

UNIVERSIDAD DE CASTILLA-LA MANCHA ESCUELA SUPERIOR DE INFORMÁTICA

GRADO EN INGENIERÍA INFORMÁTICA

**Identificación automática de personajes en textos de ficción
pertenecientes al género fanfiction**

María González Gutiérrez

Febrero, 2021

UNIVERSIDAD DE CASTILLA-LA MANCHA

Departamento de Tecnologías y Sistemas de Información

GRADO EN INGENIERÍA INFORMÁTICA

Computación

**Identificación automática de personajes en textos de ficción
pertenecientes al género fanfiction**

Autor: María González Gutiérrez

Tutor académico: José Ángel Olivas Varela

Febrero, 2021

Índice

1. PÁGINA DE CALIFICACIÓN	3
2. RESUMEN	4
3. ABSTRACT	6
4. AGRADECIMIENTOS	7
5. INTRODUCCIÓN Y OBJETIVOS	8
6. FANFICTION Y ARCHIVE OF OUR OWN	12
7. EXTRACCIÓN DE INFORMACIÓN EN OTROS TRABAJOS	15
8. RECOGIDA Y LIMPIEZA DE DATOS	17
8.1. Creando un scraper para Archive of our Own	17
8.2. Limpieza de datos y creación de datasets	25
9. EXTRACCIÓN DE DATOS A PARTIR DE TEXTO	29
9.1. Algoritmo de identificación de entidades	29
9.1.1. Extracción de entidades con NLTK	29
9.1.2. Extracción de entidades con CoreNLP	33
9.2. Algoritmo de identificación de relaciones	39
9.2.1. Primeras estrategias: Clustering y LDA	41

9.2.2. Correferencia con CoreNLP	44
10. PROGRAMA PRINCIPAL: fic_character_extractor	52
11. EVALUACIÓN DEL SISTEMA	54
11.1. Prueba 1: Funciones básicas del programa con un texto largo (Fanfic 9)	54
11.2. Prueba 2: Género distinto del canon (Fanfic 2856)	56
11.3. Prueba 3: Personajes que no son nombrados (Fanfic 2163)	57
12. CONCLUSIONES	62
12.1. Trabajos futuros	65
13. REFERENCIAS	67

1. PÁGINA DE CALIFICACIÓN

TRIBUNAL:

- Presidente:
- Vocal:
- Secretario:

FECHA DE DEFENSA:

CALIFICACIÓN:

PRESIDENTE

VOCAL

SECRETARIO

2. RESUMEN

La extracción de información es una tarea consistente en identificar las entidades presentes en un texto y qué relaciones las unen. En este trabajo se aborda una posible aplicación de este proceso en obras literarias, en particular las pertenecientes al género fanfiction, que se caracteriza por que sus autores no crean obras originales: un fanfic es un relato que toma prestados los personajes o la historia de una obra ya existente. La finalidad de este proyecto pues es crear una herramienta que pueda resultar de utilidad a la hora de realizar análisis literarios sobre los temas, conflictos y carencias que los fans de una obra perciben en la misma.

En la primera parte del proyecto consiste en descargar una parte de los relatos de la web [Archive of Our Own](#), con el objetivo de crear un corpus básico de textos que utilizar para pruebas y entrenamientos. Esto se realiza mediante un sistema de scrapers, y además se desarrollan las herramientas necesarias para organizar y manejar los archivos del corpus para que los algoritmos de procesamiento de texto puedan entender su contenido. Dichos algoritmos utilizarán tanto el texto en sí como los metadatos. Para ello se explora la estructura de la página web y se echa mano de librerías de python como *BeautifulSoup*, *HTML2Text* y *requests*.

La siguiente parte del proyecto es el procesado de texto natural. Utilizando NLTK, se entrena un modelo de regresión logística con el corpus *Groningen Meaning Bank*, dando como resultado un algoritmo de identificación de entidades nombradas que es capaz de encontrar los nombres de los personajes de un texto y contar cuántas veces se les menciona. Para desarrollar un algoritmo de identificación de relaciones se utilizan librerías como *sci-kit learn*, *gensim* y *Stanza* para explorar diferentes técnicas con el fin de hallar relaciones sociales entre personajes, incluyendo clustering, un modelo LDA y explotando la correferencia pronominal en frases en las que un personaje es mencionado para buscar patrones, pero no se llega a ninguna conclusión satisfactoria. Sin embargo, el uso de CoreNLP (mediante *Stanza*) en estas pruebas resulta en el desarrollo de otro algoritmo de identificación de personajes que tiene las ventajas añadidas de ser capaz de identificar el género del personaje y hallar más menciones que el modelo de regresión logística.

Por último, ambos algoritmos de identificación de personajes llevan a cabo una tarea de canonicalización, que en este proyecto es una tarea definida por las características del género fanfiction: puesto que estos relatos no son originales, sino que están basados en personajes ya existentes, estos algoritmos identificarán cuáles personajes se encontraban en la obra original y cuáles son originales del autor fan. Para reducir el alcance de esta tarea, todos los relatos descargados de [Archive of Our Own](#) están basados en

el libro *Good Omens* (Neil Gaiman y Terry Pratchett, 1990).

3. ABSTRACT

Information extraction is the task of recognizing the entities present in a text, and which relationships exist between them. This project attempts an application of this task on literary works, in particular to those belonging the genre of fanfiction, which is defined by not being original works: a fanfic is a story that lends the characters or the plot from an already existing work. The purpose of this project is to create a tool that helps in the literary analysis of the themes, conflicts and oversights that the fans of a particular work find on it.

The first part of the project consists in downloading texts from the website [Archive of Our Own](#), thus creating a basic corpus of stories to be used in tests and training. This is achieved with a scraper system, and developing tools to organise and handle the corpus' files so that the text processing algorithms can understand their contents. Said algorithms will use both the pure text and the metadata of the files. The structure of the website is studied in order to create the scrapers, and python libraries like *BeautifulSoup*, *HTML2Text* and *requests* are used.

The next part of the project is the natural text processing. A named entity recognition algorithm is developed using NLTK, by training a logistic regression model on the *Groningen Meaning Bank* corpus. The result is an algorithm that can recognize a character's name in a text, and count how many times it is mentioned. Several strategies are explored in order to find an algorithm for recognising relationships, and python libraries like *sci-kit learn*, *gensim* and *Stanza* are used to create clustering algorithms, a LDA model and to exploit pronominal coreference in sentences where characters are mentioned in order to find patterns that may signify a relationship, but none of these attempts are successful. However, the use of CoreNLP via *Stanza* ends up creating another named entity recognition algorithm which has the ability of identifying the gender of a character, and can find more mentions than the logistic regression model.

At last, both named entity recognition algorithms perform a task of canonicalization, which in this project is defined by the features of the fanfiction genre: since these stories are not original, but instead they are based in already existing characters, these algorithms identify which characters belong to the original work, and which are created by the fan writer. In order to reduce the scope of this task, all the downloaded works from [Archive of Our Own](#) are based on the book *Good Omens* (Neil Gaiman and Terry Pratchett, 1990).

4. AGRADECIMIENTOS

palabra. otra palabra. palabra!!!

5. INTRODUCCIÓN Y OBJETIVOS

En este proyecto se desarrollará un sistema para la identificación y extracción de personajes y relaciones en textos de ficción, en particular aquellos pertenecientes al género fanfiction. Consistirá, por un lado, en realizar un scrape parcial de la página web [Archive of our Own](#) para obtener un corpus de textos con los que entrenar y hacer pruebas, y por otro lado se desarrollarán los algoritmos encargados de la identificación de entidades y las relaciones entre ellas. De esta manera se pretende crear una herramienta que ayude en análisis literarios, en particular aquellos que se centren en investigar cuáles son las interpretaciones que los lectores fuera el mundo académico encuentran en una obra particular.

Esta herramienta para análisis literario estará basada en los personajes del texto, y las relaciones entre ellos. Al fin y al cabo, es útil para un académico comprobar cuáles son los personajes que más captan la atención de los lectores, si son los protagonistas o si optan por desarrollar personajes menores a los que el autor original no prestó tanta atención, y también puede aprender mucho de qué relaciones y a qué personajes involucran: cuando un gran número de lectores elige enfrentar a personajes que en la obra original son amigos, o crear amistades y romances entre personajes que son enemigos (o que directamente ni se conocen), un académico puede sacar conclusiones sobre qué encuentran atractivo o repulsivo de un personaje, o qué temas quieren resaltar y explorar mediante el desarrollo de estas relaciones.

Por tanto, es una tarea adecuada para la extracción de información a partir de texto natural. El análisis de lenguaje natural tiene como objetivo ahorrar el trabajo humano de resumir y sacar conclusiones de grandes conjuntos de textos, creando programas capaces de procesar información no estructurada y dotándoles de la habilidad de entender el significado del lenguaje humano. En este proyecto se pretende utilizar esta técnica para ahorrar a un académico el tener que leerse manualmente cientos de relatos para concluir cuáles son los temas, conflictos y personajes que le interesan a una comunidad de escritores en particular.

En primer lugar se va a crear un corpus de relatos pertenecientes al género fanfiction a partir de los textos alojados en la web gratuita [Archive of our Own](#); este corpus será útil para realizar pruebas, entrenamiento y experimentos que ayudarán en el desarrollo de un sistema de procesamiento de texto natural que extraerá personajes nombrados de un texto y qué relaciones sociales les unen, utilizando el libro *Natural Language Processing*, de Jacob Eisenstein [Eis18] como guía general del proceso.

El proceso de extracción de información que el proyecto va a seguir consistirá en identificar primero las entidades que se corresponden con los personajes del texto, y después identificará las relaciones entre dichos personajes. La identificación de eventos en un texto es también parte habitual del proceso de

extracción de información, pero decidí centrarme sólo en entidades y relaciones para acotar el problema. La intención es aplicar estos dos procesos a un conjunto de textos pertenecientes al género fanfiction para identificar qué personajes aparecen en el mismo y qué relaciones los unen, así como otros datos relevantes al campo del fanfiction como si el personaje aparece en la obra original, o si es un añadido del autor fan. El usuario podrá utilizar esta información para sacar conclusiones sobre la interpretación del autor sobre los personajes de la obra original.

Los textos serán ser extraídos directamente de AO3 utilizando un scraper, en una fase de recogida y limpieza de datos explicada en las secciones 8.1 y 8.2. Por tanto, los **objetivos de este proyecto** consisten en:

1. Extraer un conjunto de relatos alojados en [Archive of Our Own](#) sobre los que utilizar técnicas de extracción de información.
2. Crear módulos y herramientas para manejar y extraer información de los archivos HTML de AO3.
3. Organizar estos archivos de alguna forma y procesarlos para que sean comprensibles para los algoritmos de análisis de texto.
4. Desarrollar un algoritmo que identifique personajes en textos de ficción.
5. Desarrollar un algoritmo que identifique relaciones de carácter social entre dichos personajes.
6. Crear un programa que utilice ambos algoritmos para extraer los personajes y relaciones en los relatos recogidos de AO3, y mostrarlas al usuario.

Para consultar cualquier aspecto del código e implementación del proyecto, se puede visitar el repositorio del mismo en [GitHub](#) (Usuario: *mariaGnlz*, repositorio: *Fanfic_ontology*).

La idea para empezar el proyecto surgió de la cantidad de información que existe en internet sobre fenómenos culturales como *Los Vengadores* o *Harry Potter*, historias que claramente resuenan con mucha gente y que mueven a los fans a reunirse y crear contenido en comunidades digitales. Estas comunidades de fans suelen ser muy activas y producen una vasta cantidad de contenido muy rico en detalle; son básicamente una gran discusión entre los fans de una obra sobre qué es lo que dicha obra significa para ellos, y sobretodo, qué es lo que a ellos les hubiese gustado llegar a ver realizado en el texto de la misma. A veces, estas comunidades crean enormes proyectos de calidad profesional de forma totalmente gratuita, simplemente para mejorar el espacio y las experiencias del resto de miembros. Uno de los mayores ejemplos de este tipo de 'trabajo fan' es la [Organizaton for Transformative Works](#), una organización sin ánimo de lucro creado 'por fans y para fans' que crea y mantiene proyectos como [FanLore](#), una wiki sobre

la cultura fan, o su comité legal, que se encarga tanto de educar sobre leyes de propiedad intelectual y el *fair use* del mismo como de involucrarse en los procesos jurídicos sobre copyright de diversos gobiernos (especialmente Estados Unidos) para defender el derecho del público general a crear obras derivadas.

Uno de sus proyectos más famosos es [Archive of our Own](#), comúnmente acortado a AO3, un sitio web que aloja principalmente relatos pertenecientes al género fanfiction y cuyo objetivo es, por un lado, facilitar la tarea de encontrar fanfiction para aquellos que lo quieran leer y, por otro, funcionar como un archivo que clasifique y documente el fenómeno fanfic a nivel global. En sus datos de mayo de 2020, aparecen más de dos millones de usuarios registrados, y más de seis millones de trabajos alojados. AO3 se ha convertido en una parte fundamental de la cultura fan, especialmente la dedicada a la escritura, y en este proyecto lo utilizaré como fuente de información.

En literatura comparada se suelen tener en cuenta dos perspectivas a la hora de analizar una obra: la de la teoría del autor, que tiene en cuenta lo que el autor quería comunicar y plasmar en esa obra, y la de la 'muerte del autor' [Bar68], que tiene en cuenta el mensaje con el que los lectores se quedan tras leer la obra (independientemente de si coincide con el que el autor quería comunicar). Para entender el mensaje de una obra de forma plena [EII18], es necesario tener en cuenta tanto la intención comunicativa del autor, como el mensaje que al final los lectores acaban entendiendo. Y como cada lector es hijo de su padre y de su madre, acaban surgiendo muchas posibles interpretaciones distintas a partir de un único texto.

Tradicionalmente, los académicos solamente han tenido en cuenta las opiniones de un grupo reducido de personas (compuesto principalmente por otros académicos) a la hora de analizar una obra desde la perspectiva de la muerte del autor, ya que el lector común no suele tener a su disposición las herramientas necesarias para difundir sus interpretaciones. Sin embargo, desde que Internet y los foros como *LiveJournal* se volvieron accesibles a grandes partes de la población, miles de comunidades fan empezaron a organizarse justamente con la intención de poner en común sus interpretaciones, de expresar sus críticas y opiniones. No todas estas discusiones tienen lugar en forma de fanfiction, pero es un género muy popular en las comunidades fans, y yo personalmente estoy muy familiarizada con sus estructuras y códigos.

Estas comunidades de Internet están generando una cantidad inmensa de opiniones y perspectivas en torno a un tema común en foros públicamente accesibles, y me pareció interesante la idea de crear un sistema que sea capaz de recoger y procesar toda esta información para crear un 'abanico' de las distintas interpretaciones que existen en una comunidad fan, especialmente aquellas sobre los personajes y las relaciones entre ellos. El resultado final se podría utilizar como herramienta dentro de la propia comunidad fan, para observar cómo tienden a interpretar a ciertos personajes a nivel de comunidad y cómo estas

interpretaciones cambian a lo largo del tiempo, o en distintas subsecciones dentro de la comunidad en general. También se podría utilizar como herramienta general de análisis literario, aplicándola primero a la obra original y luego a un conjunto de fanfics representativos, y observando cuáles son las diferencias entre la perspectiva del autor original y la de los lectores (convertidos en autores fan).

6. FANFICTION Y ARCHIVE OF OUR OWN

Fanfiction (del inglés *fan fiction*, 'ficción del fan', y abreviado como 'fanfic') es el nombre que recibe un texto basado en una historia ya existente (normalmente con copyright), en particular cuando el autor es fan de la obra de la cual su texto deriva. Son, por lo tanto, textos de ficción sin ánimo de lucro que los fans escriben como expresión de su creatividad.

El concepto detrás del fanfiction es, en esencia, una ausencia percibida en la historia original. Uno se termina un libro o un videojuego y siente que le falta algo: el pasado de un protagonista, una perspectiva distinta de un conflicto, una relación que acabó o nunca empezó, qué sucede después del final, o quizás que a la historia le hacían falta doscientas páginas más, o incluso que tendría que haber sido de un género literario distinto... Hay algo en la historia que está ausente. El lector se queda con ganas de explorar más a fondo el mundo y los personajes que el autor ha creado, y de aquí nace el impulso de crear historias propias en las que se exploran dichas ausencias. Por tanto, no es sorprendente descubrir que hay muchos fanfics en los que se cambia el destino de tal o cuál personaje, que exploran qué sucede tras el final, o que llevan a cabo exploraciones exhaustivas de los conflictos, los personajes y sus motivaciones desde perspectivas distintas a las de la obra original.

Todos estos motivos hacen que el fanfiction se considere una obra derivada [Swi98], y está en su naturaleza el reflejar las opiniones y críticas que el autor tiene de la obra original: qué es lo que le gusta, qué temas siente que faltan en la obra, qué cosas tendrían que haberse explicado desde una perspectiva distinta, etc.

Por ejemplo, es evidente al leer los libros de la saga *Harry Potter* que el texto quiere que pienses que Ron Weasley, el mejor amigo del protagonista, es un chico un poco torpe y bocazas pero con buen corazón, y un buen amigo de Harry. Sin embargo, muchos fans no interpretaron a Ron como torpe y bocazas, sino como egocéntrico e insensible, y hay no pocos fanfics en los que Ron y Harry discuten y dejan de ser amigos, o en los que Ron es directamente un villano aliado con Voldemort.

Cuando los fans de una misma obra se reúnen y organizan, se crean comunidades fan llamadas "fandoms", que suelen crear foros donde intercambiar sus impresiones, teorías y, por supuesto, fanfiction y otras formas de arte fan. Es evidente que existe un intercambio de ideas en foros de discusión y otras comunidades online explícitamente creadas para conversar, pero ya que es totalmente posible inferir las opiniones de un autor a partir de sus fanfics, tanto escribir como leer fanfiction son actividades que contribuyen al discurso general del fandom, ayudando a popularizar algunas teorías y generando las suyas propias.

Cuando un fandom alcanza un cierto nivel de madurez, algunas teorías se consolidan y el fandom acaba

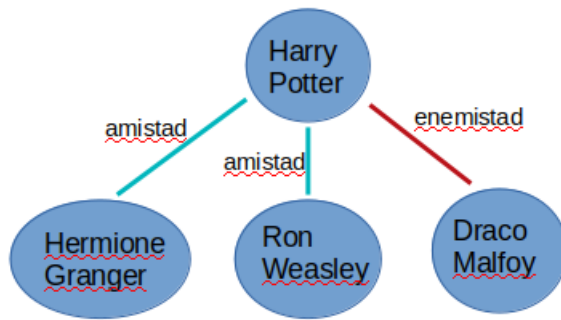
formando, a nivel de comunidad, una interpretación propia de la obra original. Para distinguir la perspectiva del fandom de la que realmente pretende transmitir la obra original, en los fandoms se distingue entre el *fanon* y el canon. Siguiendo el ejemplo de *Harry Potter*, el Ron Weasley del *fanon* es una persona egoísta que sólo es amigo de Harry por interés, mientras que el Ron Weasley del canon tiene una amistad sincera con Harry. *Fanon*, por tanto, es el 'conjunto de teorías basadas en el material original que, aunque generalmente parecen ser la interpretación 'obvia' o 'única' de los hechos canónicos, no son realmente parte del canon' [Stu17].

En resumen, las comunidades fan tienen una interpretación propia de la obra original llamada "*fanon*", que influencia los fanfics que los miembros de dicha comunidad van a escribir y, a su vez, los escritores de fanfic también crean y popularizan interpretaciones que se acaban convirtiendo en parte del *fanon*.

