como se puede utilizar para la obtención de la información requerida.
- Determinar como almacenar y acceder de forma eficiente a la gran cantidad de información que se obtendrá.
- Permitir la recolección de información según criterios de búsqueda proporcionados por el usuario.
- Habilitar la gestión, visualización y exportación de los datos obtenidos para su posterior análisis en herramientas especializadas.
- Ofrecer herramientas de visualización para el análisis y comprensión de los datos obtenidos.
- Proporcionar una interfaz de usuario usable que permita realizar las acciones requeridas por el usuario final.

Objetivos

Planificación tentativa:

- **Definición de los contenidos del trabajo:** Redacción propuesta TFG.
 - **Plan de trabajo:** Redacción planificación.
 - **Desarrollo del trabajo – Fase 1:** Instalación en servidor de explotación con funcionalidad de crawler implementada.
 - **Desarrollo del trabajo – Fase 2:** Actualización en servidor de explotación de la versión final de la aplicación con funcionalidad analítica implementada.
 - **Redacción de la memoria:** Entrega de la memoria del proyecto.
 - **Elaboración de la presentación:** Realizar exposición oral.
 - **Defensa pública:** Defender públicamente el proyecto.

NOMBRE	FECHA DE INICIO	FECHA DE FIN	DURACIÓN
• TFG: YouTubeCrawlerTool	21/02/18	25/03/18	12 días
• Definición de los contenidos del trabajo	21/02/18	5/03/18	14 días
◦ Definir la temática del trabajo	21/02/18	2/03/18	14 días
◦ Redactar propuesta de TFG	3/03/18	5/03/18	14 días
• Plan de trabajo	6/03/18	19/03/18	14 días
◦ Definir metodología y planificación proyecto	6/03/18	16/03/18	14 días
◦ Redactar planificación	17/03/18	19/03/18	14 días
• Desarrollo del trabajo - Fase 1	20/03/18	23/04/18	34 días
◦ Analizar API de Twitter	20/03/18	21/03/18	14 días
◦ Analizar API de YouTube	22/03/18	23/03/18	14 días
◦ Prueba de concepto YouTube	24/03/18	27/03/18	14 días
◦ Plantilla memoria en LaTeX	28/03/18	28/03/18	14 días
◦ Analizar librerías gráficas	29/03/18	30/03/18	14 días
◦ Prueba de concepto visualización grafo	31/03/18	3/04/18	14 días
◦ Diseño de la interfaz de usuario	4/04/18	7/04/18	14 días
◦ Analizar uso de Neo4j como base de datos	8/04/18	9/04/18	14 días
◦ Prueba de concepto Neo4j	10/04/18	13/04/18	14 días
• Diseño de la arquitectura de la solución	14/04/18	20/04/18	14 días
◦ Modelo de datos	14/04/18	15/04/18	14 días
◦ Capa de acceso a datos	16/04/18	18/04/18	14 días
◦ Capa de presentación	19/04/18	20/04/18	14 días
◦ Instalación en servidor de explotación	21/04/18	23/04/18	14 días
◦ Redacción de la memoria	20/03/18	23/04/18	34 días
• Desarrollo del trabajo - Fase 2	24/04/18	21/05/18	28 días
• Implementación de la aplicación	24/04/18	20/05/18	28 días
◦ Modelo de datos	24/04/18	26/04/18	14 días
◦ Capa de acceso a datos	27/04/18	6/05/18	14 días
◦ Capa de presentación	7/05/18	20/05/18	14 días
◦ Definición y ejecución de pruebas	24/04/18	20/05/18	28 días
◦ Actualización servidor de explotación	21/05/18	21/05/18	14 días
◦ Redacción de la memoria	24/04/18	21/05/18	28 días
• Elaboración de la presentación	22/05/18	5/06/18	14 días
◦ Preparar material presentación	6/06/18	13/06/18	14 días
◦ Exposición oral	13/06/18	13/06/18	14 días
• Defensa pública	14/06/18	25/06/18	14 días

Marco ágil de desarrollo Scrum:

- **Scrum master:** Javier Sánchez Mendoza.
 - **Product owner:** Javier Sánchez Mendoza.
 - **Desarrollador:** Javier Sánchez Mendoza.
 - **Cliente:** Johanna Milena Rodríguez (analista de datos).
 - **Stakeholder:** Carlos Luis Sánchez Bocanegra.



