UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

Escuela Superior de Informática

(TITLE)

Autor: María González Gutiérrez

Asignatura:

Grado en Ingeniería Informática

(YEAR), (MONTH)

Índice

El trabajo recogería los siguientes apartados:

- Introducción (muy recomendable aunque no obligatorio)
- Tecnología específica cursada por el alumno
- Objetivos
- IRÁN TANTAS SECCIONES CENTRALES COMO EL ALUMNO CONSIDERE
- Medios que se pretenden utilizar
- Bibliografía básica consultada en la elaboración del anteproyecto
- Contrato de propiedad intelectual (si lo hubiera)

1. INTRODUCCIÓN Y OBJETIVOS

El capítulo de introducción podrá abordar los siguientes aspectos:

- Introducción al tema, entorno en el que el trabajo desempeñará su objetivo, justificación de la importancia del trabajo abordado.
- Motivación y antecedentes (con algunas referencias bibliográficas).
- Descripción gráfica del proyecto (es aconsejable incorporar una figura que describa el trabajo a desarrollar y que mejore la comprensión del mismo).

De acuerdo a la Introducción, el alumno deberá especificar cuál o cuáles son las hipótesis de trabajo de las que se parten, qué se pretende resolver, y en base a eso formular el objetivo principal del TFG.

El objetivo principal deberá desglosarse en sub-objetivos parciales. Los sub-objetivos deberán describirse de forma breve y concisa.

Como preámbulo a la formulación del objetivo parcial, el alumno deberá discutir sobre las limitaciones y condicionantes a tener en cuenta en el desarrollo del TFG (lenguaje de desarrollo, equipos, madurez de la tecnología, etc.).

Del mismo modo, será recomendable incluir una lista preliminar de requisitos del sistema a construir.

Este trabajo fin de grado consiste en un sistema de análisis de texto. El programa procesa un conjunto de textos y extrae una ontología con los personajes, sucesos y relaciones relevantes entre ellos. Los textos escogidos son del género conocido como *fanfiction*, que se caracteriza por no ser originales: son textos de ficción basados en historias ya existentes, normalmente escritos for fans de la obra original que quieren explorar su mundo y personajes. Para este

trabajo se ha escogido un conjunto de *fanfics* basados en *Good Omens* (1990, Terry Prattchet y Neil Gaiman), de modo que se garantiza que tengan personajes, eventos, temas y tramas en común. El sitio web del que se han extraído los textos es Archive Of Our Own (AO3 de ahora en adelante), puesto que es el mayor archivo virtual de *fanfics* que existen en internet y permiten la descarga gratuita de todas las obras publicadas en él.

cantidad

AO3 es un sitio web que reúne a fans de

El objetivo del proyecto es conseguir una ontología que

2. FANFICTION, WEB ARCHIVE OF OUR OWN Y RELATOS A UTILIZAR

Un fanfic (abreviatura de "fanfiction", "ficción del fan") es una historia basada en una historia ya existente. Son, en esencia, historias creadas sin ánimo de lucro por los fans de un libro, película o videojuego. Crear fanfiction, en general, consiste en explorar temas e ideas que uno siente que faltan en la historia original. Puesto que cada fan tiene una interpretación distinta de los personajes y el mensaje que la historia original transmite, el fan convertido en autor puede añadir o quitar a la historia original lo que considere oportuno para contar su propia visión. Esto significa que cada fanfic es efectivamente una "transformación" de la historia original. El fanfic cae en una zona gris en términos de derechos de autor, pero suele considerarse "fair use". A día de hoy, la mayoría de fanfics se publican en sitios web de toda índole, destacando entre ellos Archive Of Our Own, que es un archivo open-source y sin ánimo de lucro creado expresamente para alojar obras creadas por fans. Según sus datos de mayo de 2020, tiene más de dos millones de usuarios registrados y más de seis millones de trabajos alojados. En particular elegí descargar los fanfics basados en *Good Omens*, un libro de Terry Pratchett y Neil Gaiman, tanto por la cantidad de relatos existente como por mi familiaridad con esa comunidad.

Además de la gran cantidad de relatos que aloja, los motivos por el que elegí extraer los datos de Archive Of Our Own son su herramienta de búsqueda y filtrado, su sistema de etiquetas y el hecho de que permite descargar el archivo en HTML. Archive Of Our Own permite filtrar fanfics según varios parámetros y genera un enlace a ese subconjunto de relatos particular, muy útil para descargar una gran cantidad de datos.

3. RECOGIDA DE FANFICS USANDO LA BASE DE DATOS DE ARCHIVE OF OUR OWN

En el momento en el que decidí utilizar los fanfics de *Good Omens* para el proyecto, dicho libro tenía unos 23000 fanfics en Archive Of Our Own (AO3 para abreviar). Sin embargo, de todos esos relatos sólo me interesaban los que están en inglés y los que realmente contuvieran texto (puesto que, aunque AO3 se centra en texto, permite todo tipo de archivos multimedia), de modo que utilicé el sistema de filtrado del sitio para seleccionar sólo los fanfics en inglés y que no estuvieran etiquetadas como "fanart", "podfic", etc, ya que indican que la obra no tiene texto. Tras este filtrado, quedaron unos 21000 relatos.

