



UNIVERSIDAD
POLITÉCNICA
DE MADRID



SISTEMA DE PREGUNTA-RESPUESTA

Alejandro Gabarre González
César García Cabeza

Curso académico 2020-21

Enero 2021

Asignatura:
Ingeniería Lingüística

Professor:
D. Jesús Cardeñosa Lera

Índice

1. Introducción	2
1.1. Objetivos	2
1.2. Material entregado	2
1.3. Estructura	3
2. Elección de escenarios	4
3. Análisis de dimensiones	5
4. Creación de preguntas	11
5. Obtención de reglas	15
6. Implementación	20
7. Experimentación	22
8. Conclusiones	25
A. Manual de uso	26

1. Introducción

Los sistemas de pregunta respuesta son un tipo de sistema dentro de los campos de la Recuperación y Extracción de Información encargados de responder automáticamente a preguntas en lenguaje natural formuladas por usuarios. Dentro de este tipo de sistema, podemos distinguir dos clases:

1. Pregunta-Respuesta en un dominio cerrado: consiste en construir un sistema capaz de contestar preguntas para un dominio específico, siendo incapaz de responder preguntas que se encuentren fuera del dominio.
2. Pregunta-Respuesta en un dominio abierto: consiste en construir un sistema capaz de responder preguntas sobre prácticamente cualquier cosa. Para construir estos sistemas es necesario el uso de datasets muy grandes.

Una vez explicados los dos tipos de sistemas, está claro que nuestro sistema entra dentro de la primera categoría, al tratarse de un dominio muy específico que explicaremos en el siguiente capítulo.

1.1. Objetivos

El objetivo principal de esta práctica es el de diseñar e implementar un sistema de pregunta respuesta que sea capaz de responder a una serie de preguntas sobre los dos textos escogidos. Con el fin de lograr este objetivo, se plantean los siguientes hitos:

1. Elección de escenarios.
2. Análisis de dimensiones.
3. Creación de preguntas.
4. Parafraseo de las preguntas.
5. Obtención de reglas.
6. Implementación del sistema.

1.2. Material entregado

Para cumplir con el objetivo principal y con los hitos especificados, entregamos un fichero .zip con el nombre *QA-System.zip*, como se puede ver en la Figura 1.

- En la subcarpeta *documents* se encuentran los dos documentos que analizaremos para crear nuestro sistema.
- En la subcarpeta *answers* se irán guardando los resúmenes de las preguntas y respuestas.
- En la subcarpeta *src* se encuentran todos los ficheros de código de la implementación de nuestro sistema.
- Finalmente, el documento *gabarre_garcia_qa_system.pdf*, documento que está leyendo actualmente.

Destacar que todo este material lo puede encontrar en [este](#) repositorio de Github.

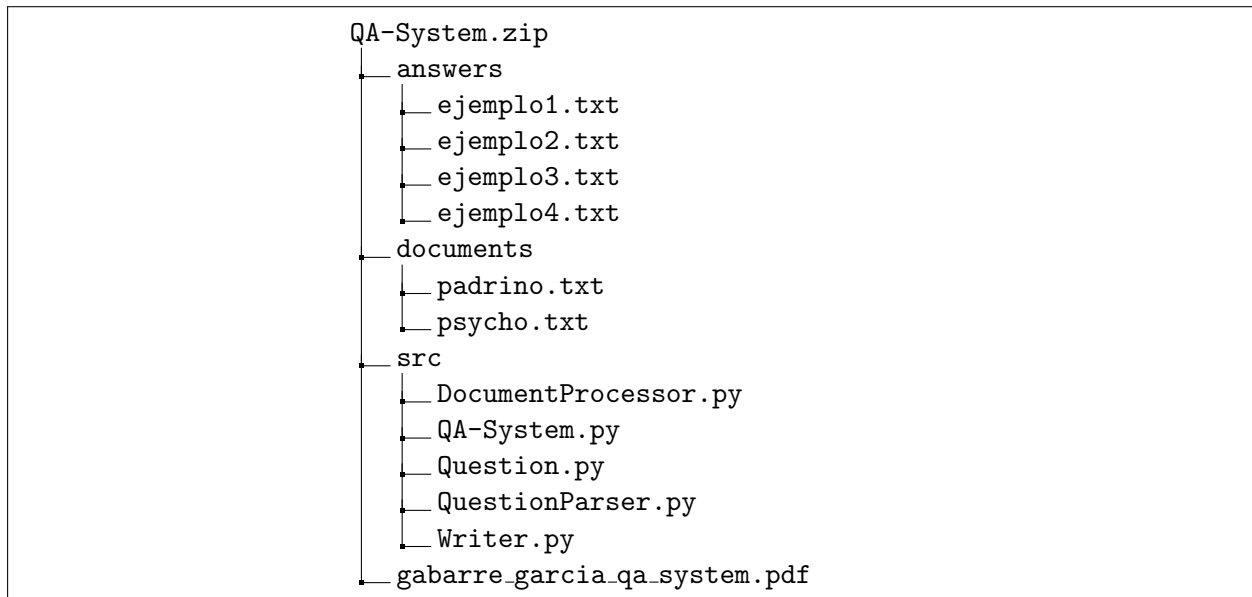


Figura 1: Estructura del zip.