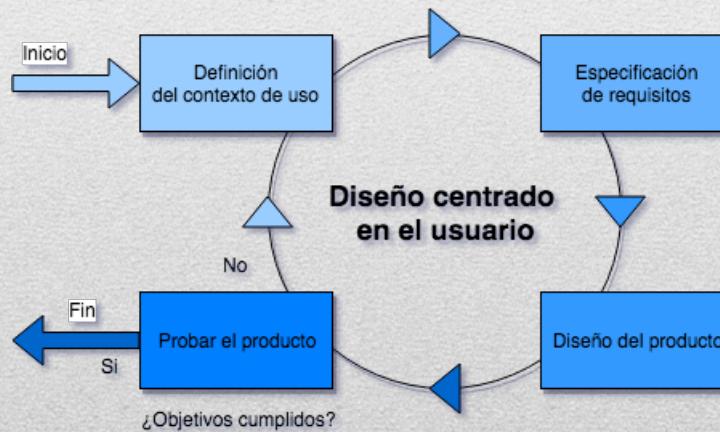
Backlog del proyecto

Planificación y proceso de desarrollo

Diseño centrado en el usuario:

Proceso iterativo por fases con el objetivo de proporcionar una aplicación que cumpla todas las expectativas de los usuarios y obtenga la máxima satisfacción posible.

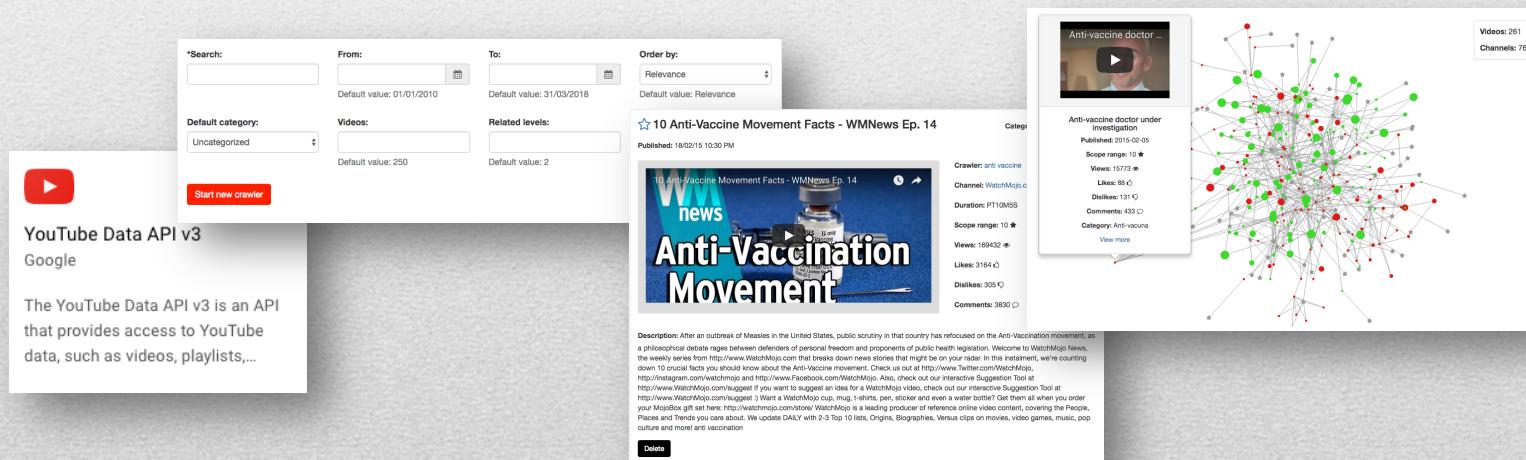
- **Definición del contexto de uso:** Entrevistas por videoconferencia con la usuaria final de la aplicación.
- **Especificación de requisitos:** Definición de casos de uso.
- **Diseño del producto:** Creación de maquetas y pruebas de concepto.
- **Probar el producto:** Pruebas de integración y demostraciones de la aplicación.