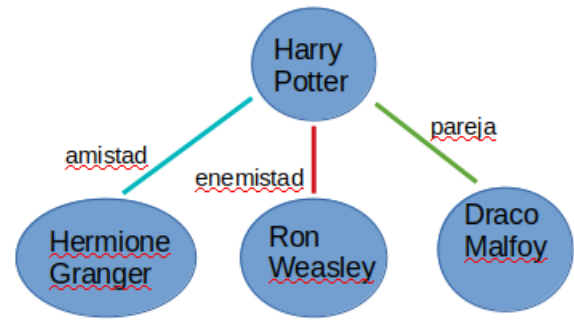
Como se ve en el ejemplo de *Harry Potter*, las relaciones entre personajes son una de las mayores fuentes de especulación entre los fans, especialmente las relaciones románticas. En general, los personajes a los cuales los fans les tienen manía acaban convertidos en villanos (o, como mínimo, enemigo de los protagonistas) en los fanfics, incluso aunque en la obra original sean aliados. Naturalmente, lo mismo sucede a la inversa: los fans tienden a convertir en amigos y aliados a los personajes que les gustan, incluso aunque en la obra original sean los villanos de la historia. Por tanto, simplemente contrastando las relaciones presentes en un fanfic con las relaciones de la obra original podemos tener una buena idea de cuál es la interpretación del autor del fanfic.

Las relaciones románticas entre personajes son una parte enorme de la especulación fan. El romance es uno de los temas más populares, y aunque las relaciones canónicas atraen naturalmente la atención de muchos fans, 'inventar' parejas en el *fanon* no sólo es común, sino una de las principales actividades de un fandom. Los fans ven parejas y conflictos amorosos tanto entre amigos como enemigos, y son felices de ignorar todos y cualquiera de los obstáculos que existan en el canon con tal de tener el escenario necesario para que su pareja preferida pueda estar junta, llegando incluso al extremo de sacar a los personajes del universo al que pertenecen para meterlos en otro más amistoso. Un villano que es muy popular entre los fans tiene garantizados fanfics en los que cambia de bando, convirtiéndose en aliado y pareja del protagonista (no necesariamente en ese orden).

Como se ha dicho anteriormente, los fans se organizan en comunidades según la obra que es el objeto de su admiración, y tienen sitios como AO3, dedicados a alojar y compartir sus creaciones fan. Evidentemente, analizar las más de seis millones de obras existentes en AO3 es una tarea imposible con los recursos a mi alcance, de modo que elegí utilizar únicamente los fanfics basados en *Good Omens*, un libro de Terry Pratchett y Neil Gaiman, en parte por mi familiaridad con esa comunidad, pero también porque tenía una cantidad de fanfics extensa pero manejable.



A) Grafo de relaciones entre personajes en los libros de la saga *Harry Potter*



B) Grafo de relaciones entre personajes que se encuentra fácilmente en los fanfics de *Harry Potter*

AO3 no es el único sitio web popular para alojar fanfiction, pero a lo largo de esta década ha desplazado a sitios como [Fanfiction.net](https://www.fanfiction.net) y [Wattpad](https://www.wattpad.com), y la característica que condicionó mi decisión es su extensivo sistema de etiquetas, que como se verá más adelante fue fundamental para filtrar y gestionar los datos del proyecto.

En términos legales y de derechos de autor, la mayoría de legislaciones considera el fanfiction como un tipo de obra derivada [Swi98] y por tanto entra dentro del *fair use*.

7. EXTRACCIÓN DE INFORMACIÓN EN OTROS TRABAJOS

La extracción de información es un problema que se aborda en el análisis de lenguaje natural, cuyo objetivo es ahorrar el trabajo humano de resumir y sacar conclusiones de grandes conjuntos de textos. Para lograrlo, se crean programas capaces de procesar información no estructurada, dotándoles de la habilidad de entender el significado del lenguaje humano para que resuman y saquen conclusiones a partir de grandes volúmenes de textos ya existentes y, por ejemplo, usen esa información para crear una base de datos de conocimiento automáticamente, sin tener que dedicar grandes cantidades de tiempo y esfuerzo a leerlos manualmente. Es especialmente útil en campos como la medicina y la química, para los cuales se han creado distintos modelos [Cra99][Man19] con el objetivo de ayudar a procesar las ingentes cantidades de estudios y artículos que se publican hoy en día.

El primero de los pasos para extraer información de un texto es identificar las entidades nombradas en el mismo. Algunos sistemas, como CoreNLP [Fin05], utilizan un clasificador estadístico (en particular, un campo aleatorio condicional [Laf01]) para identificar las palabras que denotan un nombre, lugar u organización. La desventaja de éstos métodos es que requieren ser entrenados con datos previamente etiquetados y listas de nombres bien conocidos (nomenclátor), por lo que también se han buscado alternativas no-supervisadas [not13] y semi-supervisadas [Lin09], pero las estrategias que echan mano del *machine learning* y grandes corpus de texto etiquetado para entrenar clasificadores siguen siendo las más populares. Existen por tanto muchas soluciones distintas, algunas dotadas de extractores de características muy sofisticados compuestos por varios algoritmos entrenados para realizar tareas específicas, como capturar el contexto o incluso canonicalizar entidades [Wic09].

La extracción de relaciones tradicionalmente se ha realizado tratando de identificar relaciones binarias, bien frase a frase o teniendo en cuenta dos o tres frases consecutivas [Zel03][Cra99]. Las limitaciones de este sistema no son sólo que no todos los tipos de relaciones son binarias, si no que cuanto más largo es un texto, menos probable es que ambas entidades aparezcan nombradas en una única frase, con lo que se pierde mucha información si se ignoran. Pasar a un sistema que tenga en cuenta varias frases interdependientes viene con sus propios problemas, pues las características sintácticas de una relación son muy dispersas en el texto y tratar de extraerlas automáticamente requiere de mucha memoria y muchos datos; más cuántas más frases tengas en cuenta. Sin embargo, recientemente se han creado algoritmos que identifican una relación n-aria a partir de todas las menciones en las que aparece, como el de Peng et al [Pen17], que utiliza grafos LSTM para modelar el contexto e información de cada frase y las conexiones entre ellas, logrando así beneficiarse de la interdependencia de distintas frases sin tanto coste de recursos.

Este proyecto sin embargo busca identificar relaciones de naturaleza social entre distintos personajes de

ficción, con lo que se centrará en identificar entidades del tipo 'Persona', y buscará relaciones binarias entre ellas.

Aunque el fanfiction no ha sido muy utilizado para realizar tareas de extracción de información, su disponibilidad online y la cantidad de metadatos asociados a cada obra ha hecho que algunos programadores los aprovechen para crear sistemas de recomendación, como [FanRecs](#), aunque el principal interés parece residir en crear herramientas de descarga ([FicLab](#), [FanfictionDownloader](#)).

8. RECOGIDA Y LIMPIEZA DE DATOS

8.1. Creando un scraper para Archive of our Own

En el momento en el que decidí utilizar los fanfics de *Good Omens* para el proyecto, dicho libro tenía unos 22000 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 relatos, permite alojar todo tipo de archivos multimedia).

Por suerte, AO3 fue creado con la intención específica de funcionar como archivo, por lo que tiene una herramienta de búsqueda y filtrado muy completa y sencilla de usar. Esta herramienta permite filtrar por características como título, autor, idioma y cantidad de palabras, pero su mayor utilidad viene de su sistema de etiquetado. AO3 permite a los autores añadir tantas etiquetas como quieran para que los posibles lectores puedan saber más de su obra a simple vista: temática, personajes principales, parejas en las que se centra, qué medio utiliza, si hay ilustraciones, si trata sobre un evento de la historia original particular... Las etiquetas añaden una gran cantidad de información sobre las historias a las que acompañan, y aunque no es obligatorio poner ninguna, en general los autores se preocupan de etiquetar correctamente sus obras.

AO3 tiene etiquetas específicas para indicar que una obra no es principalmente texto: '*Fanart*', para ilustraciones, y '*Podfic*' para archivos de audio, así que aproveché la herramienta de búsqueda para llevar a cabo un primer filtrado que eliminara todas las obras que las contuvieran, además de todas las que no estuviesen en inglés. El resultado fue un subconjunto de 20190 fanfics, todos en inglés y cuyos autores no habían incluido ninguna etiqueta que indicara que no fuera puro texto. La herramienta además genera un link permanente que siempre lleva a este subconjunto particular, por lo que no es necesario utilizar esta herramienta nada más que una vez.

Una vez localizado el conjunto de textos y el link a los mismos, viene la parte de crear el *scraper* en sí. Utilizando la herramienta de inspeccionar elemento de *Firefox* para explorar la estructura del sitio, y enseguida se hizo obvio que los fanfics estaban organizados en páginas con un máximo de 20 fanfics cada una. En el HTML de la página, cada fanfic se presenta dentro de una clase llamada '*work blurb group*'. No se puede extraer un link de descarga directamente de ésta clase, pero sí el identificador del fanfic.

Exclude ?

- Ratings
- Warnings
- Categories
- Fandoms
- Characters
- Relationships
- Additional Tags

Other tags to exclude

Fanart ☒

Podfic ☒

More Options

- Crossovers
- Completion Status
- Word Count
- Date Updated

Search within results ?

Language

English ▼

Sort and Filter

Figura 1: Herramienta de filtrado de AO3. Permite excluir (o incluir) obras que contengan etiquetas específicas, así como aquellas no escritas en un idioma particular

En AO3, cada fanfic tiene un número que lo identifica de forma única. Es posible acceder a la página de cualquier fanfic simplemente añadiendo ese número al final de 'https://www.archiveofourown.org/works/' en la barra de direcciones, y en esa página sí que se pueden encontrar links de descarga. Por tanto la idea básica para el *scraper* es utilizar las librerías *requests* y *BeautifulSoup* de python para explorar los veinte '*work blurb group*' de cada página, localizar el identificador de cada uno, utilizarlo para acceder a la página del fanfic y extraer el link de descarga. Y así con cada página del listado, hasta llegar a la última. La figura 2 ilustra el proceso con un esquema.

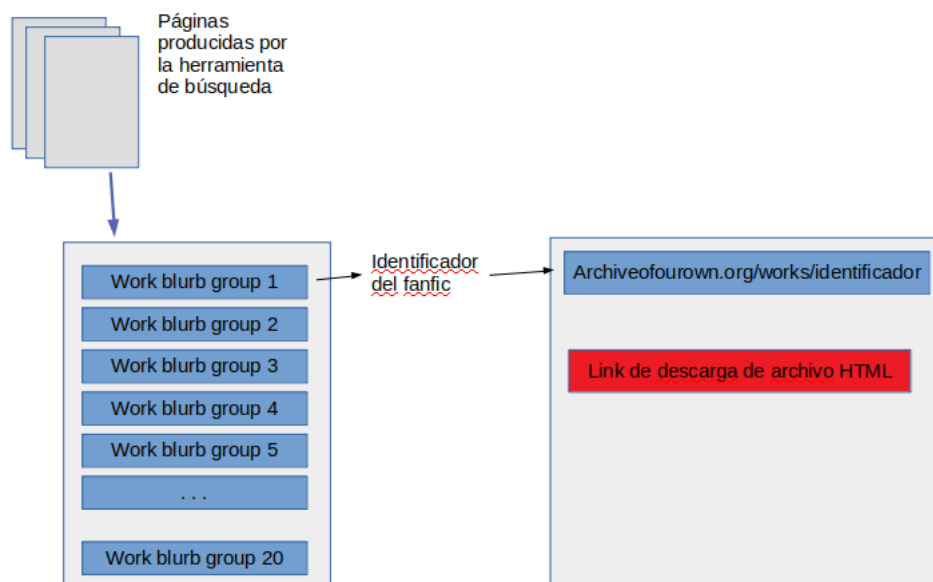


Figura 2: Concepto para el *scraper*. El objetivo es obtener los links de descarga navegando las páginas de búsqueda.

El proceso de descarga de archivos, en principio, tendría estos pasos:

1. Enviar una petición HTTP GET al link permanente del conjunto de datos, generado por la herramienta de búsqueda de AO3.
2. Iterar entre los 20 '*work blurb group*' y extraer el identificador de cada uno.
3. Usar el identificador para acceder a la página de cada fanfic, extraer el link de descarga de la página, y descargar el fanfic como archivo HTML. Hacer esto con los 20 identificadores.
4. Pasar a la siguiente página y repetir, hasta llegar a la última.

Utilizando la librería *requests* de python, el primer paso es trivial, y se puede observar en la figura 5. Encontrar los identificadores tampoco es complicado. Se puede apreciar en 2 que el identificador del fanfic también es el ID del objeto '*work blurb group*' al que pertenece, y expandiendo la clase se puede ver que el identificador completo se puede encontrar dentro del objeto, como un objeto de tipo *h4*. Por tanto, usando *BeautifulSoup* para manejar los datos resultantes de la petición HTTP GET como objeto HTML, se pueden obtener todos los objetos '*work blurb group*' usando la función *find(class_=<nombre clase>)*, cuyo resultado es una lista con los 20 objetos, sobre los cuales se itera para encontrar los identificadores usando de nuevo la función *find()*. En la figura 4 se puede observar un fragmento del código que realiza este trabajo; el código completo se puede consultar en [placeholder ref].

Las complicaciones empiezan una vez se tienen los identificadores. Para formar la dirección completa,

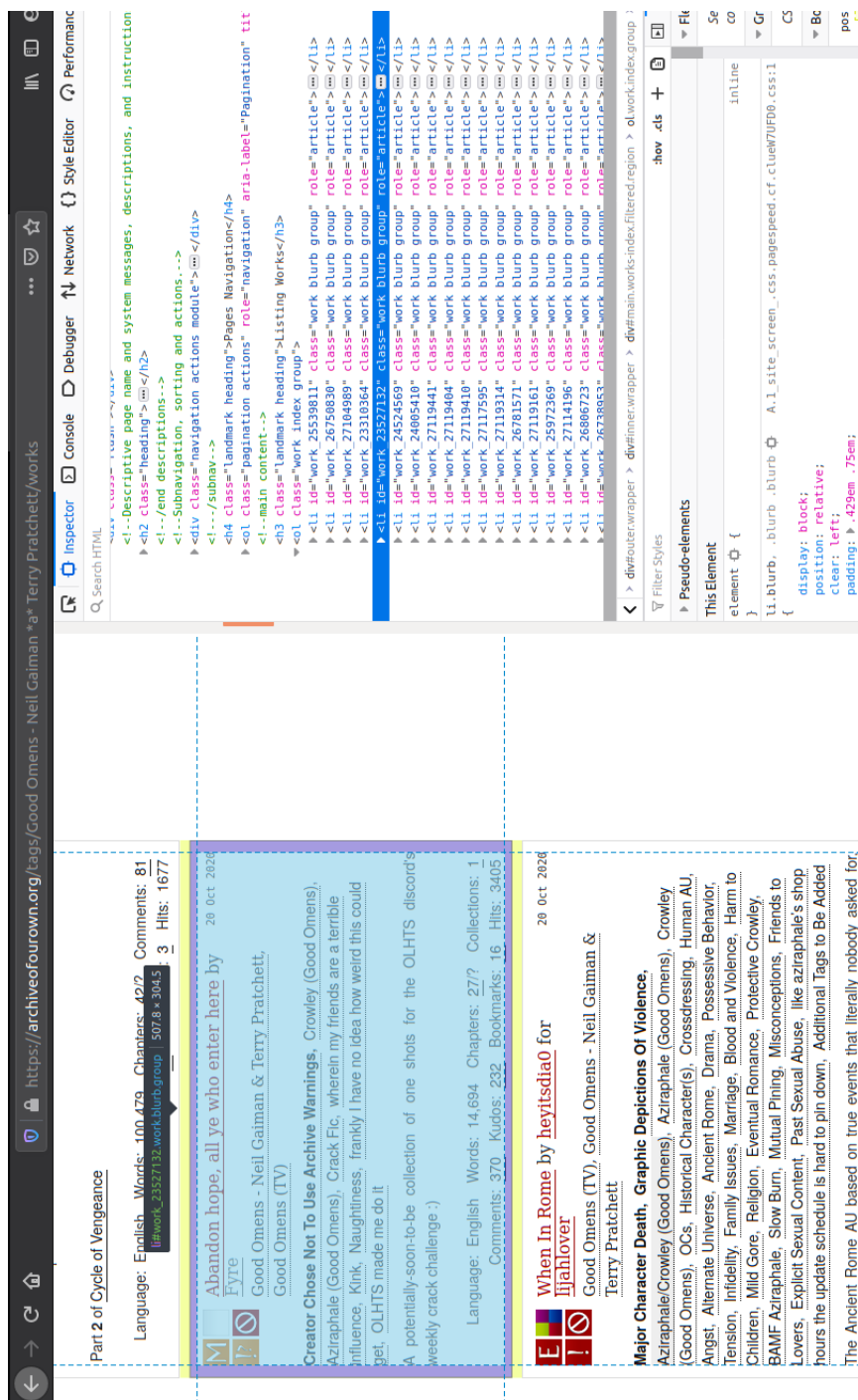


Figura 3: Exploración de la estructura de la página AO3 usando la herramienta 'Inspeccionar elemento' de Firefox. Se puede ver que el sitio utiliza una clase HTML llamada 'work blurb group' para mostrar cada obra.

```

40     current_page = 1
41     while current_page < number_of_pages:
42         blurbs = soup.find_all(class_='work blurb group')
43         #print('current page: ',current_page) #debug
44
45         for blurb in blurbs:
46             #filter out fics that don't contain text
47             contains_text = check_for_text(blurb)
48
49             work_id = (blurb.find('h4')).find('a')
50             if contains_text: work_links.append('https://archiveofourown.org'+work_id['href'])
51             else:
52                 discarded_links.append('https://archiveofourown.org'+work_id['href'])
53                 #print('out:', work_id['href'])
54         #end 'for blurb' loop
55
56         current_page +=1
57         next_page_link = page_link.replace('&page=1&', '&page='+str(current_page)+'&')
58         while True: #wait out if too many requests
59             page = requests.get(next_page_link)
60
61             if page.status_code == 429: #Too Many Requests
62                 print('Sleeping...')
63                 time.sleep(120)
64                 print('Woke up')
65
66             else: break
67
68         soup = BeautifulSoup(page.content, 'html.parser')
69
70     #end while loop

```

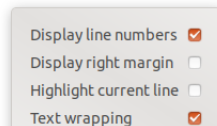


Figura 4: Código perteneciente al *scraper* 'ao3_link_scraper'. Utiliza un bucle *while* para iterar entre las páginas de la búsqueda, y en cada página, usa la librería *BeautifulSoup* para extraer los objetos 'work blurb group' en una lista llamada 'blurbs' (línea 42). De cada 'blurb' extrae el identificador del fanfic y comprueba si tiene texto (líneas 46-49), y si lo contiene forma el enlace a la página del fanfic y lo añade a una lista llamada 'work_links' (línea 50). Si no contiene texto, se añade a otra lista llamada 'discarded_links' (línea 52).

```

29 def get_work_links(page_link):
30     page = requests.get(page_link) #get first page of the archive
31     soup = BeautifulSoup(page.content, 'html.parser')
32
33     #figure out how many pages in total there are
34     page_list = (soup.find(class_='pagination actions')).find_all('li')
35     number_of_pages = int(page_list[len(page_list)-2].text) #there are number_of_pages pages in total
36

```

Figura 5: Código perteneciente al *scraper* 'ao3_link_scraper'. Utiliza la librería *requests* para enviar una petición HTTP GET al link permanente del conjunto de datos (línea 30), y *BeautifulSoup* para navegar el resultado como un objeto HTML del que poder extraer datos útiles, como la cantidad total de páginas (líneas 33-35).

hay que añadir el identificador al final de 'https://www.archiveofourown.org', mandar otra petición HTTP GET a dicha dirección, buscar ahí el link de descarga, solicitarla, esperar a que la descarga termine, y repetir todo esto otras 19 veces hasta tener descargados todos los fanfics de la página. Esto significa que por cada iteración del bucle que explora cada página es necesario introducir otro bucle que haga las descargas.