1.3. Estructura

En el Capítulo 2 se escogerán dos textos similares a los cuales se les harán preguntas. Después, en el Capítulo 3, se analizarán las dimensiones de los textos escogidos que representan qué datos se les puede preguntar. En el Capítulo 4 se crearán las preguntas concretas a las que nuestro sistema va a ser capaz de responder, incluyendo un parafraseo de alguna de ellas. Seguidamente, en el Capítulo 5, se crearán las reglas capaces de identificar las preguntas definidas anteriormente. En el Capítulo 6 se implementará el sistema una vez concluido el análisis y diseño del mismo. El Capítulo 7 contendrá unos ejemplos de las respuestas obtenidas con nuestro sistema. Finalmente, en el Capítulo 8, se incluirán unas reflexiones finales sobre el trabajo realizado.

Además, se incluye en el Anexo A un manual de uso que permite la ejecución de nuestro proyecto.

2. Elección de escenarios

Al tratarse nuestro proyecto de un sistema de pregunta respuesta de dominio cerrado, a la hora de elegir nuestros dos escenarios deberán de ser del mismo dominio.

Otro aspecto a tener en cuenta sería el tipo de documento a escoger: estructurado o no estructurado. Un documento estructurado tendría su información organizada en bloques o secciones, siendo así -a priori- más sencillo el diseño de nuestro sistema. Por el contrario, un documento no estructurado tendría toda su información en un mismo bloque sin secciones ni bloques, haciendo así más difícil la tarea de diseñar nuestro sistema.

Los tipos de documentos propuestos en los ejemplos eran recetas de cocina o prospectos de medicamentos, ambos de tipo estructurado. Tras analizar posibles opciones, nos decidimos a usar fichas de películas, también de tipo estructurado. Para ello, utilizamos la página FilmAffinity¹. Esta página es una de las más conocidas dentro del mundo del cine y contiene información sobre las películas como *trailers*, fechas de estreno, etc.

En concreto, decidimos usar las dos siguientes fichas de películas:

- *The Godfather* (El padrino): puede consultarla [aquí](#).
- *Psycho* (Psicosis): puede consultarla [aquí](#).

Estas fichas nos permitirán preguntar varios datos sobre las películas como puede ser el año de estreno, el director... La información concreta sobre qué datos será nuestro sistema capaz de responder se tratará en detalle en el próximo capítulo.

¹<https://www.filmaffinity.com/es/main.html>

3. Análisis de dimensiones

Vamos a utilizar una imagen de la página web que contiene la ficha de la película "Psycho". La ventaja de utilizar un documento estructurado se ve claramente a la hora de localizar las dimensiones, las cuáles podemos obtener con facilidad. Vemos un ejemplo en la Figura 2 donde hemos remarcado con un recuadro verde cuáles son las localizaciones que vamos a utilizar.

Título original Psycho

Año 1960

Duración 109 min.

País Estados Unidos

Dirección Alfred Hitchcock

Guión Joseph Stefano (Novela: Robert Bloch)

Música Bernard Herrmann

Fotografía John L. Russell (B&W)

Reparto Anthony Perkins, Janet Leigh, John Gavin, Vera Miles, John McIntire, Martin Balsam, Simon Oakland, Patricia Hitchcock

Productora Paramount Pictures

Género Terror. Intriga. Thriller | Película de culto. Asesinos en serie. Slasher. Thriller psicológico

Grupos Psicosis

Sinopsis Marion Crane, una joven secretaria, tras cometer el robo de un dinero en su empresa, huye de la ciudad y, después de conducir durante horas, decide descansar en un pequeño y apartado motel de carretera regentado por un tímido joven llamado Norman Bates, que vive en la casa de al lado con su madre. (FILMAFFINITY)

Premios 1960: 4 nominaciones al Oscar: director, actriz sec. (Leigh), fotografía y dir. artística (1960: Globos de Oro: Mejor actriz secundaria (Janet Leigh), 1960: Sindicato de Directores (DGA): Nominada a Mejor director, 1960: Sindicato de Guionistas (WGA): Nominada a Mejor guión drama

Críticas "Una obra maestra del entretenimiento y el horror de la que Perkins nunca pudo huir"
 Javier Ocaña: Cinemanía ●
 "Una de las grandes obras maestras del género. Tensión, terror e intriga en un filme de imprescindible revisión"
 Fernando Morales: Diario El País ●
 "Lo que hace inmortal a 'Psycho', cuando muchas otras películas las hemos casi olvidado tras salir del cine, es que conecta directamente con nuestros miedos (...)
 Puntuación: ★★★★★ (sobre 4)" [🔗](#)
 Roger Ebert: rogerebert.com ●

8,4 105.633 votos

265 críticas - por títulos

Vota esta película

Añadir a listas

Disponible en

Suscripción

Compra

Alquiler

Figura 2: Localización de las dimensiones en el documento original

Podemos obtener toda la información, a excepción de la puntuación y el número de votos, de forma trivial utilizando el portapapeles. Tras pequeñas modificaciones, se crean los ficheros txt que se adjuntan en la carpeta *documents/*, con los nombres de *padrino.txt* y *psycho.txt* haciendo referencia a ambas películas

Además de las dimensiones indicadas en la Figura 2, nos ha parecido interesante incluir una

serie de subdimensiones dentro de algunas como puede ser la dimensión premios, al parecernos correcto separarla en distintos aspectos como el año en el que fue premiada, la institución que organizaba los premios y las categorías premiadas. Por lo tanto, en la Tabla 1 mostramos las dimensiones y subdimensiones obtenidas de la información proporcionada por los documentos.