Ese filtrado genera un enlace que lleva siempre a ese subconjunto de textos, al cual puedo enviar peticiones HTTP GET e ir descargando las páginas, para lo cual creé un *scraper*. Cada página contiene un máximo de 20 fanfics, con lo que el *scraper* primero descarga la página y revisa cada fanfic en ella antes de descargar la siguiente. En este punto llevo a cabo un segundo filtrado, cuyo objetivo es descartar todos fanfics que tengan menos de 40 palabras por página, para evitar las obras sin texto que no hayan sido etiquetadas como tal por su autor. El programa entonces recoge los *links* de cada fanfic individual y los guarda en un archivo *txt*. Decidí guardar estos *links* en vez de descargar los fanfics directamente porque hay más de 20000, y pensé que sería más simple si ejecuto una vez el *scraper* que recorre las páginas, selecciona los fanfics y guarda su enlace en un *txt* que descargarlos todos de golpe. Una vez tuve una lista con todos los *links* que me interesaban, hice un segundo *scraper* que simplemente recorre los *links* de dicha lista y los descarga. hecho de tal manera que se le pueda indicar qué tramo de la lista tiene que descargar (por ejemplo, del número 5000 al 6000). De esta manera pude descargar los 20000 a lo largo de varios días. El principal problema una vez resuelto el código de la descarga en sí es cuando la página me echaba por enviar demasiadas peticiones (error 429: Too Many Requests), para lo cual simplemente utilicé un try-catch para detectar el 429 y dejar el programa durmiendo durante 2 minutos antes de reanudar su función por donde la había dejado. El otro problema es que algunos autores borraron la obra de archiveofourown.org antes de que la descargase, con lo que el *link* daba un error 404. El programa lo detecta en el try-catch y simplemente pasa al siguiente. Este scraper utiliza la librería bs4 y BeautifulSoup para descargar y manejar los archivos HTTP.

4. RECOMENDACIONES

Aquí habría que insertar tantas secciones como el alumno considere.

A continuación se indican una serie de recomendaciones que seguidas mejorarán la calificación final del trabajo.

- Agrupar párrafos. La división del texto en múltiples párrafos independientes de pequeña longitud dificulta la lectura continua del documento.
- Evitar abusar de las listas con viñetas y las enumeraciones.
- Utilizar un máximo nivel de profundidad secciones de 3 (hasta 3.1.1). Si se hace necesario una división más baja, no hacerlo con enumeración de subsecciones sino con texto en negrita y/o subrayada que represente el comienzo de cada subsección.
- Si no hay más de una sección en un nivel no crearla. Es decir, si no hay al menos una 3.2 no crear la 3.1, ya que no tendría sentido dividir la sección 3.
- Utilizar referencias absolutas numéricas a tablas, figuras, secciones, etc. Por ejemplo: en la "Figura 3, se muestra el gráfico que ..."; "Los resultados finales se encuentran en la Tabla X". Usar los ref y labels para que se actualicen las numeraciones automáticamente. No se debe hablar de “en la siguiente figura”, ¿Qué pasa si reordeamos el texto o lo hace látex de manera automática?.
- Cuando se referencia a una tabla, figura, sección, etc. hacerlo con la primera letra en mayúscula, “En la Sección 1 ...”
- El texto siempre debe tener una alineación justificada.
- En el inicio de cada sección se debe hacer una breve introducción de lo que contiene y al final, unas pequeñas conclusiones y si es posible un texto de pequeña longitud que una con el siguiente capítulo.

4.1. Inserción de Imágenes

4.2. Inserción de órdenes de línea de comandos

```
gcc -o e21 e21.c
```

4.3. Inserción de Código

Listado 1: Ejemplo de código

```
1
```

Figura 1: Árbol antiguo

```
##include <stdio.h>
#include <sys/types.h>
#include <unistd.h>
#include <stdlib.h>

int main(void){

int register i;
int numHijos=4;
pid_t childpid;

for (i=0;i<numHijos;i++)
  if (childpid=fork()==0){
    sleep(1);
    printf("Proceso %ld con padre %ld\n", (long)getpid(), (long)
       getppid());
    exit(0);
  }

printf("Soy el proceso padre %ld\n", (long)getpid());
return 0;
}
```

5. MEDIOS QUE SE PRETENDEN UTILIZAR

5.1. Medios Hardware

El alumno deberá describir los medios hardware que prevé serán necesarios para el desarrollo del proyecto.

5.2. Medios Software

El alumno deberá describir los medios software (lenguajes, entornos de desarrollo, herramientas de gestión y planificación, etc.) que prevé serán necesarios para el desarrollo del proyecto

6. REFERENCIAS

En esta sección se incluirán todas las referencias bibliográficas, ordenadas alfabéticamente por el primer apellido del primer autor, de las obras de las cuales se haya realizado alguna cita en los apartados anteriores. Las referencias deberán contener datos básicos como nombre y apellidos de los autores, título de la obra, evento al que pertenece, páginas, fecha y lugar de celebración (si se tratara de artículos de congreso), ISBN, editorial y ciudad (si se tratara de libro), nombre de revista, páginas, volumen y número (si se tratara de revista), etc.

Se empleará un formato de referencia reconocido en el ámbito académico como ACM[1][2]. Otros formatos aconsejables son, por ejemplo, IEEE, AMA, APA y AMA.

A continuación una sección de «Referencias» con ejemplos de referencias con formato ACM para:

- Un artículo de revista [**?**].
- Un informe técnico [**?**].
- Un libro [**?**].
- Un capítulo de libro [**?**].
- Un artículo en las actas de un congreso [**?**].
- Para una página web [**?**] (con autores conocidos).
- Para una página web [**?**] (con autores desconocidos).

[1] http://www.acm.org/sigs/publications/proceedings-templates
[2] http://www.cs.ucy.ac.cy/~chryssis/specs/ACM-refguide.pdf

Referencias