Metodología: fases y técnicas de desarrollo

Principales decisiones acordadas durante las entrevistas:

- Objetivos generales de la aplicación y alcance de la misma.
- Elección de YouTube como red social a utilizar en el estudio.
- Criterios de búsqueda e información a recolectar.
- Definición de visualización en formato de grafo para analizar de forma visual los datos obtenidos.
- Habilitar la categorización de contenidos mediante categorías definidas por el usuario.
- Dos perfiles de usuarios: usuarios identificados y anónimos.

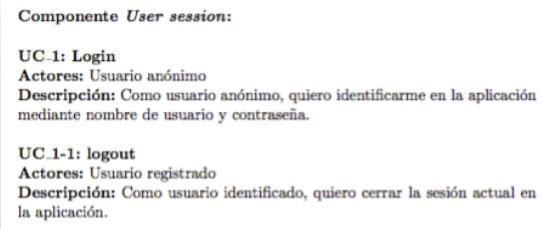


Contexto de uso

Toma de requerimientos de la aplicación como casos de uso agrupados por componentes según la funcionalidad realizada:

- **User session:** Gestión de sesión de usuario.
- **Crawler:** Procesos de recolección, acceso y gestión de los mismos.
- **Video:** Acceso y gestión a videos obtenidos mediante procesos de recolección.
- **Channel:** Acceso y gestión a canales de publicación obtenidos mediante procesos de recolección.
- **Category:** Acceso y gestión de categorías.
- **Analysis:** Exportación de contenidos y visualización en grafo

Requisitos



Casos de uso componente User session

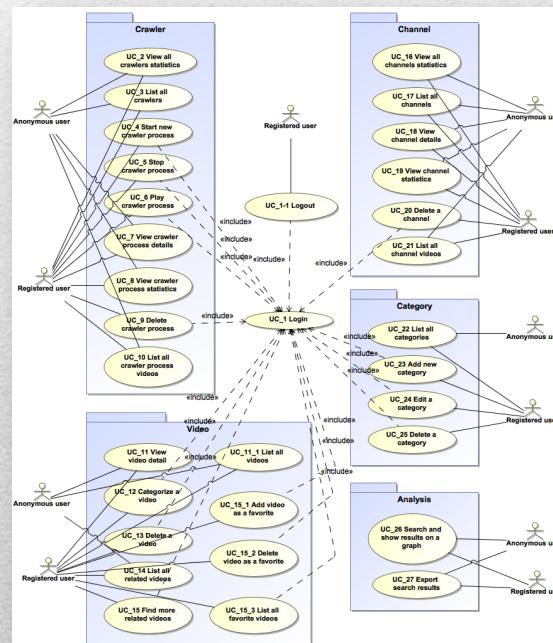


Diagrama de casos de uso

Propuestas de interfaz de usuario:

The diagram illustrates the progression of the YouTubeCrawlerTool's user interface design:

- Initial Phase:** A screenshot of the "TwitterCrawler" interface, which includes a search bar for hashtags and a table of crawled tweets.
- Intermediate Phase:** A screenshot of the "YouTubeCrawlerTool" interface, featuring a search bar for terms and a table of crawled video results.
- Advanced Phase:** A large screenshot showing the full suite of tools:
 - Login Screen:** A red-themed login page with fields for Username and Password, and a "Login" button.
 - Crawler Dashboard:** Displays "Crawler stats" with a total of 25,700 videos, categorized as 25%. It includes a search bar for "antivaccine", a date range selector from 01/01/2010 to 01/04/2018, and a "Start new crawler" button.
 - Video Detail View:** Shows a video entry for "VG423JHsw" with details like Title: Lorem ipsum, Duration: 3:10 minutes, and Crawler: sodoses.
 - Analysis Search:** A search interface for analyzing video data across categories (Antivaccine, Provacine, Repeated) and channels (sodoses, magnos).
 - Network Analysis:** A complex network graph visualization showing connections between various biological or chemical entities like APEX2, RAD50, FEN1, PCNA, MSH6, LIG1, GADD45G, POLD1, POLD3, RFC1, RFC2, RFC3, RFC4, RPL12A, UNG, and DNAC.