Tabla 1: Estructura de las dimensiones

Dimensión	Subdimensión
Título	
Estreno	
Duración	
Dirección	
Guión	
Música	
Reparto	
Productora	
Género	
Grupos	
Sinopsis	
Valoración	Puntuación
	Votos
Premios	Año
	Institución
	Categoría
Criticas	Critica
	Autores

A continuación vamos a indicar los valores que encontramos en cada uno de nuestros documentos (1 para referirnos a *El Padrino* y 2 para referirnos a *Psicosis*) para cada dimensión y subdimensión.

3.1. Dimensión: título

1. "The Goodfather"
2. "Psycho"

3.2. Dimensión: estreno

1. "1972"
2. "1960"

3.3. Dimensión: duración

1. "175 minutos"
2. "109 minutos"

3.4. Dimensión: dirección

1. "Francis Ford Coppola"
2. "Alfred Hitchcock"

3.5. Dimensión: guión

Corresponde con los guionistas que participaron en la creación del guión de cada una de las dos películas.

1. "Francis Ford Coppola, Mario Puzo"
2. "Joseph Stefano"

3.6. Dimensión: música

1. "Nino Rota"
2. "Bernard Herrmann"

3.7. Dimensión: reparto

Corresponde con los actores que formaron parte de la película

1. "Marlon Brando, Al Pacino, James Caan, Robert Duvall, Diane Keaton, John Cazale, Talia Shire, Richard S. Castellano, Sterling Hayden, Gianni Russo, Rudy Bond, John Marley, Richard Conte, Al Lettieri, Abe Vigoda, Franco Citti, Lenny Montana, Al Martino, Joe Spinell, Simonetta Stefanelli, Morgana King, Alex Rocco, John Martino, Salvatore Corsitto, Richard Bright, Tony Giorgio, Vito Scotti, Jeannie Linero, Julie Gregg, Angelo Infanti, Corrado Giampa, Saro Urzi"
2. "Anthony Perkins, Janet Leigh, John Gavin, Vera Miles, John McIntire, Martin Balsam, Simon Oakland, Patricia Hitchcock"

3.8. Dimensión: 'productora

1. "Paramount Pictures, Alfran Productions"
2. "Paramount Pictures"

3.9. Dimensión: Género

1. "Drama. Mafia. Crimen. Años 40. Años 50. Familia. Película de culto"
2. "Terror. Intriga. Thriller. Película de culto. Asesinos en serie. Slasher. Thriller psicológico"

3.10. Dimensión: Grupos

Indica la saga, trilogía... de la película.

1. "Trilogía El Padrino — Adaptaciones de Mario Puzo"
2. "Psicosis"

3.11. Dimensión: Sinopsis

Breve resumen/sinopsis sin spoilers de la película.

1. "América, años 40. Don Vito Corleone (Marlon Brando) es el respetado y temido jefe de una de las cinco familias de la mafia de Nueva York. Tiene cuatro hijos: Connie (Talia Shire), el impulsivo Sonny (James Caan), el pusilánime Fredo (John Cazale) y Michael (Al Pacino), que no quiere saber nada de los negocios de su padre. Cuando Corleone, en contra de los consejos de 'Il consigliere' Tom Hagen (Robert Duvall), se niega a participar en el negocio de las drogas, el jefe de otra banda ordena su asesinato. Empieza entonces una violenta y cruenta guerra entre las familias mafiosas."
2. "Marion Crane, una joven secretaria, tras cometer el robo de un dinero en su empresa, huye de la ciudad y, después de conducir durante horas, decide descansar en un pequeño y apartado motel de carretera regentado por un tímido joven llamado Norman Bates, que vive en la casa de al lado con su madre."

3.12. Dimensión: Valoración

3.12.1. Subdimensión: Puntuación

Puntuación media de los votos para la película.

1. "9.0"
2. "8.4"

3.12.2. Subdimensión: Votos

Número de votos (puntuaciones) de la película.

1. "170.129"
2. "105.623"

3.13. Dimensión: Premios

3.13.1. Subdimensión: Año

Año en el que la película fue premiada.

1. "1972"
2. "1960"

3.13.2. Subdimensión: Institución

Institución emisora de los premios que ganó la película.

1. "Oscars", "Globos de Oro", "Premios BAFTA", "Círculo de Críticos de Nueva York", "National Board of Review", "Sindicato de Directores (DGA)", "Sindicato de Guionistas (WGA)", "Premios David di Donatello"
2. "Oscars", "Globos de Oro", "Sindicato de Directores (DGA)", "Sindicato de Guionistas (WGA)"

3.13.3. Subdimensión: Categorías

Categorías en las que la película fue premiada.

1. "Mejor película, Actor (Marlon Brando), Guión adaptado, 10 nominaciones", "Película (Drama), Director, Actor (Brando), Guión y BSO", "Mejor música. 5 nominaciones, incluyendo Mejor actor (Brando)", "Mejor actor secundario (Duvall). 4 nominaciones", "Mejor actor sec. (Pacino) y Mejores 10 films del año", "Mejor director", "Mejor guión adaptado drama", "Mejor film extranjero y Premio Especial (Al Pacino)"
2. "4 nominaciones. director, actriz sec. (Leigh), fotografía y dir. artística (BN)", "Mejor actriz secundaria (Janet Leigh)", "Nominada a Mejor director", "Nominada a Mejor guión drama"