La última parte, la de pasar a la página siguiente, es más complicada de explicar que de ejecutar. Todas las páginas de resultados de búsqueda de AO3 contienen botones para avanzar, retroceder y saltar a páginas concretas. Es posible saber cuántas páginas en total tiene la búsqueda simplemente observando el texto del botón de la última, tal y como se ve en la figura 7. No se aprecia, pero la clase HTML a la que pertenece dicho botón se llama 'pagination actions', y es posible extraerla gracias a la función

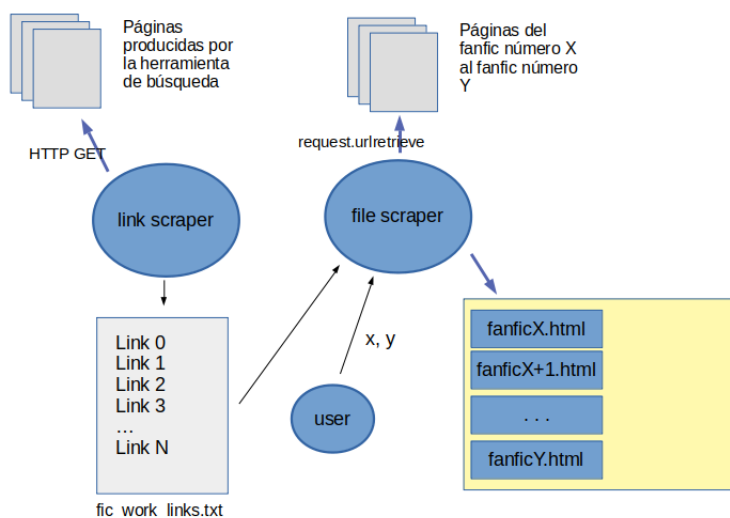


Figura 6: Proceso de descarga de fanfics de AO3 utilizando los programas 'ao3_link_scraper.py' y 'ao3_file_scraper.py'.

find(class_=<nombre_clase>) de *BeautifulSoup*. Y ya con ese objeto, se puede volver a utilizar la función *find()* para buscar todos los objetos hijos de la clase '*pagination actions*' que sean de tipo *li*. El último será el que contenga la cantidad total de páginas, y solicitar la siguiente consiste simplemente en sustituir la referencia en el link a la página 1 por una referencia a la última página. En la figura 5 se ve parte del código que realiza este proceso; el código completo se puede consultar en el anexo [placeholder ref].

Es evidente que la parte de solicitar las descargas en un bucle anidado ralentiza el programa, enturbia el código y además, hace que sea complicado parar o interrumpir el programa si hay algún error de red, pues para reanudar la ejecución por donde se quedó sería necesario almacenar en alguna parte el número de página por el que iba y el número del fanfic dentro de esa página, y programar los bucles para que salten directamente a la iteración deseada.

Ninguna de estas cosas me convenía, ya que descargar más de 20000 archivos ya iba a ser lento de por sí y hacerlo de una sentada sería prácticamente imposible, de modo que decidí dividir el programa en dos: uno que llamé '*link scraper*' y otro '*file scraper*'.

El *link scraper* se ejecutaría una vez y exploraría todas las páginas de búsqueda, extrayendo los links a los fanfics de cada una, y los almacenaría en un archivo de texto. Por tanto, al terminar su ejecución este *scraper* ha generado un archivo llamado '*fic_work_links.txt*' que almacena los enlaces a cada fanfic. El *file scraper* utiliza esta lista para saber dónde buscar las descargas, y el usuario le indica en la línea de comando los índices que acotan el tramo de la lista a descargar, tal y como se ilustra en el esquema de la figura 6. De este modo, es posible indicarle al programa que descargue desde el link 0 al link 1000 de la lista, permitiendo descargar los 20000 archivos en porciones manejables. Además, el programa

anuncia en pantalla qué link está siendo descargado en cada momento, por lo que si sucede un error de red mientras descargaba el link número 866, es posible reanudar el programa fácilmente e indicarle que continúe desde el 866 al 1000 [placeholder ref].

Esta división del trabajo en dos programas además me daba la oportunidad de introducir con sencillez un segundo filtrado durante el proceso de exploración que realiza el link *scraper*. Si el primer filtrado se encargaba de cribar los fanfics que habían sido etiquetados por sus autores como imágenes o audio, este segundo filtrado pretende detectar los fanfics que tampoco contienen texto, pero no han sido etiquetados como tal por sus autores. Para ello usé el criterio de la relación palabras/capítulo de cada fanfic: si una obra tiene menos de 40 palabras por capítulo, se considera como fanfic "sin texto", y se elimina. Escogí 40 palabras como umbral tras investigar un poco con la herramienta de búsqueda de AO3, que como se puede ver en la figura 1, tiene una opción para filtrar por cantidad total de palabras. Tras probar varios umbrales, 40 parecía ser el que descartaba todas las obras sin texto sin sacrificar muchos microrrelatos en el proceso.

Introducir este filtrado en el *scraper* fue sencillo, puesto que el número de palabras y capítulos de la obra es información que se puede extraer de la clase *work blurb group* de cada fic. Todo esto se realiza desde la función *check_for_text*, y en la figura 4 se puede ver cómo el bucle llama a dicha función; el código completo se puede consultar en [placeholder ref]. Por tanto, el link *scraper* realiza estos pasos:

1. Enviar una petición HTTP GET mediante la librería *requests* al link permanente del conjunto de datos, generado por la herramienta de búsqueda de AO3.
2. Iterar entre los 20 '*work blurb group*', comprobar si contienen texto, y descartar los identificadores de los que no.
3. Utilizar cada identificador para generar el link de la página de cada fanfic y almacenarlos en un archivo de texto.
4. Pasar a la siguiente página y repetir, hasta llegar a la última.

Por su parte, el *file scraper* realiza estos pasos:

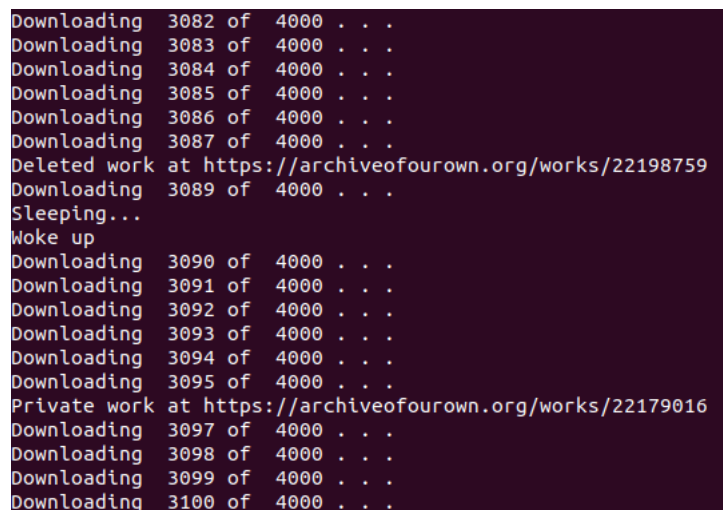
1. Abrir el archivo *fic_work_links.txt* y extraer la lista de links.
2. Mediante la librería *requests*, realizar una petición HTTP GET al primer link, saltando al siguiente si devuelve un código 404.
3. Extraer el link de descarga HTML de cada página.

4. Solicitar la descarga mediante *request.urlretrieve*. Guardar el archivo resultante en la carpeta adecuada en el sistema.
5. Repetir con todos los links de la lista.

El manejo del código de error 404 (Page Not Found) es bastante importante en este *scraper*, puesto que entre el momento en el que se almacenó el link del fanfic mediante el primer *scraper* y el momento en el que el segundo *scraper* lo utiliza para la descarga pueden haber pasado varios días. En ese tiempo, el autor del fanfic puede haber decidido borrar el fanfic de AO3, o haberlo hecho privado, y de ahí que el *scraper* reciba un 404. Un simple *try-catch* detecta el código 404 y simplemente pasa al siguiente link, como se puede consultar en el anexo [placeholder ref].

El otro error que ambos *scrapers* necesitaban manejar es, naturalmente, el error 429 (Too Many Requests). En las líneas 58-66 de la figura 4 se puede ver cómo se utiliza un *try-catch* que envuelve la petición HTTP GET para detectar el status 429 y, en vez de pasar al siguiente link, se lanza una espera de dos minutos tras la cual vuelve a solicitar la página. Antes de incorporar este código a los *scrapers* creé un pequeño programa de prueba, para ver cuánto tardaba AO3 en enviar un 429 y cuánto tiempo de espera requería antes de volver a aceptar solicitudes; dicho programa se puede consultar en [placeholder ref].

El resultado de la ejecución de estos *scrapers* es una carpeta con 818,8 MB de archivos HTML.



```
Downloading 3082 of 4000 . . .
Downloading 3083 of 4000 . . .
Downloading 3084 of 4000 . . .
Downloading 3085 of 4000 . . .
Downloading 3086 of 4000 . . .
Downloading 3087 of 4000 . . .
Deleted work at https://archiveofourown.org/works/22198759
Downloading 3089 of 4000 . . .
Sleeping...
Woke up
Downloading 3090 of 4000 . . .
Downloading 3091 of 4000 . . .
Downloading 3092 of 4000 . . .
Downloading 3093 of 4000 . . .
Downloading 3094 of 4000 . . .
Downloading 3095 of 4000 . . .
Private work at https://archiveofourown.org/works/22179016
Downloading 3097 of 4000 . . .
Downloading 3098 of 4000 . . .
Downloading 3099 of 4000 . . .
Downloading 3100 of 4000 . . .
```

Figura 8: Ejecución de *ao3_file_scraper.py*

8.2. Limpieza de datos y creación de datasets

Al terminar el proceso de descarga, acabé con un conjunto de archivos HTML y un archivo TXT con una lista de los *path* de todos ellos.

La principal tarea de limpieza de datos es, por tanto, convertir los archivos HTML a texto. Para ello utilicé la librería *HTML2Text*, que como cuyo nombre dice sirve para eso mismo. Sin embargo, los fanfics en HTML no contienen sólo el texto del fic en sí, sino que también contienen todos los metadatos del mismo: etiquetas, *rating*, resumen, comentarios del autor, entre otros, con lo que tras limpiar el archivo HTML con *HTML2Text* el resultado no era el texto puro del fanfic. Tuve que crear varias funciones y ayudarme de *Beautiful Soup* para limpiar todos estos metadatos y dejar únicamente el texto en sí, sin llevarme por delante parte del texto ni dejarme notas del autor entre capítulos.

Al principio puse estas funciones dentro de cada programa que necesitaba manejar los textos, pero obviamente enseguida se volvió muy aparatoso, por lo que consolidé todas las funciones de limpieza en un único archivo, *fanfic_util.py*, junto con tres clases para encapsular el uso de estas funciones:

- *FanficGetter*, que se encarga de proveer el texto limpio de un fanfic (o lista de fanfics) a demanda. Al principio devolvía los textos como una lista de *string*, pero luego resultó más útil que devolviera una lista de objetos *Fanfic* que pudiera devolver los capítulos por separado o juntos en un mismo *string*, además del identificador del fanfic.
- *FanficHTMLHandler* se encarga de extraer información de los metadatos del archivo HTML de un fanfic, como por ejemplo los personajes principales, las relaciones, las etiquetas, el número de capítulos y su clasificación.
- *Fanfic*, una clase que se utiliza a lo largo del proyecto para encapsular toda la información relevante sobre un fanfic, como sus capítulos, sus personajes, etiquetas, el dataset al que pertenece e incluso los objetos *Document* generados por CoreNLP.

La creación de la clase *Fanfic* surgió a mediados del proyecto, cuando el límite de 100000 caracteres de *CoreNLP*9.1.2 hizo necesario poder acceder al texto de cada fanfic dividido en capítulos. *Fanfic* por tanto tiene un atributo *chapters*, que es la lista de *string* con los capítulos, y un método *get_string_chapters()* que devuelve todos los capítulos en un único *string*. Según seguía avanzando la clase también demostró ser útil para almacenar en un único elemento toda su información relevante, por lo que termina siendo la unidad básica de trabajo del proyecto.

Sin embargo, el acceso a los datos sigue basándose en una lista de *paths* guardada en un archivo TXT. Es un sistema muy rudimentario (ilustrado en la figura 9), pero no vi necesidad de trasladarlo a una base de datos propiamente dicha, puesto que nada en la creación de una base de datos me ahorraba nada del trabajo de crear *fanfic_util*, y desde el punto de vista del resto de programas es igual acceder al texto de un fanfic a través de *FanficGetter* que de una función que recupere el texto de una base de datos. Únicamente intercambiaría el tiempo de limpiar los textos por el de conectar con la base de datos y extraer su información.

El identificador de entidades puede utilizar la lista original con todos los fanfics como fuente para eti-

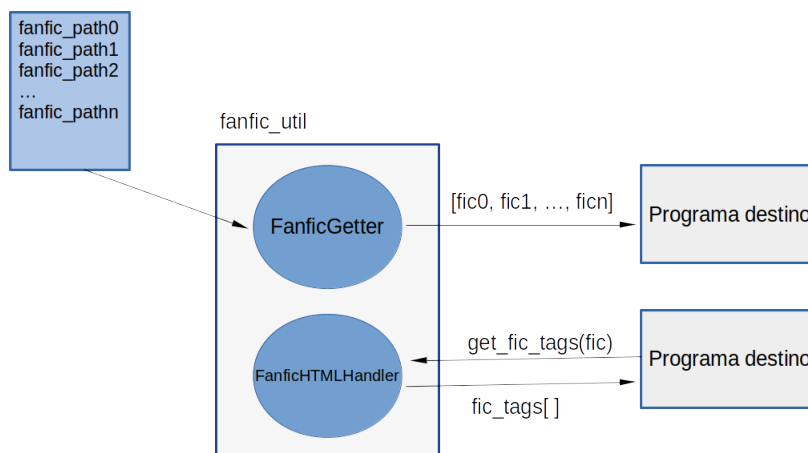


Figura 9: Esquema ilustrando la función de las clases *fanfic_util*

quetarlos, pero para la parte de identificación de relaciones del proyecto se necesita filtrar los fanfics originales en tres subconjuntos: uno centrado en el romance, otro en la amistad, y el último en la enemistad. Por lo tanto, utilicé las funciones de *fanfic_util* para crear un programa, *generate_fic_lists.py*, para crear las listas de *paths* correspondientes a cada grupo.

El criterio utilizado para crear estos tres grupos está basado en la longitud de los textos y sus etiquetas. En el caso de las etiquetas es sencillo: los autores casi siempre etiquetan las relaciones románticas en sus relatos, y a menudo también las amistades, para hacer que sus historias sean más fáciles de encontrar por aquellos que quieran leerlas. En AO3, las etiquetas románticas tienen el formato 'Personaje A/Personaje B', mientras que las etiquetas de amistad son 'Personaje A & Personaje B' o 'Personaje A and Personaje B', con lo que extraer estas etiquetas del archivo HTML es sencillo utilizando expresiones regulares [placeholder ref]. Además de tener una etiqueta que valide la expresión regular, sólo tuve en cuenta aquellos relatos que sólo tenían un capítulo. De esa forma, esperaba poder eliminar historias largas y elaboradas que contienen romance, pero que principalmente son una historia costumbrista o de aventuras. Limitando la longitud de la historia a un capítulo, todo el romance o la amistad queda condensada en dicho capítulo. Además, como en los relatos que contienen abusos sexuales es común etiquetar al perpetrador y a la víctima con el formato de 'Personaje A/Personaje B', realicé un segundo pase a la lista de romance, eliminando todos aquellos relatos que tuviesen la etiqueta 'Rape/Non-con'.

Crear un conjunto con relaciones de enemistad u odio es una tarea más complicada, ya que los usuarios de AO3 no tienen un formato oficial para las mismas y no se suelen etiquetar. Al final utilicé una lista de las etiquetas que los autores comúnmente utilizan como aviso de que su historia contiene violencia

o abusos, como 'Rape/Non-con', 'Torture', 'Graphic Depictions of Violence' o 'Dead Dove: Do Not Eat'[placeholder ref]. Esperaba que, entre esas etiquetas y la limitación de longitud, pudiese aislar un conjunto de relatos que sirvieran de modelo para la relación de enemistad.

El resultado fueron 12520 relatos en el conjunto de romance, 784 en el de amistad y 155 en el de enemistad. Para equilibrar los *datasets*, reduje el conjunto de romance a 220 y el de amistad a 180, dando un total de 555 relatos para modelar estas relaciones. A este *dataset* lo llamo 'RFE dataset' (por *Romance, Friendship, Enemy*) y se utiliza en la sección 9.2.1.

9. EXTRACCIÓN DE DATOS A PARTIR DE TEXTO

Explicar los programas que hice para familiarizarme con NLTK

En la identificación de entidades, se considera una entidad a los personajes, los lugares y las instituciones, entre otras cosas, que haya sido nombrada en el texto. Un algoritmo capaz de identificar entidades nombradas tiene que poder dividir un texto en tramos y asignarle una etiqueta de entidad ("Persona", "País", etc) a cada uno. Esta tarea además requiere que las palabras del texto hayan sido previamente etiquetadas con su rol morfológico.

Por estos motivos, la librería NLTK parecía la más idónea para la tarea. Es una librería de python que contiene herramientas básicas para el análisis de texto, y en particular me interesaba que venía con un *part of speech tagger* (es decir, un identificador de rol morfológico) ya programado y entrenado. NLTK también viene con un identificador de entidades ya entrenado, pero quería programar uno que fuera más preciso y adaptado a mi conjunto de datos.

9.1. Algoritmo de identificación de entidades

9.1.1. Extracción de entidades con NLTK

Además del identificador de rol morfológico, NLTK también tiene una clase llamada *ChunkParser* cuyo trabajo es dividir un texto en tramos. Todas las funciones de la librería que se encargan de dividir y/o etiquetar texto (como el identificador de rol morfológico) heredan de alguna versión de la clase *ChunkParser*, de modo que la idea para el algoritmo era modificar la clase *ChunkParserI* para convertirla en un identificador de secuencias basado en características. El código utilizado en este proyecto está basado en el tutorial de Ivanov en *Natural Language Processing for Hackers* [\[iva\]](#).

Un identificador de secuencias basado en características trata de asignar un peso a un tramo concreto, y según el peso, le asigna una etiqueta u otra. Este peso se calcula como una función de las características del propio tramo, así como de los tramos que le preceden. El programador puede elegir las características que considere más importantes, pero hay algunas que son bien conocidas como las más importantes para reconocer entidades, como:

- El rol morfológico de la palabra actual, las anteriores y las siguientes.
- La forma de la palabra, las anteriores y las siguientes (si empiezan por mayúscula, si tienen signos de puntuación, si son siglas, etc)
- Los prefijos y/o sufijos de la palabra actual, las anteriores y las siguientes.
- Si la palabra anterior ha sido identificada como una entidad o no.

El conjunto de características de cada tramo se llama vector de características, y se utiliza para calcular

un "peso" que se corresponde con la probabilidad de que un tramo X con un vector de características V tenga una etiqueta Y . El algoritmo al final asigna a cada tramo la etiqueta cuyo peso sea el más alto.

El cómo se calcula exactamente ese peso depende del modelo matemático a utilizar. A la versión modificada de `ChunkParserI` para la identificación de entidades la llamo *NERChunker* (NER por Named Entity Recognition), y tiene tres versiones:

- *NERChunkerv1* y *NERChunkerv3* utilizan un modelo de regresión logística (también llamado modelo de entropía máxima), a través de la clase `MaxentClassifier` de NLTK. Para que NLTK pueda utilizar esta clase correctamente, es necesario tener instalado el módulo Megam para python, que no viene incluido en NLTK. La única diferencia entre la versión 1 y la 3 de este chunker es que la 3 maneja las estructuras de NLTK para oraciones y etiquetas de forma ligeramente más rápida.
- *NERChunkerv2*, que utiliza un modelo de naïve Bayes a través de la clase `ClassifierBasedTagger` de NLTK.

Las versiones *v1* y *v3* de *NERChunker* obtuvieron los mejores resultados en la evaluación, y la *v3* es algo más rápida, por lo que es la versión definitiva del identificador de entidades. Todas estas versiones, junto con sus funciones auxiliares, se encuentran encapsuladas en el archivo *NERChunkers.py*, para ser utilizadas donde se las necesite.

Puesto que tanto los clasificadores de regresión logística como los de naïve Bayes son algoritmos de aprendizaje supervisado, antes de poder utilizar (o evaluar) cualquiera de las versiones de *NERChunker* era necesario entrenarlas con un conjunto de datos ya etiquetados. El problema aquí es que NLTK, a pesar de incluir un corpus muy extenso en la propia librería, sólo tiene dos conjuntos de datos para identificación de entidades: uno en español y el otro en holandés. Todos los textos a analizar en el proyecto están en inglés, obligándome a buscar un conjunto ajeno a NLTK y finalmente decidiéndome por *Groningen Meaning Bank* (GMB). GMB es un *dataset* para identificación de entidades específicamente en inglés, grande, con una gran variedad de etiquetas de entidad y, sobretodo, con un formato de etiquetado sencillo de entender, cosa importante puesto que al ser ajeno a NLTK, GMB utiliza etiquetas distintas que son necesario adaptar para que *MaxentClassifier* pueda trabajar con ellas.