Diseño del producto

Pruebas de concepto:

- **POCTwitterCrawler:** Integración con la API de *Twitter* y persistencia de contenidos recolectados en *MongoDB*
- **POCYoutubeCrawler:** Integración con la API de *YouTube* y persistencia de contenidos recolectados en *MongoDB*. Posteriormente también se añadió visualización en grafo.
- **POCYoutubeCrawlerNeo4j:** Persistencia de contenidos recolectados en *Neo4j*.

```

POCYoutubeCrawler
  Collections (4)
    > cafe
    > channels
    > edges
    > nodes
    > Functions
    > Users

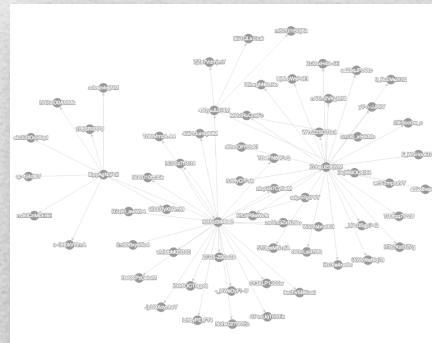
db.getCollection("cafe").find({})

Key          Value
-----      -----
ObjectID("5acc7ae08813b0073485caf") { 13 fields }
ObjectID("5acc7ae08813b0073485caf")
videoId      ZccHqg5$eo
title        Relaxing Jazz & Bossa Nova Music Radio - 24/7 CH...
description  Please Subscribe Cafe Music BGM channel https://...
publishedAt 2017-11-04T11:37:09Z
channelId   UC.JnE7wbvYaaeG25m0tHAA
someRange    172247016050
duration    7701
viewCount   11700825
likeCount   86777
dislikeCount 3429
commentCount 1
contributors  []
contributors  []

(2 ObjectID("5acc7ae08813b0073485cb7")) { 13 fields }
(3 ObjectID("5acc7ae08813b0073485cb7"))
(4 ObjectID("5acc7ae08813b0073485cb7"))
(5 ObjectID("5acc7ae08813b0073485cb7"))
(6 ObjectID("5acc7ae08813b0073485cb7"))
(7 ObjectID("5acc7ae08813b0073485cb7"))
(8 ObjectID("5acc7ae08813b0073485cb7"))
(9 ObjectID("5acc7ae08813b0073485cb7"))
(10 ObjectID("5acc7ae08813b0073485cb7"))
(11 ObjectID("5acc7ae08813b0073485cb7"))

```

POCYoutubeCrawler: consola MongoDB



POCYoutubeCrawlerNeo4j: visualización grafo

Diseño del producto

Diseño de la aplicación por capas:

- Capa de datos
- Capa de aplicación (acceso a datos, lógica de negocio y presentación)
- Capa cliente