3.14. Dimensión: Críticas

3.14.1. Subdimensión: Crítica

Contenido de la crítica *insitu*

1. "'El padrino' son palabras mayores. Las dos primeras partes están entre las 10 mejores películas de la historia del cine.", "Coppola inventa una nueva mirada para el cine y amplía los horizontes de una industria que pedía a gritos savia nueva.", "I believe in America... Así comienza este fascinante compendio de cine de Coppola. La compleja historia de una familia mafiosa liderada por un implacable padre de familia y hombre de honor (inmenso Marlon Brando) y cuyo poder hereda su hijo más pacífico (asombroso un primerizo Al Pacino de mirada gélida) es, ante todo, un ejercicio narrativo apabullante, de insuperable nivel. La brillante disección de todos los personajes, el ritmo magistral -que alterna largas secuencias familiares con creíbles escenas de acción- y una ambientación perfecta consiguen un film que entró 'violentamente' entre los mejores clásicos de todos los tiempos, para alzarse con el título de obra cumbre del cine moderno.", "Inicialmente la Academia anunció, en febrero de 1973, que 'The Godfather' tenía 11 nominaciones, más que cualquier otra película de ese año. Pero la cifra se redujo finalmente a 10 nominaciones (empatando con 'Cabaret' como la película con más nominaciones del año) tras una nueva votación por parte de los responsables de la Academia de nominar a mejor música, con motivo de una controversia sobre si la banda sonora de Nino Rota de 'The Godfather' era elegible para la nominación que recibió. Así que la nominación fue retirada y reemplazada por una nominación para 'Sleuth'.", "La secuencia de la boda es un momento cinematográfico virtuoso: Coppola lleva a su gran reparto con tanta destreza que nos hace entrar completamente en el mundo de 'El Padrino' (...) Puntuación: 4 (sobre 4)", "Brando hizo de Don Vito algo que rara vez vemos en las películas: un villano-héroe tragicómico.", "Una de las crónicas más brutales y conmovedoras de la vida americana que se haya diseñado nunca dentro de los límites del entretenimiento popular.", "Tiene una excelente producción, a ratos es emocionante, y un reparto bien escogido. Pero es también demasiado larga (...) y ocasionalmente confusa. Aunque nunca llega a ser aburrida, no fascina tanto como para ser un drama superior.", "Una gran película americana, llena de imágenes increíbles y momentos perdurables.", "Una obra de arte que perdura el paso de los años. Puntuación: 5 (sobre 5)"
2. "Una obra maestra del entretenimiento y el horror de la que Perkins nunca pudo huir", "Una de las grandes obras maestras del género. Tensión, terror e intriga en un filme de imprescindible revisión", "Lo que hace inmortal a 'Psycho', cuando muchas otras películas las hemos casi olvidado tras salir del cine, es que conecta directamente con nuestros miedos

(...) Puntuación: 4 (sobre 4)”, “Marca un antes y un después en el género.(...) ésta es de las películas que se dejan ver más de una, dos y hasta tres veces, o más.”, “Un buen e inusual entretenimiento, con la huella imborrable de Hitchcock.”, “Sentimos que sus explicaciones parecen bromas de un hombre conocido por recurrir a esas tácticas en películas anteriores. La consecuencia es que su desenlace fracasa”, “Con ‘Psicosis’, Alfred Hitchcock no sólo creó una implacable obra maestra y engendró un nuevo género cinematográfico – el slasher. También dio uno de los golpes más audaces en la historia del cine (...) Puntuación: 5 (sobre 5)”, “‘Psycho’ no es una película larga, pero lo parece. Quizás porque el director se entretiene con efectos técnicos; quizás porque es difícil, si no imposible, que nos interese alguno de los personajes.”, “El suspense de su película crece lentamente hasta un punto casi insoportable de excitación.”, “Su poder no es sólo el de una máquina de terror calibrada por un showman, sino el de una fuga sombría acerca de las criaturas atrapadas del siglo XX que habitan este mundo (...) Puntuación: 4 (sobre 4)”, “Un clásico eterno. Excelentes actuaciones y la infame escena de la ducha hacen de ella la pesadilla perfecta (...) Puntuación: 5 (sobre 5)”, “‘Psycho’ no sólo reubicó el terror, de Transylvania al corazón de la familia americana, sino que fue increíblemente irónica de principio a fin.”, “Un auténtico clásico de clásicos.”

3.14.2. Subdimensión: Autores

Autores de las críticas

1. “Carlos Boyero: Diario El Mundo”, “Luis Martínez: Diario El País”, “Pablo Kurt: FilmAffinity”, “FilmAffinity”, “Roger Ebert: rogerebert.com”, “Michael Wilmington: Chicago Tribune”, “Vincent Canby: The New York Times”, “A.D. Murphy: Variety”, “Desson Thomson: The Washington Post”, “César Albarrán Torres: Cine Premiere”.
2. “Fernando Morales: Diario El País”, “Roger Ebert: rogerebert.com”, “Lucero Solórzano: Diario Excelsior”, “Variety”, “Bosley Crowther: The New York Times”, “Mark Monahan: Telegraph”, “CA Lejeune: The Guardian”, “New York Daily News”, “Bill Weber: Slant”, “David Parkinson: Empire”, “J. Hoberman: Village Voice”, “Diego Brodersen: Diario Página 12”

4. Creación de preguntas

Obtenidas las dimensiones, el siguiente paso va a ser definir las preguntas que va a aceptar nuestro sistema. Para ello, vamos a trabajar sobre cada una de las dimensiones añadiendo una serie de preguntas admisibles. Además, realizaremos un parafraseo de las preguntas para prevenir posibles variaciones de preguntas que cuyo significado es el mismo.