GMB utiliza la notación IOB para etiquetar entidades, y separa cada palabra de la siguiente por un carácter de nueva línea, y cada frase, por dos. De modo que la frase "*Mr. Blair left for Turkey Friday from Brussels.*" en GMB tendrá el aspecto de la figura 10.

Cuando el programa detecta una entidad de tipo persona, etiqueta como 'B-PER' la primera palabra de la secuencia, mientras que el resto de palabras dentro de la secuencia son etiquetadas como 'I-PER'. Similarmente, si la entidad es de tipo geográfico las etiquetas usadas serán 'B-GEO' y 'I-GEO', si es de

Mr.	NNP	B-PER
Blair	NNP	B-PER
left	VBD	O
for	IN	O
Turkey	NNP	B-GEO
Friday	NNP	B-TIM
from	IN	I-TIM
Brussels	NNP	I-TIM
.	.	O

Figura 10: Frase etiquetada por GMB. De izquierda a derecha, las columnas representan la palabra a etiquetar, la etiqueta de rol morfológico, y la etiqueta IOB.

tiempo serán 'B-TIM' y 'I-TIM', etc. Si una palabra no forma parte de ninguna secuencia de entidad, se etiqueta como 'O'.

NLTK, por su parte, no utiliza la notación IOB ni caracteres de nueva línea, sino que utiliza una estructura de datos propia de tipo árbol que encapsula cada palabra y cada tramo con su etiqueta. La misma frase etiquetada por NTLK tiene el aspecto de la figura 11.

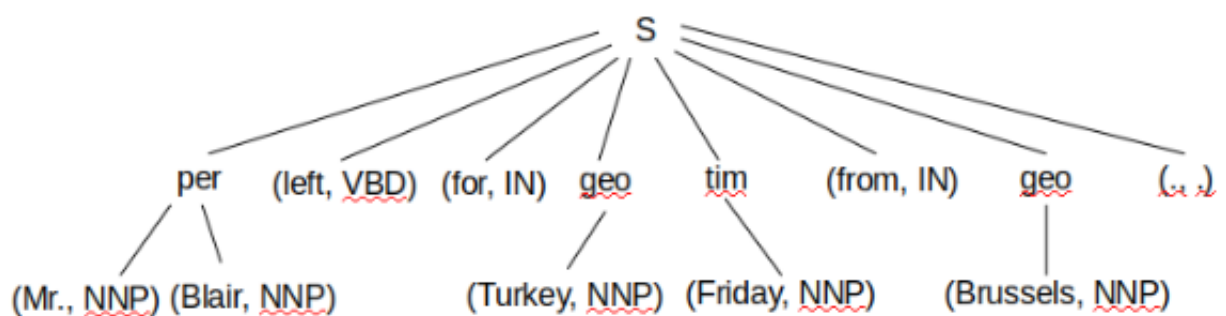


Figura 11: Frase etiquetada por NLTK. Las etiquetas de entidad se encuentran en los nodos, encontrándose todas las palabras pertenecientes a una secuencia de entidad en la profundidad 2 del árbol. Las palabras que no pertenecen a ninguna secuencia de entidad se encuentran en la profundidad 1. Cada hoja del árbol contiene una tupla formada por la palabra y su etiqueta de rol morfológico.

Como se ve, en vez de usar etiquetas IOB, NLTK organiza las palabras y su etiquetas en una estructura

```

maria@maria-P7815:~/Documents/Fanfic_ontology$ python3 NER_trainer.py
Starting training... go for a walk
optimizing with lambda = 0
NER chunker ( 126.00397671063742 mins)
  ChunkParse score:
    IOB Accuracy: 96.7%%
    Precision:    83.4%%
    Recall:       81.6%%
    F-Measure:    82.5%%
Preparing to pickle. . .
NER_chunker successfully pickled

```

Figura 12: Ejecución final de *NER_trainer.py*, mostrando su evaluación.

de árbol. La raíz, S, indica el inicio de la frase (Sentence), y las etiquetas de entidad son nodos.

En horizontal queda así:

```

(S, [(per, [(('Mr.', NNP), ('Blair', NNP))], ('left', VBD), ('for', IN), (geo, [(('Turkey', NNP))], (tim, [(('Fri-
day', NNP))], ('from', IN), (geo, [(('Brussels', NNP))], ('.', .))])

```

Fue más o menos a estas alturas del proyecto cuando decidí separar el proceso de entrenar el identificador de entidades y el de utilizarlo para etiquetar texto nuevo en dos programas distintos (NERTrainer y NERTagger, respectivamente). Acceder a los textos de GMB y transformar sus etiquetas a un formato que NLTK pueda entender y acceder a los textos de la base de datos de fanfics y preprocesarlos para su posterior etiquetado mediante el programa ya entrenado han resultado ser dos procesos muy distintos, y dividirlo parecía la mejor manera de tener un código limpio y claro.

Por lo tanto, el programa *NER_trainer.py* se encarga únicamente de entrenar el *chunker* y guardarlo en un objeto binario con la ayuda de *pickle*, mientras que el programa *NER_tagger.py* carga el objeto binario y lo encapsula en una clase NERTagger. Cuando otro programa quiere usar NERTagger, sólo tiene que importarlo y usar su función *parse(text)*, que requiere el texto completo a analizar, previamente etiquetado con el rol morfológico de cada palabra. En este aspecto, funciona de forma parecida al resto de *taggers* de NLTK.

Sin embargo, al contrario que los modelos de NLTK, NERTagger no devuelve el texto entero con los personajes etiquetados en objetos tipo árbol, sino que devuelve directamente la lista con los nombres de los personajes y el número de veces que es mencionado cada uno. Además, tras la extracción de personajes el programa realiza una canonicalización de personajes.

Vamos a aprovechar el hecho de que los relatos que estamos manejando pertenecen al género fanfic, es decir, son obras basadas en obras ya existentes. Eso quiere decir que hay una alta probabilidad de que la mayoría o incluso todos los personajes del fanfic no sean creación del autor, sino que ya aparecían en la

obra original. A estos personajes se les llama 'personajes canon'.

En este proyecto, la canonicalización de personajes consiste en descubrir qué personajes del fanfic son canon o no. Para ello necesito la lista de personajes que sí que son canon, de modo que, con la ayuda de [la página de personajes de la wiki de *Good Omens*](#)¹, creo un archivo llamado *canon_characters.csv* que sirve como base de datos de los personajes de la obra original, asignándole a cada uno un identificador numérico. La base de datos además también contiene el género canónico de cada personaje, y una lista con sus apodos.

Usando la base de datos es sencillo comprobar si un personaje pertenece al canon simplemente observando si su nombre o parte de su nombre coincide con el nombre o algún apodo de un personaje canon. Para no pasar por alto posibles erratas cometidas por los autores al escribir, se considera que dos nombres coinciden cuando la distancia de edición [Lev66] entre ellos es menor que una distancia máxima, cuyo valor se depende de lo largos que sean los nombres. Por ejemplo, un candidato a personaje llamado 'Azriaphale' será enlazado con el personaje canon 'Aziraphale', mientras que para que un candidato sea enlazado con 'War' o 'God' la distancia de edición tendrá que ser 0, puesto que son nombres muy cortos e incluso una distancia de edición de 1 añadiría ruido como 'dog', 'Got', 'warm' o 'ware'. Nombres que tengan varias palabras, como 'Mr. Anderson' y 'Jane Austen', se comparan palabra a palabra con los nombres de los personajes canon, y se escoge la menor distancia de edición.

El resultado final es un diccionario que incluye el nombre del personaje, el identificador numérico de su personaje canon (si lo tiene) y su número de menciones. Un ejemplo se puede observar en la figura 13.

9.1.2. Extracción de entidades con CoreNLP

La decisión de incluir CoreNLP en el proyecto está motivada por la posibilidad de no sólo aprovechar sus capacidades para identificar menciones a una entidad en un texto, y, sobretodo, su función de resolución de correferencia para identificar relaciones entre personajes en un texto. En la sección 9.2.2 está explicado en mayor detalle por qué esto podría ser relevante para el proyecto y cómo utilizarlo en python, pero aquí baste con mencionar que la resolución de correferencia es un método por el cual se enlaza un pronombre con el nombre propio al que se refiere. Por ejemplo, en una frase como '*I love you, Juliet*' se podría aplicar correferencia para enlazar el *you* con *Juliet*. En este ejemplo, *you* es una mención pronominal, mientras que *Juliet* es una mención propia.

Esto hace que CoreNLP sea una un complemento atractivo al NERtagger desarrollado en la sección

¹<https://goodomens.fandom.com/wiki/Category:Characters>

```

- DATA FOR FANFIC #9
Title: I Say Unto You, Love Your Enemies
NERTagger characters: CanonID Name
NO Lord Beezlebub
NO Lilith
14 Did Gabriel
NO Ashmedai
4 #endAziraphale
#endNOIF len Laradiri
if NO (all_coref_mentions) > 0:
NO for Maruton in all_coref_mentions:
NO Joe if mention.mentionType in ['PROPER', 'P
NO Amides sentence = {}
NO Dr Dudders sen = sentences[mention.sentNu
NO History
14 Gabriel string_sen = ''
NO Butter for token in sen.tokens:
NO Alisha string_sen += ' '+token
4 Aziraphale never
NO Mrs Beeton sentence['senID'] = sen.senID
15 God sentence['Sentiment'] = sen.sentiment
9 Dagon #sentence['Verbs'] = get_lemma(sentence['Verbs'])
NO Any sentence['Verbs'] = string_sen
NO So Below sentence['nerIDs'] = -1 #fill
NO Seamus Blackley sentence['clusters'] = mention
14 Sir Yes Sir Gabriel Sir
10 Death character_sentences.append(sen

```

Figura 13

anterior, ya que además del nombre del personaje también recoge información de sus pronombres, posibilitando identificar su género y llevar una cuenta más precisa de cuántas veces se menciona un personaje concreto en el texto (incluso aunque la mención sea sólo un pronombre como *he*). En esta sección se explica cómo utilizar todas estas funciones, además de los metadatos a nuestra disposición, para identificar y extraer el nombre y el género de los personajes presentes en un texto. Además, puesto que la intención del programa es que el fanfic se pueda comparar con el texto original en el que se basa, por cada personaje se identificará si aparece en la obra original o es un añadido del autor fan.

La idea básica para la extracción de personajes es buscar tanto las menciones de entidad como las de coreferencia (en particular, las menciones pronominales y propias). Como se explica en su [documentación](#)², CoreNLP tiene dos tipos de menciones:

- *NERMention*, para las menciones relacionadas con identificación de entidades. Hay varios tipos, pero este proyecto se utiliza sobretodo las menciones de tipo 'PERSON'. Cada mención tiene un `entityMentionIndex` que la identifica de forma única, y además también tiene un `canonicalEntityMentionIndex` que identifica a la entidad particular a la que hace referencia (de modo que si una entidad se llama John Smith, todas las menciones que contienen John irían indexadas a una única *NERMention*).
- *Mention*, para las menciones de coreferencia. En este proyecto se utilizan sobretodo las de tipo 'PROPER' y 'PRONOMINAL', ya que son las que identifican nombres y pronombres. Cada una tiene un `mentionID` que la identifica de forma única, además de un `corefClusterID` y un `golden`

²<https://github.com/stanfordnlp/stanza/blob/master/doc/CoreNLP.proto>

CorefClusterID que indican los clusters a los que pertenecen. Una vez procesado un texto, CoreNLP devuelve un objeto *Document* que contiene todos los NERMention y Mention detectados en el texto. Es sencillo hallar los personajes simplemente creando una lista que contenga todas las NERMention con un mismo canonicalEntityIndex, pero todas estas menciones contienen el nombre o parte del nombre de un personaje, y mi intención era hallar también los pronombres utilizados para referirse a un personaje. Ahí es donde entra la función de resolución de correferencia, cuyo resultado se almacena en las Mention: una mención de tipo PROPER contiene el nombre de un personaje, mientras que las de tipo PRONOMINAL contienen algún pronombre. CoreNLP organiza todas estas menciones en clusters, de modo que una mención PROPER y una mención PRONOMINAL comparten el mismo cluster si se refieren a la misma entidad. Para decidir en qué cluster debe ir una mención, CoreNLP tiene en cuenta factores como el género identificado de la entidad, la distancia entre menciones y cuál fue la última entidad mencionada.

La estrategia más evidente es hacer una lista con todas las menciones PROPER y PRONOMINAL, identificar sus clusters y asociarlos a las entidades de las NERMentions, de modo que por cada texto se tenga una lista de personajes únicos con su identificador, su nombre y su género.

Sin embargo, antes incluso de empezar a entender cómo se relacionan las Mention y NERMention con el texto, hay que tratar el problema de la latencia de CoreNLP, y es que es un servidor que realmente no está preparado para procesar grandes cantidades de información. Mis primeros programas manejando CoreNLP podían tardar alrededor de un minuto por texto, lo cual no es un problema terrible para procesar uno o dos textos, pero puesto que la intención inicial de utilizar CoreNLP era analizar un dataset de casi 400 textos (sección 9.2.2) me vi obligada a buscar una forma eficiente de realizar las peticiones. Además, muchos de los textos exceden el límite de 100000 caracteres del servidor, lo cual hace que la petición expire y el servidor se cierre, terminando el programa. Aunque es posible aumentar dicho límite con los parámetros de configuración del servidor, la mejora en rapidez es insuficiente.

El límite de caracteres tiene fácil solución, puesto que basta con enviar cada fanfic como una lista de capítulos (dividiendo aquellos que aún sigan excediendo el límite), y cada capítulo se envía en una petición separada. Obviamente eso significa que por cada texto se pueden recibir dos o más objetos *Document*, lo que significa que un mismo personaje puede tener distintos identificadores según el *Document* en el que se generó su mención, lo que complica ligeramente el proceso de consolidación de personajes, como se explica más adelante. Sin embargo, esta división del texto en distintas peticiones evita con éxito que expiren por ser demasiado grandes, pero la respuesta sigue siendo muy lenta para listas de más de 9 ó 10 textos, dependiendo de lo largos que sea cada uno. Finalmente la solución fue rediseñar el programa de manera que en vez de abrir y cerrar el servidor por cada texto, se abre una vez y todas las peticiones se envían juntas. Aunque el tiempo que se tarda en abrir el servidor casi siempre es menor que el necesario

para procesar el texto en sí, era obviamente la manera más simple de ahorrar tiempo de ejecución.

Para facilitar todo este proceso se crea la clase `CoreWrapper`, que se encarga de todo lo relacionado con la comprobación del límite de caracteres y manejar el servidor. `CoreWrapper` simplemente recibe una lista de objetos `Fanfic` y devuelve esencialmente la misma lista, pero ahora cada `Fanfic` tiene un atributo *annotations* que es una lista de los *Document* asociados a él. `CoreWrapper` también maneja los errores de servidor y de red que puedan ocurrir durante la ejecución de CoreNLP, avisando si uno se produce, y asegurando que los datos obtenidos hasta el momento sean almacenados. Ésta función resultó muy importante en el procesado del RFE dataset en la sección 9.2.2, puesto que CoreNLP es bastante propenso a errores de red y rara vez podía digerir más de 25 fanfics de golpe.

Una vez solucionado el problema de la latencia, el siguiente reto es entender cómo CoreNLP indexa cada palabra u oración de un texto con sus correspondientes menciones, y cómo éstas se refieren unas a otras. Repasé la documentación de CoreNLP, además de dibujar esquemas y crear varios programas para encontrar el mejor método de agrupar `Mention` en clusters y `NERMention` en listas que contuvieran toda la información necesaria para identificar cada personaje. Tras toda esta experimentación, todas las menciones y su información quedaron resumidas en listas de diccionarios de python, de forma que cada diccionario representa una mención e incluye el texto de la mención (que generalmente es el nombre del personaje), el género del personaje, y otra información como el tipo de mención y sus identificadores (de entidad, de cluster, etc).

Puesto que las menciones de correferencia pertenecen a clusters de menciones, mientras que las menciones de entidad tienen un identificador, en un principio pensé en clasificar todas las menciones de correferencia según cluster, determinar qué personaje representa qué cluster y luego asociar cada uno de estos clusters con un los identificadores de entidad, teniendo en cuenta factores como el género y el nombre de cada uno. Sin embargo, los clusters de correferencia no se corresponden fácilmente con personajes, especialmente si dos personajes del mismo género aparecen mencionados juntos (cosa a la que éste conjunto de relatos es muy susceptible). Por tanto, para consolidar las menciones de correferencia en personajes tuve aprovechar otros patrones en la información proporcionada por CoreNLP:

1. A veces las menciones de entidad y las de correferencia coinciden en una misma palabra. En particular, las menciones de tipo `'PROPER'` (que corresponden con sustantivos y nombres propios) como mínimo a veces también serán una mención de entidad.
2. Las menciones de correferencia tienen un atributo *headString*, que es la palabra que CoreNLP identifica como la más importante en la mención, y que suele corresponderse con el nombre propio del personaje (si una mención es `'Mr. Smith'`, por ejemplo, CoreNLP identifica `'Smith'` como el

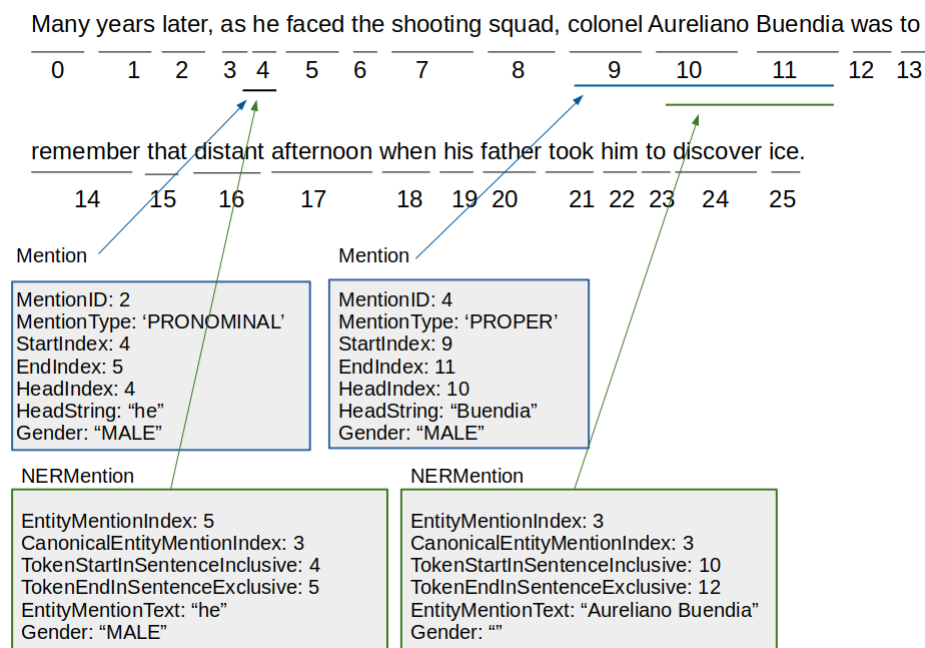


Figura 14: Solapamiento entre menciones de correferencia (Mention) y de entidad (NERMention).

headString de la mención).

3. Cuantas más menciones tenga un personaje, más probable es que dicho personaje sea un personaje real en el texto, y no un error de identificación.

En base a estas observaciones se pueden determinar algunas reglas para decidir qué menciones se refieren a qué personaje. Como en el proceso de canonicalización explicado en la sección 9.1.2, se considera que dos nombres coinciden cuando su distancia de edición [Lev66] es menor que una distancia máxima, que se fija según lo largo que sea el nombre. Así, tenemos:

1. Todas las NERMention que tengan el mismo *canonicalEntityMentionIndex* se consideran como el mismo personaje, y el atributo *entityMentionText* se considera el nombre del mismo.
2. Identificar las Mention que también son menciones de entidad. Dichas Mention se consideran como el mismo personaje que el de la NERMention si sus nombres coinciden.
3. De las Mention que hayan podido ser identificadas como pertenecientes a un personaje, obtener su cluster de correferencia y considerar a todas las Mention de dicho cluster como pertenecientes a dicho personaje, pero sólo si 1) El atributo *gender* de ambas menciones son el mismo, y 2) El atributo *headString* de ambas Mention coinciden.

