TFG - Health IT: YouTubeCrawlerTool

PEC 2: Informe de seguimiento

Año académico 2017/18-2

Javier Sánchez Mendoza

22-4-2018

TFG - Health IT: YouTubeCrawlerTool : PEC 2 – Informe de seguimiento

1. Avance del proyecto

1.1. Grado de cumplimiento de los objetivos

A continuación se facilita la relación de objetivos definidos para el proyecto y su grado de cumplimiento durante la primera fase de desarrollo:

- *Investigar que funcionalidades aportan las API públicas ofrecidas por YouTube y analizar como se pueden utilizar para la obtención de la información requerida.*
    - Se ha analizado la API de YouTube y se ha estudiado conjuntamente con el cliente que información va a ser recolectada por la aplicación. Este objetivo se considerara realizado.

- *Determinar como almacenar y acceder de forma eficiente a la gran cantidad de información que se obtendrá.*
    - Se ha tomado la decisión de utilizar una base de datos NoSQL enfocada a grafos (Neo4j) para poder acceder a la información de forma eficiente. Se ha diseñado un diagrama UML para representar el esquema invariante de la información y se ha adaptado al concepto de nodos y relaciones de las base de datos de grafos. Finalmente, se ha realizado una prueba de concepto para estudiar la viabilidad de utilizar Neo4j como base de datos en el proyecto con resultado favorable. Este objetivo se considerara realizado.

- *Permitir la recolección de información según criterios de búsqueda proporcionados por el usuario final.*
    - Conjuntamente con el cliente se han determinado los criterios de búsqueda a utilizar. Se ha desarrollado una prueba de concepto en la cual a sido posible recolectar información de videos y canales utilizando la API de YouTube. Este objetivo, a falta de su implementación en la aplicación final, se considera parcialmente realizado.

- *Habilitar la gestión, visualización y exportación de datos obtenidos en distintos procesos de extracción para su posterior análisis en herramientas especializadas.*
    - Se ha diseñado una interfaz de usuario para proporcionar estas funcionalidades y se ha diseñado la arquitectura de la aplicación para hacerlas viables. Este objetivo, a falta de su implementación en la aplicación final, se encuentra en su fase inicial.
- *Ofrecer herramientas de visualización para el análisis y comprensión de los datos obtenidos.*
    - Se ha decidido el uso de un grafo para la visualización de los datos obtenidos y se ha determinado cuales serán sus componentes. Se ha realizado una prueba de concepto en la cual ha sido posible visualizar los datos obtenidos a través de la API de YouTube en un grafo. Queda pendiente acabar de determinar como se visualizara el grafo. Este objetivo esta en su fase inicial.

- *Proporcionar una interfaz de usuario usable que permita realizar las acciones requeridas por el usuario final.*
    - Se ha realizado una propuesta de interfaz de usuario que ha sido aprobada por la cliente. A falta de su implementación en la aplicación, este objetivo se encuentra un su fase inicial.

1.2. Justificación de cambios

Durante la realización de la primera fase de desarrollo se realizaron cambios que afectaron al alcance del proyecto o su implementación, la relación de cambios es la siguiente:

- **Cambio de red social:** Debido a las restricciones de uso de la API gratuita de Twitter (https://developer.twitter.com/en/docs/tweets/search/overview), la cual solo permite obtener información con una antigüedad de 7 días haciendo inviable el estudio de la evolución del movimiento antivacuna en el tiempo, se ha decidido utilizar en substitución la red social YouTube. Hace tiempo que YouTube se ha convertido en un referente como red social donde sus usuarios suben videos sobre una gran cantidad de temáticas (incluyendo videos pro y anti vacunas) y la API publica de YouTube aunque sigue teniendo limitaciones de uso, es mucho más abierta y nos permite obtener información histórica (https://developers.google.com/youtube/v3/getting-started).