Para realizar este parafraseo, vamos a mantener siempre la estructura de una *pregunta base*, sobre la cuál vamos a establecer una serie de variaciones. Como bien hemos dicho, queremos mantener al máximo la estructura de la pregunta original, por lo que las variaciones que realicemos serán del tipo:

Pregunta original: ¿Cuándo se estrenó *El Padrino*?

Variación: ¿Cuándo salió *El Padrino*?

En ambas preguntas se mantiene la estructura: Cuándo + verbo + película.

Una variación muy típica que podemos encontrar es la aparición u omisión de “la película”, ya que se puede añadir en todas las ocasiones antes del nombre de la misma o puede ser omitida. Para evitar crear una variación que contenga esta secuencia, ya que la trataremos como irrelevante, vamos a añadirla entre corchetes [la película], tanto en la pregunta original como en las variaciones de la misma, indicando que es un valor que puede tanto aparecer como ser omitido. De igual forma, algunos de los caracteres que pueden aparecer o no los indicaremos de igual forma entre corchetes.

Una vez explicado el procedimiento que vamos a utilizar para obtener las variaciones, vamos a mostrar a continuación aquellas preguntas y sus variaciones, que aparecen como una lista numerada (a), b), c), ...) tras la pregunta original, que se van a efectuar sobre cada una de las dimensiones. Como ejemplo, trabajaremos sobre la película *El padrino*.

4.1. Dimensión título

1. ¿Cual es el nombre original de [la película] El padrino?
 - a) ¿Cual es el título original de [la película] El padrino?

4.2. Dimensión estreno

1. ¿Cuándo se estrenó [la película] El padrino?
 - a) ¿Cuándo salió [la película] El padrino?
2. ¿Se estrenó [la película] El padrino en [el año] 1960?
 - a) ¿Salió [la película] El padrino en [el año] 1960?
3. ¿Se estrenó [la película] El padrino antes de 1960?
 - a) ¿Salió [la película] El padrino antes de 1960?
4. ¿Se estrenó [la película] El padrino después de 1960?
 - a) ¿Salió [la película] El padrino después de 1960?

4.3. Dimensión duración

1. ¿Cuánto [tiempo] dura [la película] El padrino?
2. ¿Dura [la película] El padrino más de 110 minutos?
3. ¿Dura [la película] El padrino menos de 110 minutos?

4.4. Dimensión dirección

1. ¿Quién es el director de [la película] El padrino?
 - a) ¿Quién dirigió [la película] El padrino?
2. ¿Dirigió *Coppola* [la película] El padrino?

4.5. Dimensión guión

1. ¿Quién es el guionista de [la película] El padrino?
2. ¿Es *Coppola* el guionista de [la película] El padrino?
3. ¿Cuántos guionistas hay en [la película] El padrino?

4.6. Dimensión música

1. ¿Quién compuso la banda sonora de [la película] El padrino?

4.7. Dimensión reparto

1. ¿Qué actores participaron en [la película] El padrino?
 - a) ¿Qué actores tiene [la película] El padrino?
 - b) ¿Qué actores actuaron en [la película] El padrino?
2. ¿Quiénes son los 3 actores principales de [la película] El padrino?
3. ¿Participó *Capone* en [la película] El padrino?
 - a) ¿Actuó *Capone* en [la película] El padrino?

4.8. Dimensión productora

1. ¿Cuál es la productora de [la película] El padrino?

4.9. Dimensión género

1. ¿A qué género[s] pertenece [la película] El padrino?
 - a) ¿A qué categoría[s] pertenece [la película] El padrino?
 - b) ¿De qué género[s] es [la película] El padrino?
 - c) ¿De qué categoría[s] es [la película] El padrino?
2. ¿Es [la película] El padrino [una película] de genero terror?

4.10. Dimensión grupos

1. ¿A qué saga pertenece El padrino?
2. ¿Pertenece El padrino a la saga Trilogía El Padrino?

4.11. Dimensión sinopsis

1. ¿Cuál es la sinopsis de [la película] El padrino?
 - a) ¿Cuál es el resumen de [la película] El padrino?

4.12. Dimensión valoración

1. ¿Qué valoración tiene [la película] El padrino?
 - a) ¿Qué puntuación tiene [la película] El padrino?
2. ¿Tiene [la película] El padrino más valoración que 7?
 - a) ¿Tiene [la película] El padrino más puntuación que 7.5?
3. ¿Tiene [la película] El padrino menos valoración que 7?
 - a) ¿Tiene [la película] El padrino menos puntuación que 7.5?
4. ¿Cuánta gente ha valorado [la película] El padrino?
 - a) ¿Cuánta gente ha puntuado [la película] El padrino?
 - b) ¿Cuántas personas han valorado [la película] El padrino?
 - c) ¿Cuántas personas han puntuado [la película] El padrino?

4.13. Dimensión premios

1. ¿Qué premios ha ganado [la película] El padrino?
 - a) ¿Qué premios ha recibido [la película] El padrino?
2. ¿Cuántos premios ha ganado [la película] El padrino?
 - a) ¿Cuántos premios ha recibido [la película] El padrino?
3. ¿Ha ganado [la película] El padrino más de 3 premios?
 - a) ¿Ha recibido [la película] El padrino más de 3 premios?
4. ¿Ha ganado [la película] El padrino menos de 3 premios?
 - a) ¿Ha recibido [la película] El padrino menos de 3 premios?
5. ¿Cuándo ha ganado premios [la película] El padrino?
 - a) ¿Cuándo ha recibido premios [la película] El padrino?
 - b) ¿En qué años ha ganado premios [la película] El padrino?
 - c) ¿En qué años ha recibido premios [la película] El padrino?