El resultado de este proceso es una lista de diccionarios de python que representa a cada personaje,

conteniendo su nombre, su género, el número de veces que es mencionado, a qué clusters pertenece, etc.

Esta lista aún es bastante imperfecta. De entrada, los identificadores de entidad y de cluster asignados a cada mención dependen del *Document*, y dada la estructura del programa, la mayoría de relatos van a tener asociados dos o más *Document*, lo que significa que hay personajes con exactamente el mismo nombre y género que aparecen como dos personajes distintos, porque no comparten el mismo *canonicalEntityMentionIndex* ni ningún cluster de correferencia. Además, esta forma de identificar personajes significa que si alguno tiene un apodo, o si a un personaje se le llama de forma consistente por su apellido en una zona del texto y por su nombre en otro, aparecerá como dos personajes distintos. Para mitigar estos errores, utilicé el proceso de canonicalización de personajes de la sección 9.1.1. También vamos a aprovechar el proceso de canonicalización para determinar el género de un personaje de forma definitiva, y para ello vamos a aprovechar nuevamente que estamos manejando fanfics y que, por lo general, los autores etiquetan sus fanfics. No todos ellos etiquetan el género de sus personajes, pero algunos sí, ya que hay personas a las que le gusta explorar la personalidad o sexualidad de los personajes mediante técnicas como cambiarles el género o darles una expresión ambigua. Algunos ejemplos de las etiquetas que se suelen usar para indicar el género de un personaje son 'He/Him Pronouns for Crowley', 'Androgynous Crowley' o 'Female!Crowley'.

Teniendo en cuenta todos estos factores, la canonicalización de un personaje consiste en estos pasos:

1. Identificar si es canon o no, utilizando los nombres de cada candidato y comparándolos con los nombres y apodos de los personajes canon, igual que en la sección 9.1.1.
2. Identificar el género del personaje:
 - a) Obtener las etiquetas su fanfic y comprobar si hay alguna etiqueta que indique el género del personaje. Si la hay, el personaje se considera de ese género.
 - b) Si no hay ninguna etiqueta, nos quedamos con el género que CoreNLP le haya asignado.
 - c) Si CoreNLP no le ha asignado ningún género ('UNKNOWN', o simplemente el *string* vacío ''), le asignamos el género del personaje canon. Por tanto, si el personaje no es canon, su género seguirá siendo desconocido.

Este proceso resuelve los problemas relacionados con tener los mismos personajes procesados en distintos *Document*, y asegura que una mención será identificada con el personaje canon tanto si menciona el nombre, el apellido o sólo un apodo. Discrimina por género de tal manera que si se le asignan géneros distintos al mismo personaje, se cuentan las menciones de cada uno de modo que al final cada personaje tiene una cuenta de menciones masculinas, femeninas, neutras o desconocidas. De esta manera es más

probable identificar correctamente a un personaje con su correspondiente en el canon aunque el género no concuerde.

Todos este proceso de extracción de personajes se lleva a cabo mediante una clase `CoreNLPPDataProcessor`, que se encuentra encapsulada junto a `CoreWrapper` en un programa llamado `corenlp_util.py`. `CoreNLPPDataProcessor`, además de extraer personajes, también tiene función de análisis de sentimiento, como se explicará en la sección 10.

En la figura 15 hay un esquema que explica de forma visual todo este proceso de extracción de personajes.

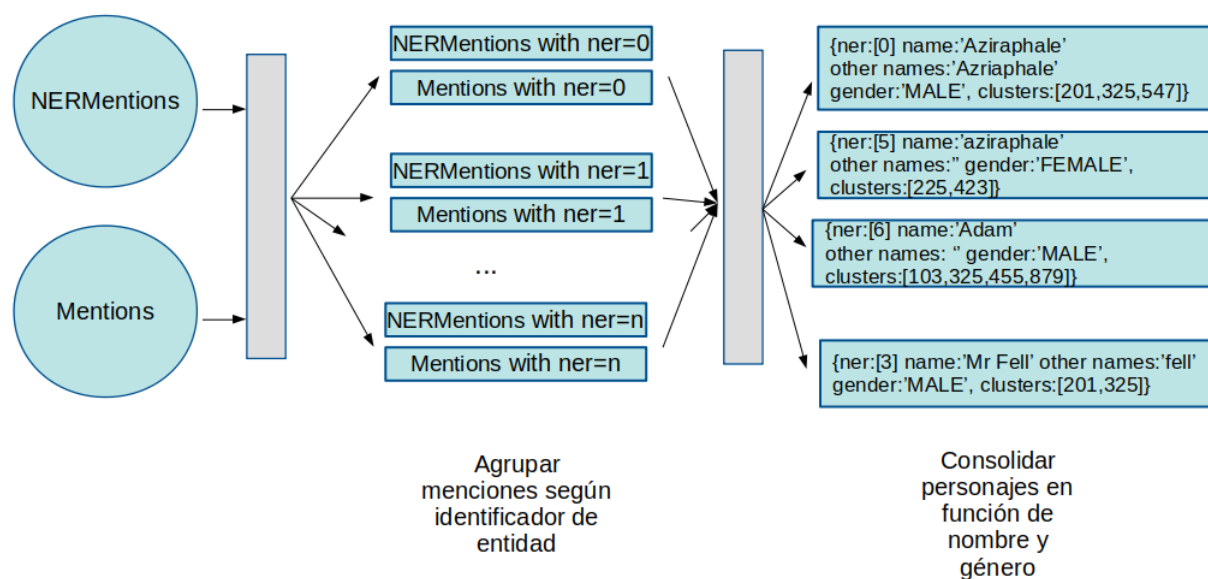


Figura 15: Esquema sobre el proceso de extracción de personajes con CoreNLP.

9.2. Algoritmo de identificación de relaciones

Las relaciones que definen un fanfic son las amistades, los enemigos y, principalmente, los amantes. Extraer relaciones a partir de texto natural es una tarea compleja de por sí, y tratar de detectar este tipo concreto de relaciones en literatura de género puede presentar un reto mayor, puesto que las relaciones sentimentales tienden a representarse de forma implícita, de modo que el lector aprende quiénes son amigos y quiénes enemigos a través de las acciones de los personajes. Además, la ambigüedad, la heterogeneidad y la experimentación son partes naturales de cualquier proceso creativo, por lo que un conjunto de obras no representará una misma relación de forma uniforme, incluso si es entre los mismos personajes.

Encontrar una estrategia para abordar este problema requiere exploración y creatividad, y ya que había

empezado el proyecto con NLTK, me pareció natural comenzar la búsqueda por ahí.

El extractor de relaciones de NLTK funciona mediante reglas: después de extraer las entidades nombradas del texto, se puede utilizar el módulo *relextract* para dividir el texto en listas de fragmentos del texto que contienen dichas entidades, y aplicar reglas basadas en expresiones regulares que definan la relación entre las entidades. La regla puede incluir etiquetas de rol morfológico en la expresión regular, y *relextract* permite filtrar por etiqueta IOB, lo que le da algo más de flexibilidad.

Por ejemplo, para extraer una relación de lugar entre una organización y una localización, se puede crear una expresión regular que busque la palabra clave 'in' en el texto, e indicarle a *relextract* que sólo te interesan los fragmentos de texto que tengan una entidad de tipo 'ORG' seguida de una entidad de tipo 'LOC':

```
1 IN = re.compile(r'.*\bin\b(?:\b.+ing\b)')
2
3 for doc in parsed_docs:
4     for rel in nltk.sem.extract_rels('ORG', 'LOC', doc, pattern=IN):
5         print(nltk.sem.show_raw_rtuple(rel))
```

Listado 1: Ejemplo de código que utiliza el módulo *regex* de NLTK para extraer relaciones de lugar y mostrarlas por pantalla. Adaptado del capítulo 7 de Natural Language Processing with Python [Bir12]

Existen proyectos que utilizan este módulo de NLTK para extraer relaciones como *DateOfBirth* y *Has-Parent* [jos17], pero es evidente que es un método poco adecuado para el tipo de proyecto que estaba intentando hacer.

Estos programas basados en reglas dependen de localizar palabras claves en el texto, y aunque existen palabras clave para identificar relaciones sociales ("love", "kiss", "hug", "friend", "kill", "hate", etc), lo cierto es que la naturaleza de la expresión literaria hace que este método, incluso a simple vista, parezca bastante ingenuo. No sólo es perfectamente posible expresar amor, amistad y odio sin usar "palabras clave".³sociadas con dichos sentimientos, sino que en un texto literario raramente se escribe explícitamente *Romeo loved Juliette*, si no que es más normal encontrar estructuras como '*I love you*', *said Romeo*. En una frase así, no se menciona explícitamente a Julieta, pero un lector humano sabe si se refiere a ella por el contexto de la escena. Pero un programa que únicamente se preocupa de las etiquetas IOB de una frase no será capaz de unir ese *you* con Julieta (ni, ya puestos, el *I* con Romeo).

Descartado el extractor de relaciones de NLTK, empiezo a buscar opciones en otras librerías. El Stanford Natural Language Processing Group publicó un extractor de relaciones como parte de las funciones

de CoreNLP, pero las relaciones que está entrenado para detectar (*Live_In*, *Located_In*, *OrgBased_In*, *Work_For*, *None*) no parecen útiles para el proyecto. Por tanto, entrenar mi propio modelo para relaciones sociales parece la única solución.

Crear un modelo de regresión logística con NLTK, similar al identificador de entidades, requería que el texto ya estuviera etiquetado con las relaciones. Los autores de AO3 usan etiquetas que es posible extraer las relaciones a partir del archivo HTML de cada fanfic, pero es una etiqueta a nivel del texto completo, no a nivel de frase, que es como trabaja NLTK. Dejando de lado NLTK por el momento, decidí explorar soluciones usando clustering y modelado de temas.

9.2.1. Primeras estrategias: Clustering y LDA

Decidir si dos personajes son amigos, enemigos o amantes (sin tener ninguna información previa sobre la obra) puede requerir leer el texto completo, con lo cual es razonable utilizarlo para el análisis y asignar una etiqueta de 'romance', 'amistad' o 'enemistad' al texto en su conjunto, más que etiquetar ciertas palabras y personajes del mismo.

Por tanto, dado un conjunto de textos al azar, podría llegarse a la conclusión de cuáles contienen romance, cuáles amistad y cuáles enemistad observando si hay similitudes entre ellos. Con este enfoque, parece una tarea adecuada para un algoritmo de clustering.

Para comprobar cómo de útil sería esta estrategia utilicé el RFE dataset, cuya creación está explicada en la sección 8.2. Esperaba que teniendo tres conjuntos claros en los que el tema era evidente sirviese para ver cómo de eficaz es el clustering para la tarea general, que sería poder decir si hay romance en un texto incluso si no es el tema central del mismo.

Una vez creados los conjuntos, utilicé la librería *Scikit-Learn* en conjunto con NLTK para crear el programa de clustering. Para preprocesar el texto se utilizan los métodos de NLTK para *tokenizar* y *lemmatizar* los textos, además de crear un conjunto de *stopwords*, palabras comunes en inglés pero que no aportan mucha información sobre el mismo (preposiciones, pronombres, puntuación, demostrativos, etc). Después de preprocesar el texto se procede a extraer las características relevantes del mismo, para lo cual se utiliza el módulo *TfidfVectorizer* de *Scikit-Learn*. Su trabajo es 'vectorizar' el texto de manera que sus características principales queden expresadas en un formato que el algoritmo de clustering pueda entender, cosa que hace asignando un peso a cada palabra dependiendo del número de ocurrencias de la misma (esto se llama *bag of words*). Hay vectorizadores que asignan más peso cuanto más frecuencia, pero esto hace que palabras muy comunes pero con poco valor informativo roben protagonismo a palabras menos

$$\text{idf}(t) = \log \frac{1 + n}{1 + \text{df}(t)} + 1 \quad (1)$$

Figura 16: Componente idf del vectorizador Tf-Idf. t se refiere al término cuyo peso está determinando, n al número total de documentos y df a la frecuencia de t en este documento.

frecuentes pero más interesantes. El vectorizador 'Tf-Idf' (*Term frequency-Inverse document frequency*), en cambio, multiplica la frecuencia de una palabra en un documento por un componente *idf* que, como se ve en la fórmula 1, está basado en la frecuencia de ese término en todos los documentos. Los vectores resultantes se normalizan usando la norma de Euclides; más información está disponible en la guía de usuario de *Scikit-Learn* [skl].

Con los textos ya preprocesados y convertidos en vectores *tf-idf* se puede crear un modelo de clustering, en este caso utilizando el módulo *KMeans* de *Scikit-Learn*. *KMeans* es un algoritmo que crea clusters de tal forma que cada uno tenga la misma varianza, minimizando la suma de los cuadrados de las distancias entre los miembros de cada grupo (fórmula 2).

$$\sum_{i=0}^n \min_{\mu_j \in C} ||x_i - \mu_j||^2 \quad (2)$$

Figura 17: Criterio de la suma de cuadrados. n es el total de textos, con x perteneciendo a n . C es el número de clusters, con μ siendo la media de las muestras x de cada cluster. El algoritmo *KMeans* reduce esta suma todo lo posible.

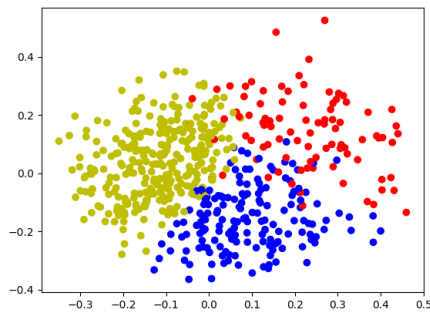
Además del código necesario para crear el modelo *KMeans* y procesar los datos, añadí código para evaluar el modelo e imprimir un diagrama de puntos con los clusters. Tras probar dos tokenizers distintos y varias combinaciones con los parámetros del vectorizador y el modelo, los resultados se pueden ver en la figura 18.

Los resultados no son muy buenos. Aunque los términos de cada cluster parecen prometedores, ninguna métrica sube del 0.1, lo que indica que las categorías del clustering son sólo un poco mejores que haberlas asignado al azar, y que hay mucho solapamiento entre clusters.

Puesto que para crear el modelo he utilizado datos filtrados a propósito para modelar cada categoría tan bien como fuera posible, con la esperanza de poder usarlo como posible semilla para un sistema más general, esto supone un gran problema.

Busqué entonces otra estrategia, utilizando un modelo de temas más que uno de clustering. El modelado de temas con el algoritmo LDA parece también una buena opción para este problema, puesto que al

³<https://stackoverflow.com/questions/57626286/how-to-plot-text-clusters>



(a) El código utilizado para crear este diagrama fue escrito por [Matt L en StackOverflow](#)³.

```
marla@marla-P7815:~/Documents/Fanfic_ontology$ python3 toy_relex_kmeans.py
Fetching fic texts...
...fics fetched. Elapsed time: 0.5578843951225281 mins
Creating stopwords and removing character names from texts...
...done in 0.032831732432647525 mins
Creating vectorizer, transforming texts to tf-idf coordinates...
...done in 3.929908561706543 mins
Creating the K-Means model and fitting data...
...done in 0.0027207175890604654 mins
n_samples: 555, n_features: 3

=== Metrics ===
Homogeneity: 0.140
Completeness: 0.150
V-measure: 0.145
Adjusted Rand-Index: 0.115
Silhouette Coefficient: 0.233

Top terms per cluster:
Cluster 0:
kiss, skin, bed, tongue, press, neck, touch, lean, slide, morning,
Cluster 1:
man, child, friend, boy, wine, god, bookshop, young, understand, sat,
Cluster 2:
wings, pain, water, blood, bed, heaven, floor, black, man, form,
```

Figura 18: A la derecha, ejecución de *toy_relex_kmeans*, mostrando su evaluación. La homogeneidad se cumple cuando ningún cluster contiene miembros que pertenezcan a categorías distintas en los datos reales. La completitud se satisface si todos los miembros de una de las categorías reales pertenecen al mismo cluster. A la izquierda, el diagrama creado por el programa.

contrario que el clustering clásico, LDA asigna a cada documento una distribución de temas en cada uno. Esto se ajusta a este análisis, puesto que aunque he intentado crear *datasets* 'perfectos' que traten un único tipo de relación en cada uno, lo cierto es que lo normal en un relato es que estén mezcladas.

LDA es un algoritmo que descubre temas de forma no supervisada. Trabaja bajo la asunción de que cada documento es un conjunto de temas, y que cada tema es un conjunto de palabras. Empieza asignando cada a palabra a un conjunto al azar de temas, y en cada iteración mejora la asignación.

Al igual que en el programa anterior, LDA requiere un preprocesado del texto, con lo que utilizo los dos *tokenizers* del algoritmo de clustering. *Scikit-Learn* carece de modelo LDA, por lo que utilizo el de la librería *gensim*. LDA también trata los textos como un *bag of words*, pero no es necesario vectorizarlos antes de usarlos para entrenar el modelo, que devolverá lista de palabras por tema, junto con la probabilidad de que esa palabra pertenezca a dicho tema.

Tras preprocesar los textos de forma similar a como se hizo con clustering y entrenar el modelo LDA, se observan los resultados en la figura 19a.

Los resultados son un poco decepcionantes, pues ninguno de los 10 términos más relevantes por tema tiene siquiera un 0.1 % de probabilidad de pertenecer a su tema. Tampoco es sorprende que no sean muy relevantes, pues aparecen muchos pronombres, determinantes e incluso algún número. Aprovechando el etiquetado de rol morfológico de NLTK, se retiran esas palabras y se crea un nuevo modelo, cuya ejecución está en la figura 19b.

```

maria@maria-P7815:~/Documents/Fanfic_ontology$ python3 toy_relex_lda.py tokv1
tokv1
Fetching fic texts...
...fics fetched. Elapsed time: 0.551144790649414 mins
Preprocessing fanfics and creating dictionaries...
Processing elapsed time: 1.7671716809272766 minutes
Training LDA model...
...LDA elapsed time: 2.681595873832703 minutes
LDA Coherence score: 0.23684957439734447
Topics in LDA model:
(0, '0.034*"I" + 0.017*"Crowley" + 0.014*"He" + 0.013*"Aziraphale" + 0.012*"say" + 0.007*"The" + 0.007*"know" + 0.006*"You" + 0.006*"It" + 0.006*"like"')
(1, '0.034*"Crowley" + 0.030*"Aziraphale" + 0.018*"He" + 0.015*"I" + 0.008*"The" + 0.007*"angel" + 0.007*"back" + 0.006*"It" + 0.006*"like" + 0.006*"say"')
(2, '0.015*"I" + 0.008*"He" + 0.008*"The" + 0.008*"Crowley" + 0.007*"Aziraphale" + 0.006*"one" + 0.006*"angel" + 0.005*"know" + 0.005*"It" + 0.005*"like"')
maria@maria-P7815:~/Documents/Fanfic_ontology$

```

(a) Filtrado: sólo puntuación.

```

maria@maria-P7815:~/Documents/Fanfic_ontology$ python3 toy_relex_lda.py UNITokv2
UNITokv2
Fetching fic texts...
...fics fetched. Elapsed time: 0.5545525550842285 mins
Preprocessing fanfics and creating dictionaries...
Processing elapsed time: 4.393664975961049 minutes
Training LDA model...
...LDA elapsed time: 2.5472853938738504 minutes
LDA Coherence score: 0.25293098120527674
Topics in LDA model:
(0, '0.008*"say" + 0.007*"know" + 0.006*"get" + 0.005*"Adam" + 0.005*"back" + 0.005*"Crawly" + 0.005*"look" + 0.004*"make" + 0.004*"go" + 0.003*"time"')
(1, '0.035*"Crowley" + 0.032*"Aziraphale" + 0.009*"say" + 0.007*"back" + 0.007*"angel" + 0.006*"know" + 0.005*"look" + 0.005*"eyes" + 0.004*"demon" + 0.004*"get"')
(2, '0.027*"Crowley" + 0.019*"Aziraphale" + 0.008*"say" + 0.008*"angel" + 0.007*"know" + 0.007*"demon" + 0.006*"..." + 0.006*"get" + 0.006*"look" + 0.005*"back"')
maria@maria-P7815:~/Documents/Fanfic_ontology$

```

(b) Filtrado: puntuación, pronombres, determinantes, etc.