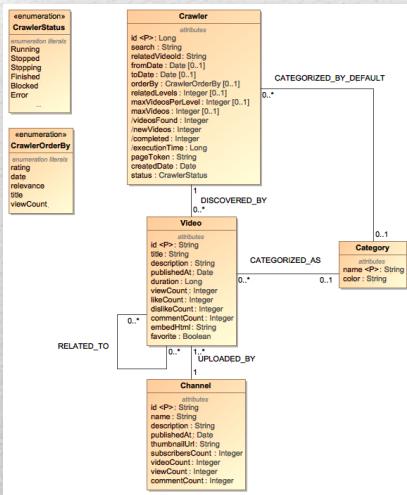


Diagrama de clases UML capa de datos

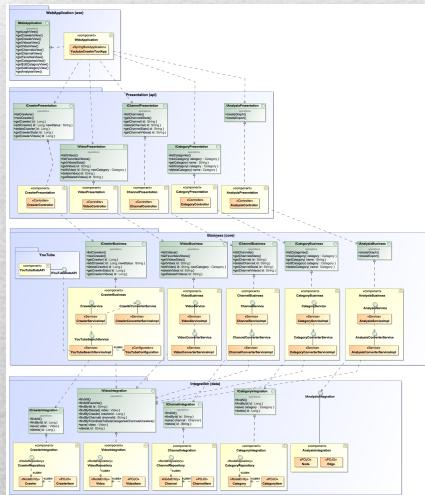
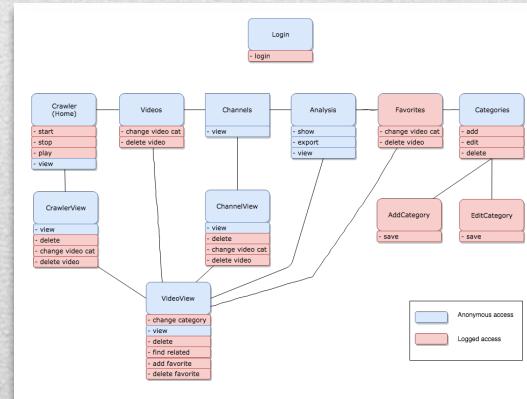


Diagrama de componentes capa aplicación

Componente crawler:

Método	Uri	Descripción
GET	api/crawlers	Lista todos los procesos de recolección.
POST	api/crawlers	Inicia un nuevo proceso de recolección. Requiere autenticación.
GET	api/crawlers/{id}	Devuelve un proceso de recolección.
PUT	api/crawlers/{id}	Editar un proceso de recolección. Requiere identificación.
DELETE	api/crawlers/{id}	Borra un proceso de recolección. Requiere identificación.
GET	api/crawlers/{id}/stats	Devuelve las estadísticas de un proceso de recolección.
GET	api/crawlers/{id}/videos	Lista todos los videos descubiertos por un proceso de recolección.