6. ¿Ha ganado algún premio [la película] El padrino?
 - a) ¿Ha recibido algún premio [la película] El padrino?
7. ¿Ha ganado algún premio [la película] El padrino en [el año] 1960?
 - a) ¿Ha recibido algún premio [la película] El padrino en [el año] 1960?

4.14. Dimensión críticas

1. ¿Cuántas críticas tiene [la película] El padrino?
 - a) ¿Cuántas reseñas tiene [la película] El padrino?
 - b) ¿Cuántas críticas ha recibido [la película] El padrino?
2. ¿Quiénes son los autores de las críticas de [la película] El padrino?
 - a) ¿Quiénes son los críticos de [la película] El padrino?

5. Obtención de reglas

Una vez hemos formulado las preguntas, tenemos que identificar la estructura por la cuál podemos identificar una determinada pregunta y, por tanto, dar una respuesta a la misma. Para ello, tenemos que definir un conjunto de reglas que identifiquen a todas las preguntas que tengan una misma respuesta. Gracias al parafraseo que hemos realizado, en el que mantenemos una misma estructura, la creación de estas reglas va a resultar mucho más sencilla.

Estructura En el momento de desarrollar las reglas, hemos pensado en dos posibles enfoques: utilización de una **estructura exacta** o una **estructura libre**.

Estructura exacta La utilización de una estructura de reglas exacta va a identificar de manera inequívoca todas las preguntas. Esto va a permitir que no encontremos colisiones entre dos reglas independientemente del nivel al que estén situadas. Esto es, si introdujésemos la pregunta *¿dónde en que año se estrenó El Padrino?*, no vamos a encontrar ninguna regla que sea capaz de identificarla correctamente y dar una respuesta.

Sin embargo, el uso de estructuras exactas nos limita las preguntas a una exactitud que puede provocar una experiencia de usuario no satisfactoria, al ser incapaces de reconocer las preguntas con pequeñas variaciones sobre la estructura de la regla y no reconocer preguntas con mayor facilidad.

Estructura libre Plantear una estructura libre, donde las reglas reconocen elementos característicos de la pregunta y no la estructura total, nos sitúa en una tesitura donde no encontramos sensibilidad al ruido. Reduciendo la pregunta a una estructura básica, podemos admitir la presencia de ruido (como puede ser una sintaxis incorrecta de la pregunta, palabras conjuntas, adición de caracteres incorrectos al final de palabras...) y nuestro sistema será capaz de responder correctamente a la pregunta. En el siguiente ejemplo, mostramos cómo nuestro sistema identifica dos preguntas como iguales siguiendo unas reglas de estructura libre, dando la misma respuesta válida para cada una de ellas.

Pregunta correcta: ¿En qué año se estrenó El Padrino? -> 1972

Pregunta con ruido: ¿En qué año aaa se estrenó aaabn bdfbd El Padrino? jajaja -> 1972

Por otra parte, siguiendo esta estructura libre es más posible la aparición de colisiones entre las reglas. Siguiendo este ejemplo, si tenemos una regla que identifique “en qué año” (cuándo) se estrenó y otra regla que identifique “dónde” se estrenó, estamos permitiendo preguntas como *¿dónde en que año se estrenó El Padrino?*, ante la cuál nuestro sistema devolverá siempre la primera regla detectada.

Tras un breve análisis, entendemos que el modelo adecuado implicaría un modelo híbrido que permitiese pequeñas variaciones sobre una pregunta de forma que podamos seguir utilizándola, pero donde llegado un punto, la aparición de ruido en la pregunta haga que no se reconozca la misma: podríamos reconocer *¿En qué año se estrenó El Padrino?* pero nunca reconoceríamos *¿En qué año aaa se estrenó aaabn bdfbd El Padrino?* jajaja. Sin embargo, el planteamiento de este sistema híbrido complica demasiado el sistema de reglas, por lo que hemos optado por permitir la aparición de ruido y optar por una **estructura libre**.

Input En el momento que se quiera indicar un valor que puede ser variable en la entrada, lo indicaremos añadiendo delante del mismo un símbolo de \$. Se añadirá además, en cada caso, una nota al pie de página realizando una breve explicación de dicha variable.

Nombre película En cada una de las reglas que se muestran a continuación, se debe incluir una cláusula conjuntiva (**and** nombre_pelicula), la cuál vamos a omitir para facilitar la visualización de las reglas, que nos permitirá indicar la película a la que se hace referencia.

Elementos de las reglas La estructura de nuestras reglas va a contener los siguientes elementos:

- El antecedente de la regla, es decir, el conjunto de premisas que identifican una regla, comenzará con la palabra en rojo **IF**.
- El consecuente de la regla, es decir, el valor que vamos a devolver, se indicará a continuación de la palabra en rojo **THEN**.
- El operador lógico *AND* se indicará con un color azul claro.
- El operador lógico *OR* se indicará con un color azul oscuro.
- Elementos opcionales del antecedente se indican entre llaves [].
- El valor del consecuente siempre hace referencia al valor de la dimensión, indicada entre llaves a continuación de *valor*, para la película que se debe introducir como *nombre_pelicula*.
- Aquellas palabras que se escriben con acento vamos a permitir tanto una escritura con acento como una escritura sin acento.
- En aquellas preguntas donde pueden existir variaciones morfológicas de la palabra, como puede ser *actuó* y *actúa*, incluiremos la raíz de las mismas como elemento de la regla.
- Indicaremos la inclusión de un valor en otro (*Coppola* está incluido en *Francis Ford Coppola*) mediante el exto en color verde **in**.