Figura 19: Ejecución de *toy_relex_lda*, con distintos criterios de filtrado. Además se muestran los 10 términos más relevantes de cada tema, y su probabilidad de pertenecer a dicho tema.

Estos resultados, sin embargo, tampoco son muy convincentes. La puntuación de coherencia, que indica cómo de adecuado es el número de temas para los datos analizados, es 0.23 en el primer caso y 0.25 en el segundo. Es un resultado que se puede mejorar afinando los hiperparámetros de LDA, para lo cual se utiliza el programa *topic_evaluate.py*, que prueba diferentes valores para *alpha* (densidad documento-tema) y *beta* (densidad palabra-tema) del modelo LDA de *gensim*. Los resultados de su ejecución se guardan en el archivo *lda_evaluation.csv*, y en la figura 20a se puede ver que la puntuación de coherencia se puede mejorar bastante para tres temas con los hiperparámetros correctos, pero como evidencia la figura 20b, queda muy lejos del 0.42 de coherencia que se puede conseguir si se permite subir el número de temas a nueve. No parece, por tanto, que este modelo sea el más adecuado para buscar las relaciones que se buscan.

9.2.2. Correferencia con CoreNLP

Tras varias pruebas con los algoritmos de clustering y LDA, encontré *CoreNLP*, un servidor de Stanford NLP Group que realiza diversos tipos de extracción de información y análisis de lenguaje natural, entre ellos, resolución de correferencia.

Volviendo al ejemplo de Romeo y Julieta, una frase como '*I love you*', *said Romeo*, extraída de un texto más largo en la que el contexto es que Romeo está hablando con Julieta, da poca información a un algoritmo que no es capaz de entender a qué personajes se refieren los pronombres *I* y *you*. Sin embargo, si se aplicase resolución de correferencia sobre el texto, a ojos del algoritmo la frase se convertiría en

```
>>> cdf.sort_values(by=['Coherence'], inplace=True, ascending=False)
>>> cdf[:5]
```

	Validation_Set	Topics	Alpha	Beta	Coherence
48	75% dataset	3	0.9099999999999999	0.9099999999999999	0.355668
313	15% dataset	3	0.61	0.9099999999999999	0.353786
53	75% dataset	3	symmetric	0.9099999999999999	0.316054
318	15% dataset	3	0.9099999999999999	0.9099999999999999	0.313049
307	15% dataset	3	0.61	0.61	0.309767

(a) Mejores puntuaciones de coherencia para 3 temas.

```
AttributeError: 'DataFrame' object has no attribute 'sort_valuse'
>>> df.sort_values(by=['Coherence'], inplace=True, ascending=False)
>>> df[:5]
```

	Validation_Set	Topics	Alpha	Beta	Coherence
233	75% dataset	9	symmetric	0.9099999999999999	0.420302
102	75% dataset	5	0.61	0.61	0.390992
338	15% dataset	4	0.31	0.9099999999999999	0.381947
408	15% dataset	6	0.9099999999999999	0.9099999999999999	0.377929
123	75% dataset	6	0.61	0.9099999999999999	0.374808

(b) Parámetros con las mejores puntuaciones de coherencia de toda la prueba.

Figura 20: Resultados de *topic_evaluate.py*, visualizados y ordenados con la ayuda de *pandas*. Se pueden observar qué número de temas y qué hiperparámetros dan mejores puntuaciones de coherencia para el modelo LDA creado a partir del RFE dataset.

'Romeo love Juliette', said Romeo, con lo que el algoritmo entiende mucho mejor qué está pasando en esta frase y a quiénes afecta.

CoreNLP da acceso a esa posibilidad. Aunque está programado en java, cuenta con un módulo de python llamado *Stanza*, con lo que creé algunos programas para familiarizarme con su funcionamiento y la idea que quería desarrollar. El primero de estos programas, *toy_relex.py*, no hace más que buscar las frases que contengan el verbo 'to love' y al menos una entidad, elegida de antemano. El segundo programa, *toy_relex_v2.py* expande este concepto usando *CoreNLP*, utilizando tanto su función de identificación de entidades como la de solución de coreferencia para obtener las 'cadenas de coreferencia' del texto, utilizándolas para enlazar cada pronombre del texto con la entidad a la que se refiere. La idea sigue siendo seleccionar las frases que tengan el verbo 'to love', pero en vez de quedarme únicamente con aquellas en las que se nombren explícitamente a un personaje, también me quedo con aquellas en las que los pronombres formen parte de una cadena de coreferencia.

Para poder llevar a cabo todo esto, fue necesario estudiar con atención las propiedades de los objetos que utiliza *Stanza*, que por suerte están bien [documentadas](#), y hacer muchas pruebas y visualización de datos, como se puede ver en las figuras 21 y 22.

Tras analizar toda esta información, entender la indexación de los objetos de *Stanza* y ganar experiencia manejándolos, se hizo evidente que el siguiente paso sería crear un programa cuya función fuese resumir la información proporcionada por *CoreNLP* de forma que sea útil para identificar relaciones entre per-

Sentence 92	tokens 0-1	PRONOMINAL	cluster 912	text: He
Sentence 212	tokens 8-9	PROPER	cluster 912	text: Aziraphale
Sentence 272	tokens 14-15	PRONOMINAL	cluster 912	text: his
Sentence 393	tokens 6-7	PROPER	cluster 912	text: Aziraphale
Sentence 333	tokens 3-4	PRONOMINAL	cluster 912	text: He
Sentence 273	tokens 0-1	PRONOMINAL	cluster 912	text: He
Sentence 2	tokens 0-3	PROPER	cluster 912	text: Aziraphale
Sentence 243	tokens 3-4	PRONOMINAL	cluster 912	text: his
Sentence 394	tokens 0-1	PRONOMINAL	cluster 912	text: He
Sentence 303	tokens 12-13	PRONOMINAL	cluster 912	text: he
Sentence 333	tokens 15-16	PRONOMINAL	cluster 912	text: his
Sentence 62	tokens 16-17	PRONOMINAL	cluster 912	text: him
Sentence 334	tokens 0-1	PRONOMINAL	cluster 912	text: His
Sentence 213	tokens 10-11	PRONOMINAL	cluster 912	text: him
Sentence 333	tokens 18-19	PRONOMINAL	cluster 912	text: his
Sentence 304	tokens 0-1	PRONOMINAL	cluster 912	text: He
Sentence 334	tokens 4-5	PRONOMINAL	cluster 912	text: he
Sentence 333	tokens 22-23	PRONOMINAL	cluster 912	text: his
Sentence 244	tokens 0-1	PRONOMINAL	cluster 912	text: He
Sentence 33	tokens 4-5	PRONOMINAL	cluster 912	text: his
Sentence 274	tokens 3-4	PRONOMINAL	cluster 912	text: his
Sentence 304	tokens 6-9	PROPER	cluster 912	text: Aziraphale
Sentence 333	tokens 25-26	PRONOMINAL	cluster 912	text: his
Sentence 274	tokens 6-7	PRONOMINAL	cluster 912	text: him
Sentence 244	tokens 5-6	PRONOMINAL	cluster 912	text: him
Sentence 334	tokens 11-12	PRONOMINAL	cluster 912	text: his
Sentence 425	tokens 0-1	PROPER	cluster 912	text: Aziraphale
Sentence 63	tokens 14-15	PROPER	cluster 912	text: Aziraphale
Sentence 394	tokens 20-21	PRONOMINAL	cluster 912	text: he
Sentence 244	tokens 11-12	PRONOMINAL	cluster 912	text: he
Sentence 335	tokens 0-1	PROPER	cluster 912	text: Aziraphale
Sentence 274	tokens 15-16	PROPER	cluster 912	text: Aziraphale
Sentence 335	tokens 2-3	PRONOMINAL	cluster 912	text: his
Sentence 395	tokens 7-8	PRONOMINAL	cluster 912	text: he
Sentence 335	tokens 6-7	PRONOMINAL	cluster 912	text: his
Sentence 214	tokens 16-17	PRONOMINAL	cluster 912	text: his
Sentence 305	tokens 5-6	PRONOMINAL	cluster 912	text: he
Sentence 394	tokens 29-30	PRONOMINAL	cluster 912	text: his
Sentence 395	tokens 12-13	PRONOMINAL	cluster 912	text: his
Sentence 425	tokens 14-15	PRONOMINAL	cluster 912	text: he
Sentence 4	tokens 8-9	PROPER	cluster 912	text: Aziraphale
Sentence 394	tokens 34-35	PRONOMINAL	cluster 912	text: he
Sentence 396	tokens 0-1	PROPER	cluster 912	text: Aziraphale
Sentence 64	tokens 13-14	PROPER	cluster 912	text: Aziraphale
Sentence 336	tokens 0-1	PRONOMINAL	cluster 912	text: He
Sentence 65	tokens 0-1	PRONOMINAL	cluster 912	text: He
Sentence 95	tokens 2-3	PRONOMINAL	cluster 912	text: his
Sentence 306	tokens 0-1	PROPER	cluster 912	text: Aziraphale

Figura 21: Visualización de la información proporcionada por *CoreNLP*

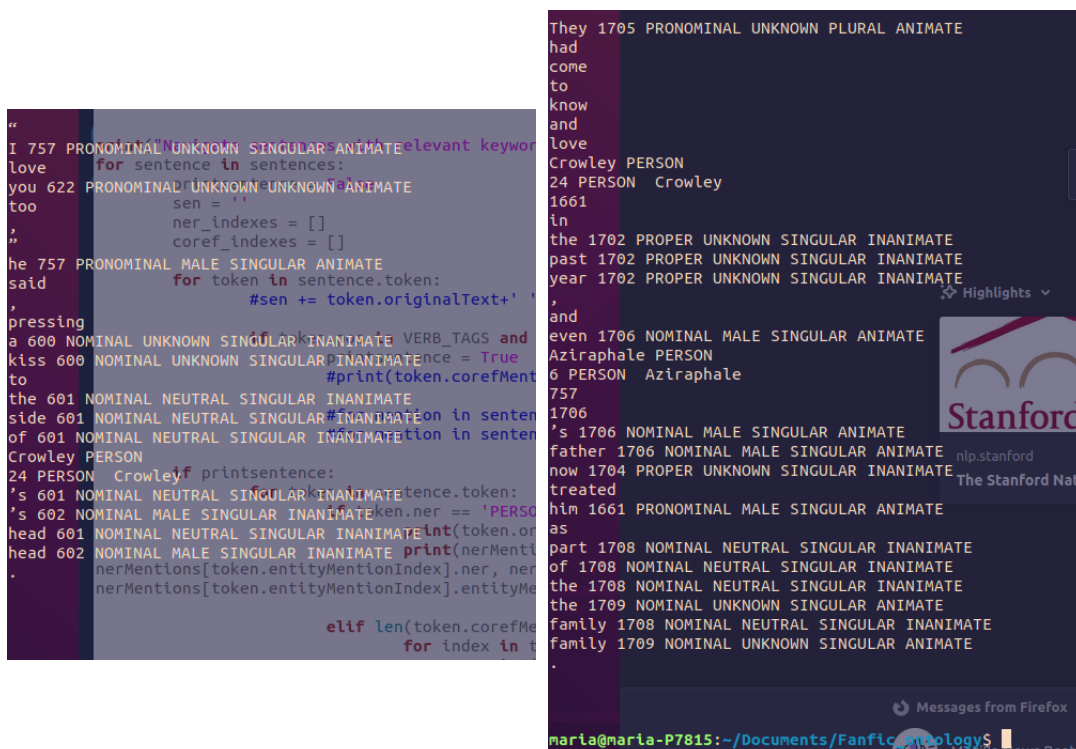


Figura 22: Visualización de frases particulares con anotaciones de correferencia de *CoreNLP*

sonajes. En particular, mi intención era usarlo con el RFE dataset y tratar de identificar relaciones en él. El problema era que el RFE dataset es bastante grande, y enseguida se hizo evidente que era necesario guardar la información proporcionada por CoreNLP en un archivo para no malgastar horas de trabajo. Ya que los objetos de *Stanza* no tienen una forma sencilla de ser almacenados directamente, acabé creando un programa que, tras recoger las respuestas del servidor, resumía la información más relevante y la guardaba en dos archivos csv:

- `fic_characters.csv` (tabla 2) almacena los personajes de cada fanfic, junto con su correspondiente en el canon y toda la información necesaria para identificar en qué frases de qué fanfic aparece (clusters, identificador, etc).
- `fic_sentences` (tabla 1) almacena los fanfics frase a frase, junto con los identificadores y clusters de los personajes mencionados en ellas. Decidí almacenar sólo las frases que mencionan personajes en vez de todas las del fanfic porque estaba orientando este archivo a la identificación de relaciones entre personajes, por lo que parecía razonable asumir que las frases con más información serían aquellas en las que se mencionan personajes.

El desarrollo de este programa no fue trivial, debido a la complejidad de la indexación de *Stanza* y otros problemas derivados de la naturaleza del proyecto, como por ejemplo, si un personaje aparece marcado como masculino en 20 menciones y femenino en 3, ¿Deberían considerarse personajes distintos? ¿Asumo que todas las menciones de un cluster se refieren a un único personaje, incluso si algunos nombres son radicalmente distintos del resto? Para resolver todas estas preguntas y alcanzar un programa funcional fui pasando por varias etapas y distintos programas, que con el tiempo refinaría y acabaría encapsulando en

ficID	Dataset	senID	Sentiment	Verbs	nerIDs	Clusters
57	ROMANCE	2	Positive	Crowley flashed a grin over his shoulder .	0	452
57	ROMANCE	5	Positive	He grabbed the long pole that served as a door handle and pulled it open .	0	452, 18
57	ROMANCE	10	Negative	He cautiously side - stepped around an especially small one .	0	33, 452
57	ROMANCE	11	Neutral	“ They ’re not going to bite you , ” Crowley laughed .	0	452, 36, 37, 113

Cuadro 1: Estructura de *fic_sentences.csv*

ficID	nerID	Name	Gender	Mentions	clusterID	canonID
57	0	ng	UNKNOWN	193	279, 452, 116, 251, 435, 448, 517, 545	-1
57	15, 80, 95	aziraphale	MALE	54	279	4
57	146, 151	crowley	MALE	23	452	8
70	0	crowley	NEUTRAL	9	1	8

Cuadro 2: Estructura de *fic_characters.csv*

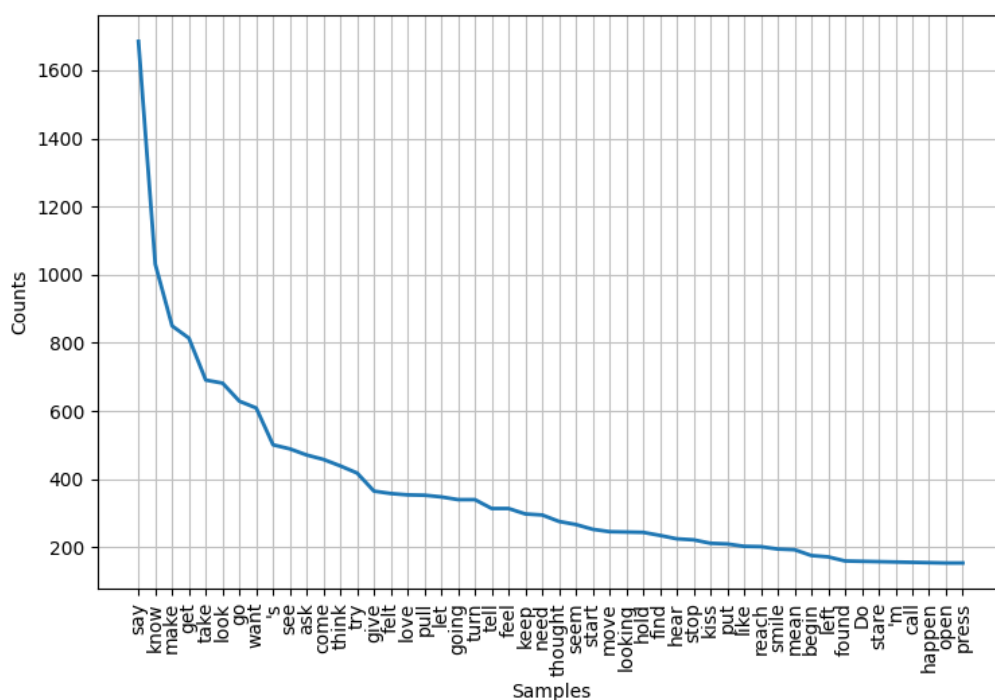
corenlp_util.py para su uso posterior en el programa final. En esta parte del proyecto sin embargo utilicé los programas de *corenlp_wrapper.py* y *ner_and_sen_extraction_v2.py*, que se pueden considerar versiones anteriores a las utilizadas en *corenlp_util.py*, y por tanto muchos de los problemas mencionados en la sección 9.1.2 relacionados con la latencia y errores de servidor fueron resueltos durante su desarrollo. Además de estos problemas, hubo que limpiar y revisar el formato de cada *string* y lista, ya que el archivo CSV estaba configurado para procesar cada punto y coma como un separador entre columnas, lo cual añadía columnas extra cada vez que encontraba un punto y coma en cualquier parte del texto.

Una vez procesado todo el RFE dataset y completados los CSV, utilicé la información almacenada ellos para realizar varias técnicas de análisis de texto natural, con la esperanza de hallar algún patrón, bigrama o palabra significativa que me ayudara a identificar relaciones entre personajes.

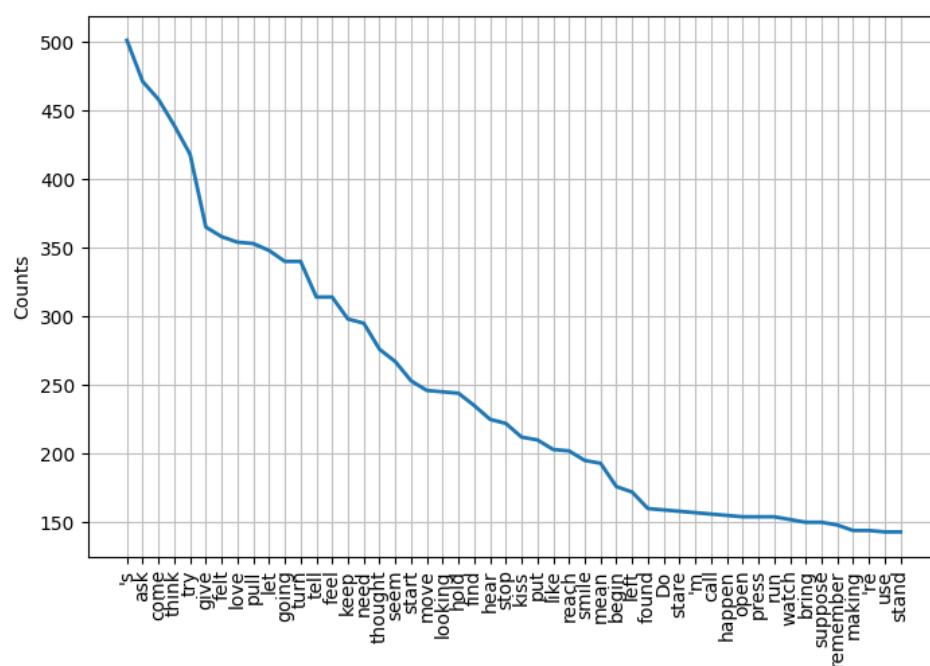
En primer lugar quise buscar si había algún verbo que fuera característico de cada dataset, por lo que extraje y lematicé los verbos de cada frase usando las herramientas de NLTK. En la primera pasada los verbos más repetidos en todos los dataset fueron verbos muy comunes en inglés, como 'say', 'go', 'get', 'make', 'look', etc. de modo que para eliminar ruido hice otro análisis eliminándolos. Para hacer más fácil la comparación y a modo de ejemplo, se muestra la distribución de frecuencias de los verbos en el dataset de romance en la figura 23a-23b, el resto de distribuciones se puede consultar en el anexo [placeholder ref].