- **Cambio de base de datos:** Inicialmente se había planeado utilizar la base de datos documental NoSQL MongoDB (https://www.mongodb.com/) para persistir los datos obtenidos y realizar consultas mediante aplicación web, pero con la importancia que a cobrado especialmente en el proyecto la visualización de los datos obtenidos en forma de grafo y el descubrimiento de bases de datos de grafos NoSQL como Neo4j (https://neo4j.com/product/), se ha decidido utilizar este ultimo sistema gestor de bases de datos para desarrollar el proyecto. Entre las ventajas que aporta Neo4j al proyecto, se encuentra el modelaje de la base de datos como grafo y mayores capacidades analíticas para grafos.

2. Relación de actividades realizadas

2.1. Actividades previstas en el plan de trabajo

De las actividades previstas en el plan de trabajo para la primera fase de desarrollo se llevaron a cabo las siguientes:

- Analizar API de Twitter
- Diseño de la interfaz de usuario
- Diseño de la arquitectura de la solución
    - Modelo de datos
    - Capa de acceso a datos
    - Capa de presentación
- Instalación en servidor de explotación
- Redacción de la memoria

2.2. Actividades no previstas y realizadas o programadas

Como era de esperar, durante la realización de la primera fase de desarrollo se detectaron o surgieron actividades no previstas inicialmente en la planificación del proyecto, el surgimiento de estas nuevas actividades se debieron principalmente al cambio de red social a utilizar (de Twitter a YouTube) y al cambio de arquitectura de base de datos (de una base de datos documental como a una basada en grafos).

La relación de las nuevas actividades que se han realizado en la actual fase de desarrollo ha sido la siguiente:

- Estudiar el uso de LaTeX para la redacción de la memoria.
- Analizar API de YouTube (e investigar funcionalidades añadidas como el uso de hashtags, detección de video repetidos o detección de influencers).
- Analizar librerías para la visualización de gráficas (se descarto el uso de librerías basadas en Java a favor del uso de librerías JS).
- Analizar los pros y contras en el uso de Neo4j (base de datos de grafos) en substitución a MongoDB (base de datos documental).
- Desarrollo de pruebas de concepto para analizar la viabilidad de la solución
    - Crawler de YouTube
    - Visualización de grafo con librería JS
    - Uso de Neo4j en la capa de persistencia.

3. Relación de desviaciones en la temporización, acciones de mitigación y actualización del cronograma

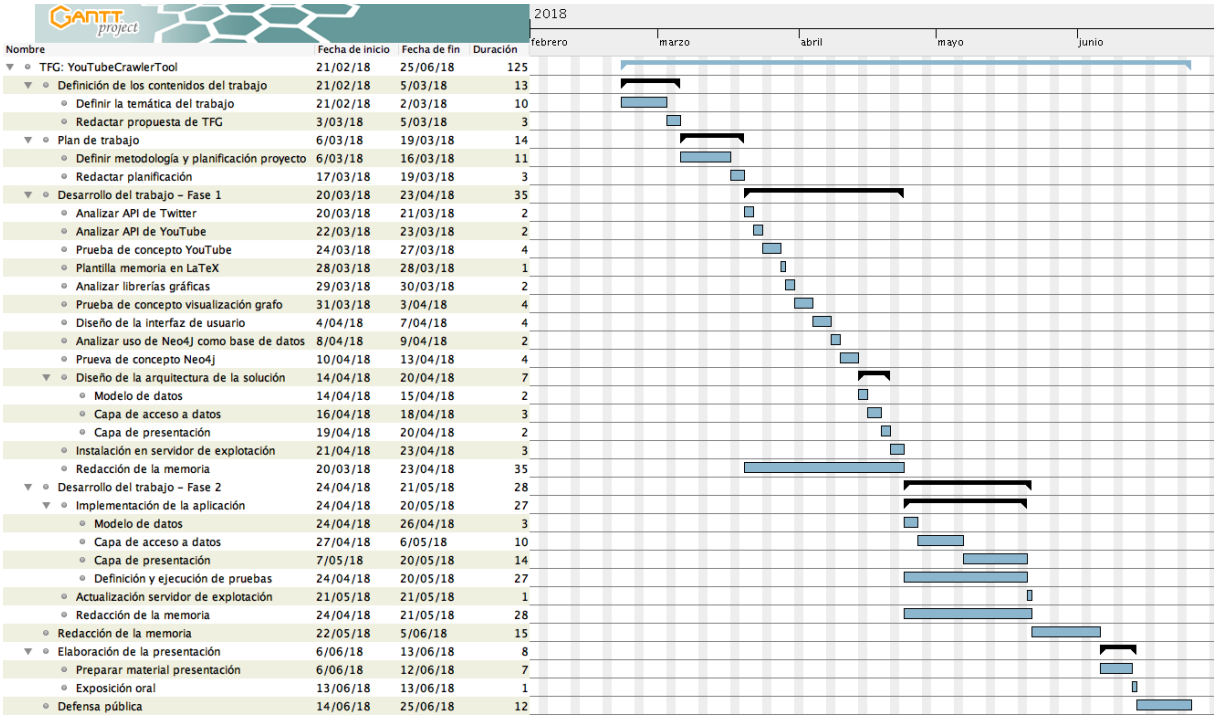
Debido a las actividades no previstas detalladas en el punto anterior, de la planificación inicial no se pudieron ejecutar las actividades relacionadas con la primera fase de implementación de la aplicación correspondientes a la funcionalidad del crawler:

Implementación del crawler	2/04/18	20/04/18	19
Modelo de datos	2/04/18	4/04/18	3
Capa de acceso a datos	5/04/18	13/04/18	9
Capa de presentación	14/04/18	20/04/18	7
Definición y ejecución de pruebas	2/04/18	20/04/18	19

Aunque es una desviación importante de un total de 19 días, gracias al detallado diseño de la aplicación, el exhaustivo análisis de la API de YouTube y, sobretodo, a las diferentes pruebas de concepto realizadas, se cree posible poder realizar estas tareas dentro de la fase 2 de desarrollo al estimarse que el tiempo necesario tanto para la implementación del crawler como de la herramienta analítica se ven reducidos gracias a las acciones comentadas anteriormente.

Aun así y con el propósito de reducir riesgos, se ha decidido priorizar algunas de las funcionalidades de la aplicación por tal de asegurar su implementación. En concreto y después de acordarlo con el cliente, se ha decidido priorizar el proceso de crawler y la visualización de los datos obtenidos en un grafo. Otras funcionalidades como la visualización de estadísticas o las vistas de canales entre otras tendrán una prioridad secundaria, al igual que la funcionalidad de exportación a fichero externo debido a que esta función se puede realizar directamente realizando consultas en el motor de base de datos Neo4j y por lo tanto no supone un riesgo para el cliente. De todas formas el objetivo del proyecto sigue siendo implementar la aplicación propuesta con todas las funcionalidades que han sido diseñadas y acordadas en la fase anterior.

De este modo, la planificación inicial para la segunda fase de desarrollo se ve afectada tal y como se recoge en este nuevo cronograma ya adaptado a la nueva planificación y en donde se han actualizado también las tareas realizadas en la primera fase:



Nombre	Fecha de inicio	Fecha de fin	Duración
TFG: YouTubeCrawlerTool	21/02/18	25/06/18	125
Definición de los contenidos del trabajo	21/02/18	5/03/18	13
Definir la temática del trabajo	21/02/18	2/03/18	10
Redactar propuesta de TFG	3/03/18	5/03/18	3
Plan de trabajo	6/03/18	19/03/18	14
Definir metodología y planificación proyecto	6/03/18	16/03/18	11
Redactar planificación	17/03/18	19/03/18	3
Desarrollo del trabajo – Fase 1	20/03/18	23/04/18	35
Analizar API de Twitter	20/03/18	21/03/18	2
Analizar API de YouTube	22/03/18	23/03/18	2
Prueba de concepto YouTube	24/03/18	27/03/18	4
Plantilla memoria en LaTeX	28/03/18	28/03/18	1
Analizar librerías gráficas	29/03/18	30/03/18	2
Prueba de concepto visualización grafo	31/03/18	3/04/18	4
Diseño de la interfaz de usuario	4/04/18	7/04/18	4
Analizar uso de Neo4J como base de datos	8/04/18	9/04/18	2
Prueva de concepto Neo4j	10/04/18	13/04/18	4
Diseño de la arquitectura de la solución	14/04/18	20/04/18	7
Modelo de datos	14/04/18	15/04/18	2
Capa de acceso a datos	16/04/18	18/04/18	3
Capa de presentación	19/04/18	20/04/18	2
Instalación en servidor de explotación	21/04/18	23/04/18	3
Redacción de la memoria	20/03/18	23/04/18	35
Desarrollo del trabajo – Fase 2	24/04/18	21/05/18	28
Implementación de la aplicación	24/04/18	20/05/18	27
Modelo de datos	24/04/18	26/04/18	3
Capa de acceso a datos	27/04/18	6/05/18	10
Capa de presentación	7/05/18	20/05/18	14
Definición y ejecución de pruebas	24/04/18	20/05/18	27
Actualización servidor de explotación	21/05/18	21/05/18	1
Redacción de la memoria	24/04/18	21/05/18	28
Redacción de la memoria	22/05/18	5/06/18	15
Elaboración de la presentación	6/06/18	13/06/18	8
Preparar material presentación	6/06/18	12/06/18	7
Exposición oral	13/06/18	13/06/18	1
Defensa pública	14/06/18	25/06/18	12

Figure 1 Listado de tareas y diagrama de Gantt

Nombre	Fecha de inicio	Fecha de fin	Duración
TFG: YouTubeCrawlerTool	21/02/18	25/06/18	125
Definición de los contenidos del trabajo	21/02/18	5/03/18	13
Definir la temática del trabajo	21/02/18	2/03/18	10
Redactar propuesta de TFG	3/03/18	5/03/18	3
Plan de trabajo	6/03/18	19/03/18	14
Definir metodología y planificación proyecto	6/03/18	16/03/18	11
Redactar planificación	17/03/18	19/03/18	3
Desarrollo del trabajo – Fase 1	20/03/18	23/04/18	35
Analizar API de Twitter	20/03/18	21/03/18	2
Analizar API de YouTube	22/03/18	23/03/18	2
Prueba de concepto YouTube	24/03/18	27/03/18	4
Plantilla memoria en LaTeX	28/03/18	28/03/18	1
Analizar librerías gráficas	29/03/18	30/03/18	2
Prueba de concepto visualización grafo	31/03/18	3/04/18	4
Diseño de la interfaz de usuario	4/04/18	7/04/18	4
Analizar uso de Neo4J como base de datos	8/04/18	9/04/18	2
Prueva de concepto Neo4j	10/04/18	13/04/18	4
Diseño de la arquitectura de la solución	14/04/18	20/04/18	7
Modelo de datos	14/04/18	15/04/18	2
Capa de acceso a datos	16/04/18	18/04/18	3
Capa de presentación	19/04/18	20/04/18	2
Instalación en servidor de explotación	21/04/18	23/04/18	3
Redacción de la memoria	20/03/18	23/04/18	35
Desarrollo del trabajo – Fase 2	24/04/18	21/05/18	28
Implementación de la aplicación	24/04/18	20/05/18	27
Modelo de datos	24/04/18	26/04/18	3
Capa de acceso a datos	27/04/18	6/05/18	10
Capa de presentación	7/05/18	20/05/18	14
Definición y ejecución de pruebas	24/04/18	20/05/18	27
Actualización servidor de explotación	21/05/18	21/05/18	1
Redacción de la memoria	24/04/18	21/05/18	28
Redacción de la memoria	22/05/18	5/06/18	15
Elaboración de la presentación	6/06/18	13/06/18	8
Preparar material presentación	6/06/18	12/06/18	7
Exposición oral	13/06/18	13/06/18	1
Defensa pública	14/06/18	25/06/18	12