Siguiendo esta estructura definida, mostramos a continuación el conjunto de reglas obtenidas para cada una de las dimensiones.

5.1. Dimensión título

1. **IF** cual **and** (nombre **or** titulo) **and** original
THEN valor[título]

5.2. Dimensión estreno

1. **IF** cuando **and** (se estreno **or** salio)
THEN valor[estreno]
2. **IF** (se estreno **or** salio) **and** en [el año] \$año²
THEN valor[estreno] == \$año

²Variable que hace referencia a un año numérico. Ejemplo: 1997, 2020...

3. IF (se estreno or salio) and antes de [el año] \$año
THEN valor[estreno] <= \$año
4. IF (se estreno or salio) and despues de [el año] \$año
THEN valor[estreno] >= \$año

5.3. Dimensión duración

1. IF cuanto and dura
THEN valor[duración]
2. IF dura and mas de and \$minutos³
THEN valor[duración] > \$minutos
3. IF dura and menos de and \$minutos
THEN valor[duración] < \$minutos

5.4. Dimensión dirección

1. IF quien and (director or dirig)
THEN valor[dirección]
2. IF dirig \$nombre⁴
THEN \$nombre in valor[dirección]

5.5. Dimensión guión

1. IF quien and guionista
THEN valor[guión]
2. IF ser⁵ and \$nombre el guionista
THEN \$nombre in valor[guión]
3. IF cuantos and guionistas
THEN contar(valor[guión])

5.6. Dimensión música

1. IF quien and compuso and banda sonora
THEN valor[música]

5.7. Dimensión reparto

1. IF que and (actor or actri) and (tiene or particip or actua)
THEN valor[reparto]
2. IF quienes and \$num_actores⁶ and actores principales
THEN primeros_\$num_actores(valor[reparto])

³Variable que hace referencia a una cantidad de tiempo expresada en minutos. Ejemplo: 60,120,150...

⁴Esta variable hace referencia realmente a cualquier palabra que continúe a una palabra cuya raíz es *dirig*

⁵Admite las siguientes variantes del verbo ser: es, fue, y ha sido

⁶Variable que hace referencia al número de actores que se mostrarán en la respuesta

3. IF (participo \$nombre or actuo \$nombre)
THEN \$nombre in valor[reparto]

5.8. Dimensión productora

1. IF cual and productora
THEN valor[productora]

5.9. Dimensión género

1. IF que and (genero or categoria)
THEN valor[género]
2. IF ser and genero \$nombre
THEN \$nombre in valor[género]

5.10. Dimensión grupos

1. IF que and saga
THEN valor[grupos]
2. IF (pertenece or es) and saga \$nombre_saga⁷
THEN \$nombre_saga in valor[grupos]

5.11. Dimensión sinopsis

1. IF cual and (resumen or sinopsis)
THEN valor[sinopsis]

5.12. Dimensión valoración

1. IF que and (valoracion or puntuacion)
THEN valor[valoración][puntuación]
2. IF tiene mas (valoracion or puntuacion) de un \$puntuacion⁸
THEN valor[valoración][puntuación] > \$puntuacion
3. IF tiene menos (valoracion or puntuacion) de un \$puntuacion⁹
THEN valor[valoración][puntuación] < \$puntuacion
4. IF ((cuanta and gente) or (cuantas and personas)) and (puntu or valor)
THEN valor[valoración][votos]

⁷En esta ocasión, esta variable cogerá cualquier palabra hasta el final de la pregunta

⁸Variable que hace referencia a una posible nota de la película, correspondiente a un valor entero. Ejemplo: 5, 10, 2...

⁹Variable que hace referencia a una posible nota de la película, correspondiente a un valor entero. Ejemplo: 5, 10, 2...

5.13. Dimensión premios

1. IF que and premios
THEN concatenar(valor[premios][institución], valor[premios][categorías])
2. IF cuantos and premios
THEN contar_premios(valor[premios][institución])¹⁰
3. IF mas de \$numero premios
THEN contar_premios(valor[premios][institución]) > \$numero premios
4. IF menos de \$numero premios
THEN contar_premios(valor[premios][institución]) < \$numero premios
5. IF (cuando or en que años) and premios
THEN valor[premios][año]
6. IF (gan or recib) and algun premio
THEN existe(valor[premios][año])
7. IF (gan or recib) and algun premio and en \$año
THEN \$año in valor[premios][año]

5.14. Dimensión críticas

1. IF cuantas and (reseñas or criticas)
THEN contar(valor[criticas])
2. IF quienes and ((autores and criticas) or criticos)
THEN contar(valor[criticas][autor])

¹⁰Por la forma de obtener los datos y guardarlos, el valor que contiene la dimensión institución incluye también el número de premios ganados en la misma. Ejemplo: 3 Oscars, 2 Globos de Oro...

6. Implementación

Una vez diseñadas las reglas de nuestro sistema es el momento de implementarlo. Nuestro sistema cuenta con las siguientes características:

- Implementación en **Python**.
- **Interfaz básica por comandos** que nos permite **preguntar ilimitadamente** a nuestro sistema sin tener que volver a ejecutarlo.
- Opción de **guardar** las preguntas y respuestas en un archivo de texto a parte una vez finalizado el proceso de preguntar.
- Se ha implementado de tal forma que si una pregunta contiene ruido, nuestro sistema sigue siendo capaz de dar la respuesta.