Desafortunadamente, ni siquiera eliminar los verbos más frecuentes parecía dar un resultado claramente distintivo para cada dataset, por lo que decidí utilizar NLTK para buscar bigramas y trigramas característicos (figuras 24a-24c). Pero tampoco parece haber una alguna combinación distintiva aquí, los n-gramas siguen siendo bastante parecidos de un dataset a otro y utilizan casi las mismas palabras.

En un último intento de buscar algún patrón, decidí volver extraer la distrución de frecuencia de verbos,



(a) Verbos más frecuentes en el dataset de romance.



(b) Verbos más frecuentes en el dataset de romance, tras retirar los más comunes.

bigramas y trigramas, pero esta vez utilizando únicamente las frases en las que se mencione a dos personajes en concreto, en vez de todas las frases del relato. Estos personajes son elegidos de antemano, utilizando los identificadores propocionados por CoreNLP (tal y como se ve en la tabla 1).

```

===== Bigram collocations of Romance dataset =====

['had been', 'It was', 'Of course', 'You 're', 'The angel', 'I 'll', 'six thousand', 'I know', 'They were', 'I think', 'The demon', 'Mistress Fell', 'thousand years', 'shook head', 'dear boy', 'You know', 'Mr. Fell', 'I suppose', 'arms around', 'six years']

===== Trigram collocations of Romance dataset =====

['I 'm sorry', 'Crowley 's face', 'Crowley 's eyes', 'Crowley 's hand', 'I 'm afraid', 'Crowley 's head', 'Crowley 's neck', 'Crowley 's hair', 'I 'm going', 'think I 'm', 'Crowley 's shoulder', 'I 'm sure', 'I 'm glad', 'Crowley 's mouth', 'But I 'm', 'Crowley 's lips', 'Crowley 's chest', 'I 'm fine', 'Crowley 's cheek', 'Crowley 's voice']

```

(a) Bigramas y trigramas con mayor likelihood en el dataset de romance.

```

===== Bigram collocations of Friendship dataset =====

['had been', 'It was', 'Of course', 'Aziraphale said', 'Crowley said', 'thousand years', 'He was', 'I think', 'gon na', 'I suppose', 'six thousand', 'holy water', 'shook head', 'six years', 'There was', 'You 're', 'cleared throat', 'did want', 'The angel', 'You know']

===== Trigram collocations of Friendship dataset =====

['I 'm sorry', 'think I 'm', 'I 'm sure', 'I 'm afraid', "I 'm sorry", 'Crowley had been', 'I 'm going', 'But I 'm', "think I 'm", 'I 'm saying', 'I 'm glad', 'know I 'm', 'I 'm getting', 'I 'm warning', 'I 'm supposed', 'thinks I 'm', 'All I 'm', 'What I 'm', 'He had been', "I 'm afraid"]

```

(b) Bigramas y trigramas con mayor likelihood en el dataset de amistad.

```

===== Bigram collocations of Enemy dataset =====

['had been', '- -', 'It was', 'Mr. Fell', 'Port Talbot', '° •', '° °', '• °', '• °', 'Az gon', '## Chapter', 'Chapter 1', '## 1', 'The Teacher', 'I think', 'Of course', 'His eyes', 'The angel', 'At least', 'I 'll']

===== Trigram collocations of Enemy dataset =====

["I 'm sorry", "And I 'm", "I 'm going", "I 'm afraid", "But I 'm", '° • °', '• ° •', "? I 'm", "I 'm asking", "I 'm trying", 'I 'm sorry', 'Crowley 's voice', 'And I 'm', 'He had been', 'Crowley had been', 'I 'm', 'think I 'm', 'Aziraphale had been', 'Crowley 's hand', 'I 'm afraid']

```

(c) Bigramas y trigramas con mayor likelihood en el dataset de enemistad.

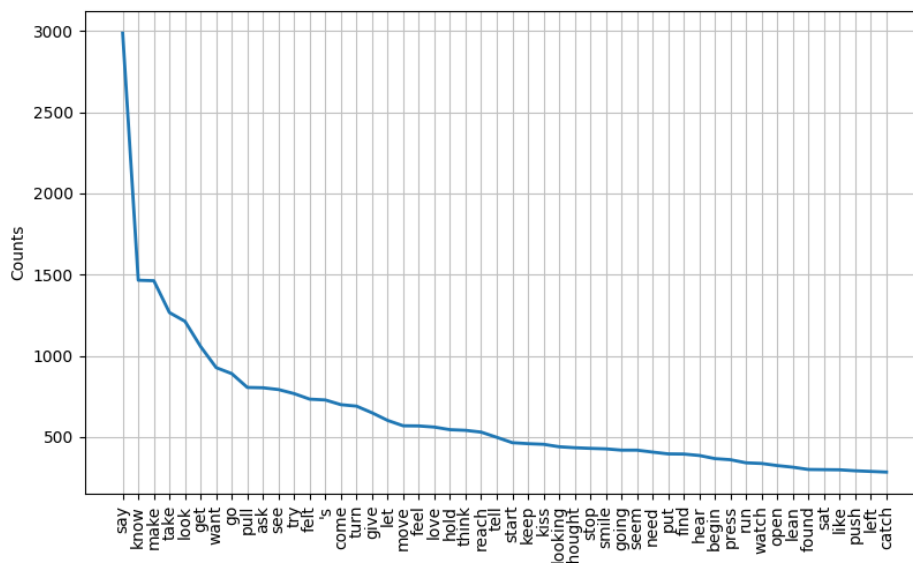


Figura 25: Verbos más frecuentes en el dataset de romance en las frases que mencionan a dos personajes concretos.

La distribución de frecuencias se puede ver en la figura 25, el resto de resultados se pueden consultar en el anexo [placeholder ref].

Sobra decir que los resultados tampoco fueron muy prometedores. Por tanto dejo para un trabajo la parte de extracción de relaciones, y el resto del proyecto se centra en la extracción de personajes nombrados y determinación de su género.

10. PROGRAMA PRINCIPAL: `fic_character_extractor`

El programa principal se lanza desde la terminal y tiene dos posibles comandos:

- `fic_character_extractor <fic_index>`, que analizará el fanfic cuyo identificador sea `fic_index`. Por tanto, se comprueba que el usuario no introduzca un identificador menor que 0 ni mayor que 20190.
- `fic_character_extractor`, que analizará un fanfic elegido al azar del total de fanfics disponibles.

Debido a la latencia de CoreNLP, evita elegir un fanfic que tenga más de 50000 caracteres. En ambos casos el programa utiliza la clase `FanficGetter` del módulo `fanfic_util` para extraer el fanfic elegido, obteniendo así el objeto `Fanfic` que encapsula toda la información necesaria del mismo. A continuación, se siguen los siguientes pasos:

1. Identificación de entidades con NERTagger: El texto del fanfic es *tokenizado* en frases y palabras antes de etiquetar cada palabra con su rol morfológico, usando para ello las herramientas de procesamiento de texto de NLTK. A continuación, se utiliza la función `parse()` de la clase `NERTagger` del módulo `NER_tagger(9.1.1)` para extraer los personajes del texto.
2. Identificación de entidades con CoreNLP: se utiliza la clase `CoreWrapper` para enviar el texto al servidor de CoreNLP, y la clase `CoreNLPPDataProcessor` extrae los personajes a partir de la respuesta. Ambas clases pertenecen al módulo `corenlp_util(9.1.2)`.
3. Análisis de sentimiento con CoreNLP: se utiliza la clase `CoreNLPPDataProcessor` de `corenlp_util` para extraer el sentimiento del fanfic y mostrar si es principalmente positivo o negativo.
4. Mostrar en pantalla los resultados, junto con el título y etiquetas de personaje del fanfic.

Las etiquetas de personaje de un fanfic son simplemente la forma del autor de indicar qué personajes aparecen en él. Por motivos técnicos evidentes, los autores se limitan a etiquetar a los personajes más importantes del texto, por lo que es esperable que NERTagger y CoreNLP encuentren más personajes en el texto. Sin embargo, como mínimo no deberían dejar sin identificar ninguno de los personajes etiquetados. Las etiquetas de personaje, como el título, son metadatos del fanfic que se extraen directamente del mismo usando la clase `FanficHTMLHandler` del módulo `fanfic_util`.

Los personajes detectados por NERTagger y CoreNLP serán algo distintos, pero por lo general las coincidencias son bastante consistentes. Las menciones de los personajes de NERTagger siempre serán más bajas que los de CoreNLP, ya que éste último no sólo cuenta cuando al personaje se le menciona por el nombre, sino que también puede detectar menciones puramente pronominales.

Las menciones de CoreNLP aparecerán fraccionadas según género, de modo que un personaje puede

tener 100 menciones masculinas, 3 femeninas, 0 neutras y 12 desconocidas (cuando no ha sido posible asignar ningún género a dicha mención). La intención de mostrar estos datos así es poder cuantificar cómo de seguro está el programa sobre el género de un personaje.



fig:pantallazo ejecucion main program

11. EVALUACIÓN DEL SISTEMA

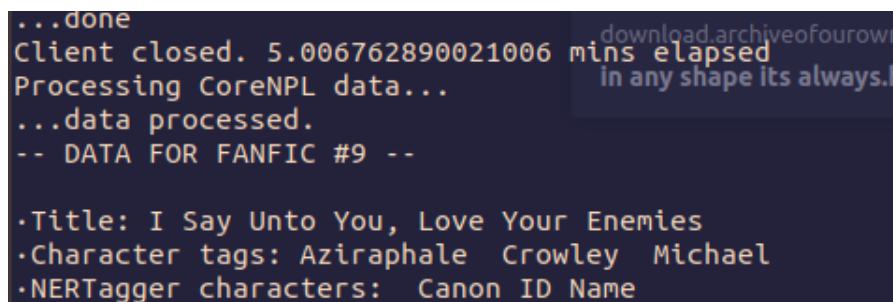
Vamos a evaluar el programa según la cantidad de personajes correctamente identificados y si su género concuerda con el género que el personaje realmente tiene en el texto. Los motivos que me llevaron a elegir cada fanfic para cada prueba serán explicados en la misma, aunque todos tienen en común que no son demasiado largos, lo cual ayuda tanto a que yo como CoreNLP no tardemos mucho en obtener la información que necesitamos de ellos.

La mayoría de información aparece transcrita en tablas vez de ser pantallazos del programa, para ahorrar espacio, pero la información aparece en pantalla con el mismo formato.

11.1. Prueba 1: Funciones básicas del programa con un texto largo (Fanfic 9)

Escogí este fanfic para la primera prueba porque tiene muchos personajes, algunos de los cuales sólo se les menciona una o dos veces, y con 8821 palabras es un texto medianamente largo, útil para recabar información sobre los personajes. Además, este fic particular tiene varias erratas, lo que también pondrá a prueba la capacidad del programa para identificar un personaje con el nombre ligeramente incorrecto.

Ejecutar el comando `fic_character_extractor 9`. En la figura ?? se muestran los metadatos del fanfic, y en las tablas 3 y 4 se han transcrito los personajes identificados por NERTagger y CoreNLProcessor.



```
...done
Client closed. 5.006762890021006 mins elapsed
Processing CoreNLP data...
...data processed.
-- DATA FOR FANFIC #9 --

·Title: I Say Unto You, Love Your Enemies
·Character tags: Aziraphale Crowley Michael
·NERTagger characters: Canon ID Name
```

Figura 26: Título y etiquetas de personaje del fanfic 9.

De entrada, tenemos los personajes de las etiquetas: Aziraphale, Crowley y Michael. Son los personajes que el autor decidió etiquetar, probablemente por considerarlos los más importantes, y los tres son identificados correctamente tanto por NERTagger como por CoreNLProcessor.

NERTagger ha identificado 35 personajes distintos, de los cuales 8 identifica como canon. Sus resultados, mostrados en la tabla 3, permite observar que lista en entradas distintas a personajes que son claramente el mismo pero con una errata (Ashemedai y Ashmedai, Azriaphale y Aziraphale), pero los identifica correctamente con el mismo personaje (Azriaphale y Aziraphale tienen ambos el identificador 4). Destacan

las entradas 'Lord Beezlebub' y 'Beezlebub': ambas tienen la misma errata, pero el primero no aparece identificado como personaje canon, mientras que el segundo sí. Esto probablemente sea por la distancia de edición y el tamaño de los nombres: 'Lord Beezlebub' son dos palabras y la más corta es Lord, con sólo 4 letras. Esto significa que el algoritmo explicado en la sección 9.1.1 sobre cuándo coinciden dos nombres habrá considerado que la distancia máxima de edición sólo puede ser 1. Cuando hay dos palabras en un nombre, como en este caso, se escoge la que tenga la distancia de edición más pequeña, que en este caso sería la distancia entre 'Beezlebub' y 'Beelzebub', que es 2, excediendo el límite de distancia. Sin embargo, cuando la única palabra en el nombre es 'Beezlebub' y se compara directamente con 'Beelzebub', al ser un nombre más largo la distancia máxima de edición es 3, y NERTagger por tanto lo identifica correctamente como canon.

También hay algunos personajes con nombres 'raros' como 'Did Gabriel' o 'Aziraphale never', en los que claramente NERTagger ha incluido en el nombre del personaje una palabra que no correspondía. Sin embargo, puesto que al menos parte del 'nombre' concide con un personaje canon, son identificados correctamente.

Los 8 personajes que identifica como canon están correctamente identificados excepto 'Heaven', que confunde un personaje canon llamado 'Raven'. En el texto aparecen otros tres personajes canon que el programa no ha detectado, y sólo dos personajes no canon (mencionados cada uno sólo una vez).

CoreNLProcessor por su parte identifica 9 personajes canon, incluyendo a Ligur, que se le escapó a NERTagger. En la tabla 4 también podemos ver a los personajes Agnes Nutter y Newton Pulsifer, que en realidad no aparecen en el texto pero aparecen incorrectamente identificados como la versión canon de Agares y Beeton (de nuevo, debido a la distancia de edición). Todo el resto de personajes listados en la tabla son personajes reales que aparecen en el texto, y sus géneros están correctamente asignados en todos los casos excepto Beelzebub, que en esta historia tiene pronombres femeninos. También hay que destacar el personaje Laradiri, inventado por el autor y cuyo género no es definido en ningún momento, siendo referido únicamente con el pronombre neutro inglés *they*. Como consecuencia, sus menciones aparecen marcadas como neutras o 'desconocidas'.

Sólo hay un personaje canon y un personaje no canon que ninguno de los dos programas han detectado.

Otro problema claro es que CoreNLProcessor no parece capaz de consolidar correctamente los personajes que no son canon, apareciendo repetidos en vez de en una única entrada que recoja todas las menciones.

11.2. Prueba 2: Género distinto del canon (Fanfic 2856)

En este fanfic, el autor decidió cambiar el género de los protagonistas, por lo que en su historia los personajes Crowley y Aziraphale son dos mujeres. Es un relato con más de 3000 palabras, por lo que no debería de ser poca información.

Cambiar el género de personajes es bastante común en el género fanfic, y es éste uno de los motivos por los cuales analizar fanfiction puede ser tan interesante: los autores juegan con los personajes y exploran su psique desde muchos ángulos, incluido el de género. Nuestro programa debería identificar que, en este texto, la mayoría de menciones a Crowley y Aziraphale son femeninas.

```
...done
Client closed. 1.9697413285573324 mins elapsed
Processing CoreNPL data...
...data processed.
-- DATA FOR FANFIC #2856 --

·Title: Like the Sweet Apple
·Character tags: Aziraphale Crowley
```

Figura 27: Título y etiquetas de personaje del fanfic 2856.

En la tabla 6 podemos ver que la detección de género ha fallado totalmente. A pesar de que no son referidos jamás en masculino en todo el texto, CoreNLProcessor sigue asignándoles menciones masculinas de forma casi exclusiva.

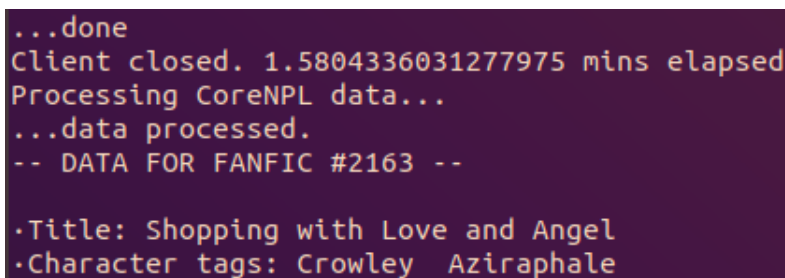
Por lo demás, las identificaciones no son incorrectas. Como se pueden comprobar con los metadatos de la figura ??, Crowley y Aziraphale son correctamente identificados en el texto por ambos programas. En la tabla 5 se muestra que además de los dos protagonistas ha identificado otros 16 nombres, pero sólo 'Eisheth' es realmente otro personaje en el texto ('Mesopotamia' aparece, pero es obviamente un lugar). También confunde 'Rather' con un personaje canon, Aziraphale (Canon ID = 4), probablemente porque uno de sus apodos es '*Brother* Francis', e inexplicablemente, confunde 'Azi Eisheth' con el personaje canon con ID = 38, cuyo nombre es 'Jeremey Wendsleydale' y que no se le menciona por ninguna parte en el texto.

CoreNLProcessor, a pesar del fracaso con el género, identifica correctamente a Aziraphale, Crowley y Adam como personajes canon, e incluso que Crowley en este relato adopta el nombre 'Zadkiel' brevemente. También identifica a Eisheth y a 'Serpent of Eden', que en principio tendría que haber sido vinculado con Crowley, al ser 'Serpent' uno de sus apodos. No añade ningún personaje que no aparezca en el texto.

11.3. Prueba 3: Personajes que no son nombrados (Fanfic 2163)

Decidí probar este fanfic porque los protagonistas Crowley y Aziraphale aparecen, pero son vistos desde la perspectiva de dos extrañas que no les conocen, y por tanto, no saben sus nombres. Además, este fanfic tiene tan sólo 876 palabras, con lo que a la falta de nombres se le añade una información limitada.

Ejecutamos el comando `fic_character_extractor 2163`:



```
...done
Client closed. 1.5804336031277975 mins elapsed
Processing CoreNPL data...
...data processed.
-- DATA FOR FANFIC #2163 --

·Title: Shopping with Love and Angel
·Character tags: Crowley Aziraphale
```

Figura 28: Título y etiquetas de personaje del fanfic 2163.

Ambos programas han identificado correctamente a Lauren y Olivia, los personajes originales del autor desde cuya perspectiva se cuenta este relato. Ambas son también correctamente identificadas como personajes que no son canon (Canon ID es 'NO'), y en 8 podemos ver que todas sus menciones son femeninas excepto las de una de las entradas repetidas de 'Angel Olivia', que tiene 24 menciones 'desconocidas'. Como en la prueba 2, tanto Crowley como Aziraphale son mujeres en esta historia, pero todas las menciones de Aziraphale son identificadas como masculinas. La buena noticia es que tanto NERTagger como CoreNLProcessor han sido capaces de identificar correctamente a Aziraphale por su apodo 'angel', y, curiosamente, en la tabla 7 se puede ver que NERTagger también ha identificado que 'Love' en este relato es un apodo cariñoso para un personaje. Sin embargo, también identifica incorrectamente la palabra 'Well' como refiriéndose a Aziraphale (probablemente por parecido con uno de sus apodos, 'Fell'). Esto significa que NERTagger ha identificado 11 personajes, de los cuáles sólo 4 son correctos (aunque un humano puede detectar fácilmente qué nombres son los falsos positivos), mientras que CoreNLProcessor ha identificado 4 personajes distintos, de los cuales dos son correctos y uno está correctamente identificado con su correspondiente canon, pero el género es incorrecto.

Como se puede ver en los metadatos de la figura ??, ambos programas han identificado correctamente a Aziraphale pero ninguno a Crowley; no es de extrañar, ya que ni su nombre ni ninguno de sus apodos es mencionado en todo el texto, y al propio Aziraphale sólo se le ha identificado por su apodo 'angel'. Sin embargo, NERTagger encuentra el nombre 'Love' y sabe que es un nombre; simplemente no tenía forma de saber que se refería a Crowley. CoreNLProcessor no ha sido capaz de identificar 'Love' como un nombre.