Diseño API REST componente crawler



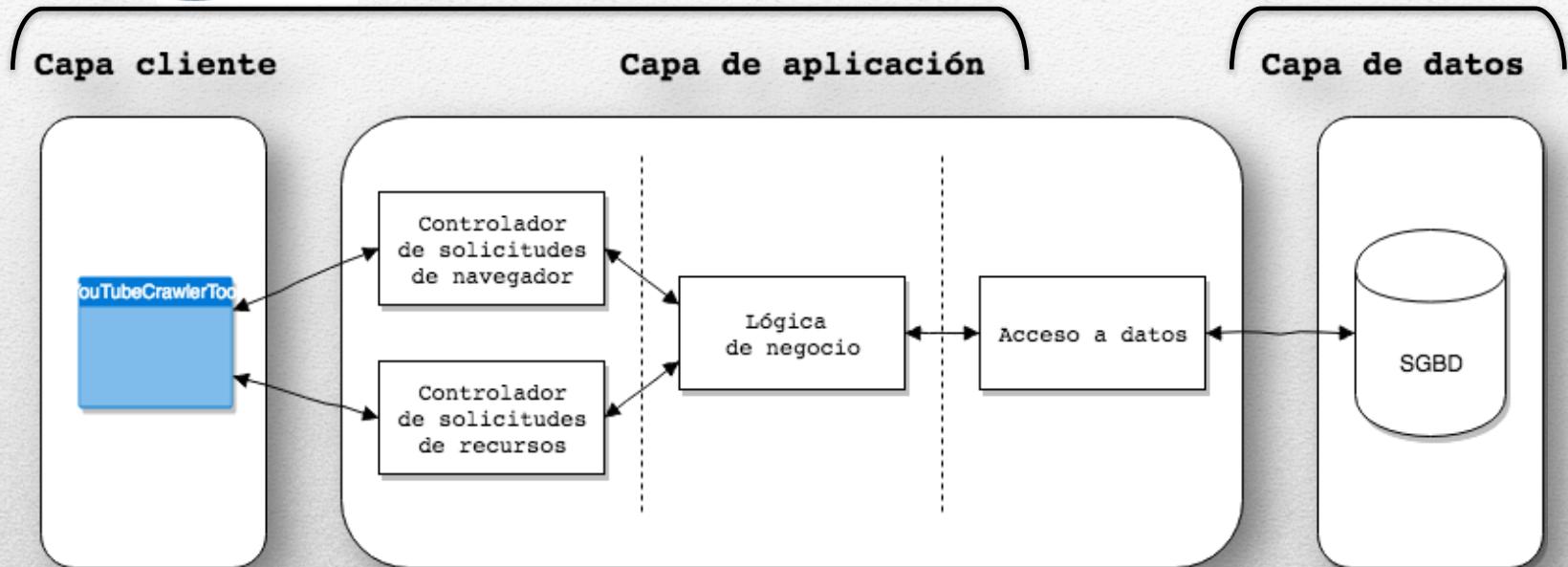
Diseño vistas capa cliente

Diseño del producto

Arquitectura y tecnologías:



Maven™



Spring MVC
Spring Security



Spring Data Neo4j

Diseño del producto

Pruebas de integración:

- Definidas durante el proceso de desarrollo.
- Realizadas a partir de los casos de uso.
- Permitieron evaluar la calidad de los componentes por separado tan pronto como eran desarrollados.

Responsable	Javier Sánchez Mendoza		
Componente	User session		
Caso de uso	UC_1: Login		
Código	Acciones a verificar	Resultado esperado	Verificación
1	Identificarse al sistema como usuario válido.	Usuario identificado.	V
2	Identificarse al sistema ya estando identificado.	Acción no permitida por la aplicación.	V
3	Introducir datos nuevos o incorrectos.	Usuario no identificado.	V
EVALUACIÓN 100 %		V X	3 0

Pruebas de integración para el caso de uso UC_1: Login

Demostraciones de la aplicación:

- Realizadas durante las ultimas etapas de desarrollo
- Permitieron obtener la aprobación de la clienta y ayudaron a su formación.

<https://youtu.be/ETWtbywxOml>

Pruebas

The screenshot shows a web application window titled "YouTubeCrawlerTool" running in a QuickTime Player browser on a Mac OS X system. The window title bar includes the application name, file menu, and system status indicators. The main content area has a red header bar with the title "YouTubeCrawlerTool". Below the header is a navigation bar with links for "Categories" and "Login". On the left, there is a sidebar with icons for "Crawlers" (selected), "Videos", "Channels", and "Analysis". The main content area is titled "Crawlers" and displays a table of crawler processes. The table has columns: Search, Videos found, New videos, Completed, Created, Execution time, and Status. It lists three entries:

Search	Videos found	New videos	Completed	Created	Execution time	Status
anti vaccine	775	236	100%	27/05/18 4:19 PM	4:39 mins	✓ Finished View
pro vaccine	775	293	100%	27/05/18 4:19 PM	4:35 mins	✓ Finished View
vaccine	1696	683	28%	09/06/18 10:31 PM	5:54 mins	■ Stopping View

Below the table, it says "Showing 1 to 3 of 3 entries" and has navigation buttons for "Previous", "1", and "Next".

Demostración

Cumplimiento de objetivos:

- Objetivo principal del proyecto realizado con éxito.
- Todos los objetivos específicos se han cumplido.
- La aplicación cumple con las expectativas de la clienta permitiendo iniciar el estudio del movimiento antivacuna en YouTube.

Propuesta de mejoras

- Mejoras en accesibilidad y usabilidad.
- Uso de inteligencia artificial para la categorización automática de vídeos (por ejemplo, modelo basado en reglas).

Opinión personal

- Oportunidad de poner a prueba conocimientos adquiridos durante el grado en ingeniería informática.
- Buena experiencia de trabajo colaborativo junto con la clienta final de la aplicación.
- Obtención de visión global en el desarrollo de software al desarrollar una aplicación de principio a fin.
- Adquisición de conocimiento en el uso de nuevas tecnologías.

Conclusiones

Gracias!

Código fuente:

<https://github.com/jsanchezmend/TFGAntivacunas>

Preguntas?

jsanchezmend@uoc.edu



Universitat
Oberta
de Catalunya

Estudios
de Informática,
Multimedia
y Telecomunicación