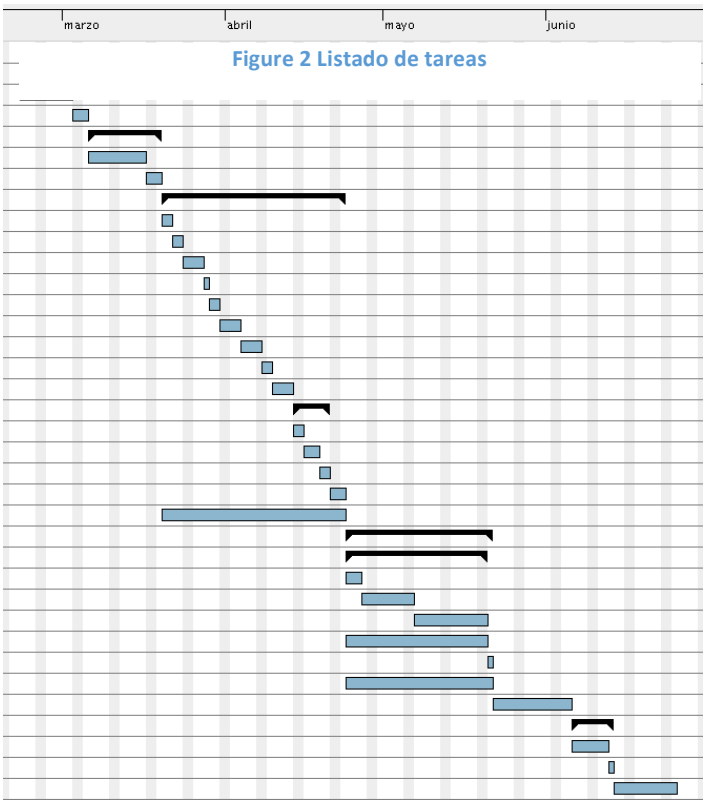


Figure 2 Listado de tareas

Figure 3 Diagrama de Gantt

4. Listado de resultados parciales obtenidos hasta el momento

A continuación se detalla el listado de los resultados parciales obtenidos durante la primera fase de desarrollo del trabajo:

- Pruebas de concepto desarrolladas:
  - Prueba de concepto YouTube + visualización grafo:
    https://github.com/jsanchezmend/TFGAntivacunas/tree/master/POCYouTubeCrawler
  - Prueba de concepto YouTube + visualización grafo + Neo4j:
    https://github.com/jsanchezmend/TFGAntivacunas/tree/master/POCYouTubeCrawlerNeo4j

- Diseño de la interfaz de usuario:
  https://github.com/jsanchezmend/TFGAntivacunas/tree/master/Mockups/15:04

- Diseño de la arquitectura de la solución:
  https://github.com/jsanchezmend/TFGAntivacunas/tree/master/Diseño

- Instalación en servidor de explotación:
  - Aplicación web (con POCYouTubeCrawlerNeo4j instalado):
    Url: http://youtubecrawlertoolwebapp.azurewebsites.net/
  - Servidor base de datos Neo4j:
    Url: http://51.136.48.142:7474/browser/
    Usuario: neo4j
    Password: Y01t1b3cr4wl3rt00l
    Consulta de ejemplo: MATCH (n) RETURN n

- Memoria del trabajo (inacabada actualizada parcialmente con los avances realizados en la primera fase de desarrollo del proyecto):
  https://github.com/jsanchezmend/TFGAntivacunas/blob/master/Memoria/TFG_JavierSanchezMendoza.pdf