Canon ID	Name	Mentions
NO	Lord Beezlebub	2
NO	Lilith	1
14	Did Gabriel	1
NO	Ashmedai	11
4	Aziraphale	42
NO	Laradiri	16
NO	Every Woman	1
NO	Marut	1
NO	Joe	1
NO	Amides	1
NO	Dr Dudders	1
NO	History	1
14	Gabriel	1
NO	Butter	1
NO	Alisha	1
4	Aziraphale never	1
NO	Mrs Beeton	1
NO	Richard	1
NO	Heavenly	1
4	Azriaphale	1
4	Mr Fell	2
NO	Seamus Blackley	1
24	Michael	1
NO	Do NOT	1
10	Death	1
14	Sir Yes Sir Gabriel Sir	1
13	Heaven	1
NO	Ashemedai	1
9	Dagon	1
NO	Pride	1
8	Crowley	45
NO	So Below	1
NO	Any	1
NO	Inside	1
15	God	1
NO	Got	1
5	Beezlebub	1
NO	Word	1
8	Mr Crowley	1
4	Aziraphale	42
8	Crowley One	1
15	God Herself	1
NO	Mr Solomons	3
NO	Cookery Reformed	1

Cuadro 3: Resultados de la ejecución de fic_character_extractor para analizar fanfic 9

Canon ID	Name	MALE	FEMALE	NEUTRAL	UNKNOWN	Other names
1	Agnes Nutter	0	1	0	0	Agares
4	Aziraphale	147	0	4	0	aziraphale, Azriaphale, Fell, azirphale, azriaphale, consume Aziraphale, Azirphale
5	Beelzebub	4	0	3	0	Beezlebub
8	Crowley	104	0	7	0	crawley, crowley, Shop Crowley, Crawley
14	Gabriel	6	0	0	0	
21	Ligur	1	0	0	0	
24	Michael	2	0	0	0	
25	Newton Pulsifer	0	1			Beeton
36	Uriel	1	0	0	0	
NO	Bentley	1	0	0	1	
NO	Somolons	3	0	0	0	
NO	Laradiri	0	0	0	1	
NO	Eden	0	0	0	1	
NO	Petronius	1	0	0	0	
NO	Ashmedai	0	0	0	1	
NO	Angelo	1	0	0	0	
NO	Ashemedai	0	0	0	4	
NO	Laradiri	0	0	0	1	
NO	Angelo	1	0	0	0	
NO	Richard	1	0	0	0	
NO	Ashmedai	0	0	0	3	
NO	Jane Austen	0	0	0	1	
NO	Richards	1	0	0	0	
NO	Marut	1	0	0	0	
NO	Joe	5	0	0	0	
NO	Laradiri	1	0	0	0	
NO	Alisha	0	2	0	0	
NO	Solomons	1	0	0	0	
NO	Perkins	1	0	0	0	
NO	Hannah Glasse	0	1	0	0	
NO	Jane Austen	0	1	0	0	
NO	Seamus Blackley	1	0	0	0	
NO	Jane Austen	0	2	0	0	
NO	Laradiri	0	0	0	42	
NO	Ashmedai	0	0	0	13	
NO	Ashmedai	0	0	0	2	

Cuadro 4: Resultados de la ejecución de fic_character_extractor para analizar fanfic 9

Canon ID	Name	Mentions
NO	Mesopotamia	1
NO	How	1
NO	Green	3
NO	out	1
NO	Eisheth	3
4	Aziraphale	48
NO	Unhand	1
38	Azi Eisheth	1
NO	Always	1
8	Crowley	62
NO	See	1
NO	No	1
4	Rather	3
NO	gorgeous	1
NO	Just	1
NO	Villain	1
NO	Oh	4
NO	Shouldn	1

Cuadro 5: Resultados de la ejecución de fic_character_extractor para analizar fanfic 2856

Canon ID	Name	MALE	FEMALE	NEUTRAL	UNKNOWN	Other names
0	Adam Young	1	0	0	0	Adam
4	Aziraphale	217	0	1	0	
8	Crowley	445	0	4	0	Crowley Zadkiel
NO	Eisheth	0	0	0	2	
NO	Serpent of Eden	0	0	0	1	

Cuadro 6: Resultados de la ejecución de fic_character_extractor para analizar fanfic 2856

Canon ID	Name	Mentions
NO	Really inaccurate	1
NO	Love	1
4	Angel	8
NO	Yeah	1
4	Well	1
NO	sort	1
NO	Did	1
NO	Oh	3
NO	Really	1
NO	Pardon	1
NO	Olivia	7
NO	Hey	1
NO	Lauren	3
NO	Um	1

Cuadro 7: Resultados de la ejecución de fic_character_extractor para analizar fanfic 2163

Canon ID	Name	MALE	FEMALE	NEUTRAL	UNKNOWN	Other names
4	Aziraphale	7	0	0	0	Angel
NO	Angel Olivia	3	0	0	0	
NO	Olivia	0	7	0	0	
NO	Lauren	0	4	0	0	
NO	Olivia	0	1	0	0	
NO	Olivia	0	1	0	0	
NO	Angel Olivia	0	0	0	24	
NO	Angel Olivia	0	6	0	0	

Cuadro 8: Resultados de la ejecución de fic_character_extractor para analizar fanfic 2163

12. CONCLUSIONES

En este trabajo se ha desarrollado un sistema de extracción y análisis de textos de internet, utilizando técnicas de *scraping* y procesamiento de texto natural.

En primer lugar se ha creado un corpus de relatos pertenecientes al género fanfiction a partir de los textos alojados en la web [Archive of our Own](#), y se ha utilizado este corpus para realizar pruebas, entrenamientos y experimentos que ayudaron en el desarrollo de un sistema de procesamiento de texto natural. Se han seleccionado textos en inglés que se basaran en el libro *Good Omens* (Neil Gaiman y Terry Pratchett, 1990) y que tuvieran una cantidad mínima de palabras para asegurar que la obra estaba compuesta principalmente por texto (en vez de imágenes o audio), y se ha desarrollado un sistema de scrapers capaces de extraer los relatos de la web que tuvieran esas características. El resultado ha sido un corpus compuesto por archivos HTML para cuyo acceso y manejo se han desarrollado diversos objetos y funciones en python, encapsuladas en un módulo que he llamado *fanfic_util*. Además, aprovechando los metadatos de los relatos del corpus se han creado tres datasets, cada uno consistente en relatos de un sólo capítulo centrados en romance, amistad y enemistad, respectivamente. Estos datasets se utilizan para experimentos durante la búsqueda de estrategias para la identificación de relaciones.

En el procesamiento de texto natural se han creado dos algoritmos de identificación de entidades, uno basado en naïve Bayes y otro en un modelo de regresión logística. Ambos fueron desarrollados con la librería NLTK de python (en conjunto con el módulo *megam* en el caso del modelo de regresión logística), y entrenados con el corpus *Groninger Meaning Bank*, que fue creado expresamente para entrenar algoritmos para identificar entidades en textos en inglés. La versión con el modelo de regresión logística consiguió mejores resultados, como se puede ver en la figura 12, con lo que se decide utilizar esta versión bajo el nombre *NER_tagger*.

Para la parte de extracción de relaciones del procesamiento de texto natural, se han explorado varias estrategias. Usando *sci-kit learn* y *gensim* se han creado algoritmos de clustering y un modelo LDA, con la esperanza de que fueran capaces de clasificar correctamente a qué dataset pertenecía cada relato (romance, amistad o enemistad). La idea era que si podían clasificarlos correctamente, las características más relevantes de cada cluster o del modelo podrían utilizarse para refinar el algoritmo y crear una versión que pudiese identificar romance, amistad o enemistad incluso en textos largos donde no fueran el tema central, pero lamentablemente no se obtuvieron resultados mucho mejores que el azar.

La última estrategia para identificar relaciones consistía en explotar la correferencia pronominal en frases donde se nombraran a los personajes, utilizando para ello el servidor de Stanford CoreNLP mediante su biblioteca para python, *Stanza*. La idea era identificar todas las menciones de un personaje en un texto,

incluso aquellas que sólo se les mencionaran por el pronombre (*he, she, they*), aumentando por tanto la cantidad de frases relevantes para el algoritmo. CoreNLP tiene, entre otras funciones, la capacidad de resolver esta correferencia pronominal en textos, por lo que se ha creado un conjunto de funciones para enviar peticiones de forma sencilla y controlada a su servidor, sin inundarlo, y manejar errores de red de forma que se pierda la menor cantidad posible. Todas estas funciones han sido encapsuladas en un módulo que he llamado *corenlp_util*. Mediante este módulo los relatos de los datasets de romance, amistad y enemistad han sido procesados y etiquetados, y los resultados se han guardado en un archivo con el que se realizaron aún más experimentos para buscar alguna forma de identificar relaciones entre entidades. Estos últimos experimentos consistieron en buscar patrones relacionados con los verbos, adjetivos, bigramas o trigramas más frecuentes en las frases que mencionasen al menos a dos personajes, pero tampoco se pudo encontrar ningún patrón que fuese más relevante que el azar.

Decidí entonces dejar el desarrollo de un algoritmo de identificación de relaciones sociales entre personajes para el futuro, y me centré en el algoritmo de identificación de personajes. Ya tenía uno, basado en regresión logística, pero como durante el desarrollo de las pruebas con CoreNLP para realizar correferencia también tuve que manejar las funciones de CoreNLP para identificación de entidades, esto resultó en un segundo algoritmo de identificación de personajes que he llamado *CoreNLPPDataProcessor*, encapsulado también dentro del módulo *corenlp_util*. A diferencia de la versión basada en un modelo de regresión logística, esta versión puede identificar no sólo el nombre sino también el género del personaje, además de aprovechar la resolución de correferencia para identificar más menciones.

El resultado final por tanto, ha sido un programa que he llamado *fic_character extractor*, que utiliza ambos algoritmos de identificación de personajes para extraer información de un texto. Este programa ofrece al usuario la posibilidad de extraer personajes de cualquiera de los relatos del corpus creado en la primera parte del proyecto, dando la posibilidad de analizar uno al azar o introducir el identificador numérico de un relato particular. Entonces, el programa utiliza tanto *NER_tagger* como *CoreNLPPDataProcessor* para analizar el texto y mostrar al usuario el nombre de los personajes, si son canon o no, el número de menciones y, en el caso de este último, su género.

De los objetivos originales expuestos en la introducción (5), se han logrado con éxito cuatro de los seis objetivos, uno que ha fallado y el último, que se ha logrado en parte:

1. Se ha conseguido extraer un conjunto de relatos de la web [Archive of Our Own](#) sobre los que utilizar técnicas de extracción de información, mediante la creación de dos scrapers que seleccionan y descargan los archivos allí alojados.
2. Se ha creado el módulo *fanfic_util* para manejar y extraer información de los archivos HTML de

AO3, de modo que los algoritmos de procesamiento de texto natural pueden acceder tanto al texto puro como a los metadatos de cada archivo de forma simple.

3. Utilizando las herramientas de *fanfic_util*, se han creado tres datasets sobre los que realizar pruebas y experimentos.
4. Se han desarrollado dos algoritmos que identifican personajes en textos de ficción, uno basado en un modelo de regresión logística creado con NLTK y *megam* y entrenado con *Groningen Meaning Bank* para identificar entidades en inglés, y otro que aprovecha las capacidades para identificar entidades de CoreNLP, a través de la biblioteca *Stanza*. Además de identificar sus nombres, también pueden identificar si los personajes son originales del autor fan o si ya existían en la obra original.
5. No se ha conseguido desarrollar un algoritmo de identificación de relaciones entre entidades, aunque se han probado diversas técnicas no-supervisadas y las capacidades de resolución de co-referencia de CoreNLP.
6. No se ha conseguido crear un programa que utilice tanto identificación de entidades y relaciones para los personajes y relaciones en los relatos recogidos de AO3, y mostrarlas al usuario. Sin embargo, sí se ha conseguido un programa que utiliza dos métodos de identificación de entidades para extraer los personajes de un texto, así como información relevante sobre los mismos (número de menciones, género, si son canon). Por lo tanto, este objetivo se ha conseguido parcialmente.

El objetivo final de este proyecto, que era el desarrollo de una herramienta para ayuda al análisis literario, ha quedado conseguido tan sólo en parte. La identificación automática de los personajes de un conjunto de relatos puede ser muy útil para saber qué personajes son más populares entre los fans y pueden dar pie a teorías interesantes según si se corresponden o no con los protagonistas de la obra original, o dividiendo la comunidad fan a estudiar en subcategorías: si se tiene datos sobre los autores, se podría buscar si todos los personajes son populares de forma homogénea en toda la comunidad o si hay algunos que son más populares entre los hombres, si algunos son más populares entre los escritores que además de ser fans de esta obra lo son también de *Harry Potter*, etc. Como el programa final también arroja datos sobre cosas como el género de los personajes y si son canon, un investigador podría utilizar estos datos para explorar, por ejemplo, si hay tendencias entre los fans a interpretar a ciertos personajes con un género diferente al de la obra original, o qué tipos de autores tienen más tendencia a escribir personajes originales. Incluso, se podrían investigar qué personajes no originales se vuelven populares en toda la comunidad, en vez de quedar limitado a los fanfic de un único autor.

Sin embargo, la falta de información sobre las relaciones entre personajes limita mucho los análisis posibles. Sin tener una idea de qué personajes tienen relación de enemistad o romances, es difícil dilucidar

qué personajes son interpretados como los "buenos" o los "malos" de una historia, con lo que un investigador tendría que leer los relatos en más profundidad para sacar una conclusión sobre las tendencias de los fans relacionadas con la redención o corrupción de personajes, y sobre la moralidad que les transmiten ciertos arcos o características de personajes.

Para consultar cualquier aspecto del código e implementación del proyecto, se puede visitar el repositorio del mismo en [GitHub](#) (Usuario: *mariaGnlz*, repositorio: *Fanfic_ontology*).

Durante todo este proceso he tenido que aprender las bases de la extracción de información, técnicas de *machine learning* y cómo crear, organizar y preprocesar un conjunto de archivos para que su información sea comprensible para los algoritmos que los utilizan como entrada. En cierto aspecto este proyecto me ha hecho perderle el miedo al procesado del lenguaje humano, que siempre había visto como algo extremadamente complicado, y aunque ciertamente no es una tarea trivial, he podido ver de primera mano que existen métodos bien establecidos para la extracción de información y que pueden ser aprendidos y entendidos. El modelo de regresión logística utilizado en la sección 9.1.1, por ejemplo, me llevó a repasar funciones y gradientes para poder entender su base matemática, y me di cuenta de que sí, es complejo, pero no es magia.

Mientras me encontraba con que la parte relacionada con el procesado de lenguaje natural en sí no era tan complicada como temía, los problemas relacionados con el manejo de archivos HTML y extraer el texto puro me pillaron por sorpresa en lo retorcidos y frustrantes que podían llegar a ser. Detalles como puntos y coma que destrozan el formato de una base de datos, o la etiqueta HTML que utilizaba para extraer un cierto metadato funciona en la mayoría de archivos pero está misteriosamente ausente en otros, no hizo más que recordarme que la mayoría del esfuerzo en ciencia de datos suele ir a limpiar los datos y darles un formato uniforme.

El diseño de ciertas del proyecto también es algo que hubiese planificado mejor; la clase *Fanfic* de la sección 8.2 acaba siendo la unidad de información básica, pero se desarrolló orgánicamente a medida que el proyecto avanzaba, especialmente mientras buscaba alguna forma de aprovechar las funciones de CoreNLP para identificar relaciones en el texto (ya que es la parte en la que los metadatos de un fanfic adquirieron más importancia). En retrospectiva, crear una clase que encapsule todas las características de un fanfic tenía mucho sentido para el proyecto, y pensar que simplemente con el texto puro de cada obra iba a ser suficiente fue un poco ingenuo. Todo el módulo *fanfic_util* podría haberse beneficiado de haber planificado la clase *Fanfic* desde el principio, por no hablar de ahorrarme trabajo.

12.1. Trabajos futuros

Un añadido evidente para este proyecto sería mejorar el manejo de archivos HTML, introducidos en una base de datos que facilite su filtrado según sus etiquetas o autor o cualquiera de sus metadatos, ya que ahora mismo no tiene un mecanismo de búsqueda generalizado, y los *datasets* fueron creados mediante comandos de python en la terminal. Otra mejora es generalizar la canonicalización de personajes. Este proyecto utiliza una base de datos con los personajes de *Good Omens* para decidir la canonicidad de los personajes de un fanfic dado, por lo que ahora mismo sólo los fanfics de *Good Omens* pueden tener sus personajes marcados como canon. Sin embargo, se podría crear un método que consultara el título de la obra original en los metadatos del fanfic y comprobase si hay alguna wiki dedicada a esta obra en internet, y usar la página de personajes de dicha wiki para decidir qué personajes del fanfic son canon o no. De esta manera el proceso de canonicalización funcionaría para cualquier fanfic cuya obra original tenga una wiki.

El proceso de consolidación de menciones en personajes también podría mejorarse, ya que ahora mismo es un proceso basado en la distancia de edición de los nombres y, en el caso del extractor de personajes que utiliza CoreNLP (9.1.2), también en el género. Un programa más sofisticado podría utilizar más información del contexto de la mención para captar más características de un personaje particular (títulos, nombre y apellidos, especie, país); en otras palabras, adoptar una estrategia que se centre en los personajes como entidad [Wic09] que trate de rellenar una "ficha" para cada candidato a personaje podría suponer una mejora para todo el proceso.

Para continuar con los esfuerzos de identificación de relaciones sociales se podría explorar el entrenar un modelo de aprendizaje supervisado, aprovechando las etiquetas de los metadatos de un fanfic para etiquetar la relación principal de un relato. También se podrían explorar técnicas de aprendizaje no supervisado distintas a las ya vistas, como las redes neuronales [Pen17], que pueden acceder a una mayor cantidad de contexto a la hora de decidir si existe una relación entre dos entidades particulares.

De cara al usuario, una mejora sería modificar la entrada del programa de modo que acepte un link de una obra de AO3, evitando así que tenga que descargar el archivo y configurar el *path* para que el programa lo encuentre.

13. REFERENCIAS

Referencias

- [Bar68] Ronald Barthes. La mort de l’auteur. *Manteia*, (5), 1968.
- [Bir12] Steven Bird. Natural language processing with python. https://www.nltk.org/book_1ed/ch07.html, October 2012.
- [Cra99] Mark Craven. Constructing biological knowledge bases by extracting information from text sources. *SMB-99 Proceedings*, 1999.
- [Eis18] Jacob Eisenstein. *Natural Language Processing*. MIT Press, Nov 2018.
- [Ell18] Lindsay Ellis. Death of the author. https://www.youtube.com/watch?v=MGN9x4-Y_7A, December 2018. Youtube.
- [Fin05] Jenny Rose Finkel. Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 2005.
- [iva] Complete guide to build your own named entity recognizer with python. <https://nlpforhackers.io/named-entity-extraction/>. NLP for Hackers.
- [jos17] Information extraction. <https://github.com/rohitjose/InformationExtraction>, 2017.
- [Laf01] John Lafferty. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. June 2001.
- [Lev66] Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10, 1966.
- [Lin09] Dekang Lin. Phrase clustering for discriminative learning. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 151 – 213, August 2009.
- [Man19] Matteo Manica. An information extraction and knowledge graph platform for accelerating biochemical discoveries. *Workshop on Applied Data Science for Healthcare at KDD*, 2019.
- [not13] Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151 – 175, 2013.
- [Pen17] Nanyun Peng. Cross-sentence n-ary relation extraction with graph lstms. *Arxiv*, 2017.
- [skl] Feature extraction: Tf-idf term weighting. https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction. Scikit Learn User Guide.
- [Stu17] Alasdair Stuart. Dean winchester and commander shepard walk into a bar: Why fanon matters. *Uncanny Magazine*, July/August 2017.
- [Swi98] Jonathan Swift. Copyright 101: A brief introduction to copyright for fan fiction writers. <http://www.whoosh.org/issue25/lee1a.html#41>, October 1998. Woosh Magazine, Birthplace of the International Association of Xena Studies.
- [Wic09] Michael Wick. An entity based model for coreference resolution. 2009.
- [Zel03] Dimitri Zelenko. Kernel methods for relation extraction. *Journal of Machine Learning Research*, (3):1083–1106, 2003.