Ejemplo: *¿En qué año se estrenó abcd el padrino real Madrid? 1972*

A continuación, iremos explicando las funcionalidades que se encuentran implementadas en cada archivo:

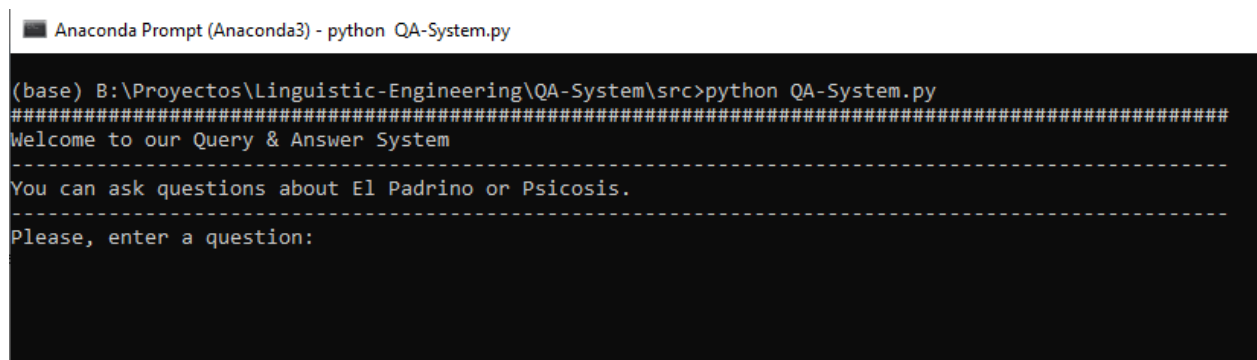
- **QA-System**: este archivo contiene la lógica general de la aplicación, permitiendo preguntar, mostrando las respuestas y preguntando al final al usuario si desea guardar los resultados.
- **DocumentProcessor**: en este archivo está implementado todo lo necesario para preprocesar las fichas de las películas, obteniendo los valores de las dimensiones explicadas en el Capítulo 3.
- **QuestionParser**: en este archivo se implementa toda la lógica de procesar las preguntas para determinar qué se está preguntando concretamente. Por lo tanto, para cada dimensión se comprueban las reglas explicadas en el Capítulo 5. Para aquellas reglas en las que era necesario obtener un valor introducido por el usuario (como en *¿ganó premios en 1992?*) se han usado expresiones regulares mediante la librería *re*¹¹ de Python.

Es importante destacar que a pesar de que en el capítulo anterior las reglas no seguían un orden concreto, a la hora de implementarlas tenemos que poner primero aquellas más complejas para así evitar errores.

- **Question**: este archivo contiene las clases necesarias para representar las preguntas propuestas en el Capítulo 4. Se ha diseñado de tal forma que si se quiere añadir una nueva pregunta no contemplada anteriormente, únicamente tendremos que añadir una nueva clase o subclase que represente a esa nueva pregunta, sin verse afectada el resto de la implementación.
- **Writer**: este archivo es el encargado de escribir en un fichero de texto el resumen de las preguntas y respuestas.

Cómo se vería nuestra interfaz se puede comprobar en la Figura 3. En ella podemos ver cómo se indica sobre qué películas puede el usuario preguntar y se indica al usuario que introduzca una pregunta.

¹¹<https://docs.python.org/3/library/re.html>



```
Anaconda Prompt (Anaconda3) - python QA-System.py
(base) B:\Proyectos\Linguistic-Engineering\QA-System\src>python QA-System.py
#####
Welcome to our Query & Answer System
-----
You can ask questions about El Padrino or Psicosis.
-----
Please, enter a question:
```

Figura 3: Visualización de la interfaz por consola.

7. Experimentación

Tras la implementación, mostraremos con algunos ejemplos las preguntas que nuestro sistema es capaz de responder. En los dos primeros ejemplos podemos comprobar cómo el sistema es capaz de distinguir sobre qué película se está preguntando. En los siguientes ejemplos se muestra la respuesta para una pregunta de cada dimensión. Finalmente, en los dos últimos ejemplos podemos ver cómo se comporta nuestro sistema en las dos posibles situaciones de fallo: se pregunta algo sobre otra película externa a nuestro sistema y se pregunta algo sobre una película que nuestro sistema no sabe responder.

- ¿Cuál es el título original de El Padrino?

```
-----  
Please, enter a question: ¿Cuál es el título original de El Padrino?  
Answer: The Godfather  
-----
```

Figura 4: Pregunta de ejemplo: título.

- ¿Cuál es el título original de la película Psicosis?

```
-----  
Please, enter a question: ¿Cual es el nombre original de la película Psicosis?  
Answer: Psycho  
-----
```

Figura 5: Pregunta de ejemplo: título 2.0 .

- ¿Cuándo se estrenó El Padrino?

```
-----  
Please, enter a question: ¿Cuándo se estrenó El Padrino?  
Answer: 1972  
-----
```

Figura 6: Pregunta de ejemplo: estreno.

- ¿Dura Psicosis más de 500 minutos?

```
-----  
Please, enter a question: ¿Dura Psicosis más de 500 minutos?  
Answer: No  
-----
```

Figura 7: Pregunta de ejemplo: duración.

- ¿Quién es el director de la película El Padrino?

```
-----  
Please, enter a question: ¿Quién es el director de la película El Padrino?  
Answer: Francis Ford Coppola  
-----
```

Figura 8: Pregunta de ejemplo: dirección.

- ¿Cuántos guionistas hay en Psicosis?

```
Please, enter a question: ¿Cuántos guionistas hay en Psicosis?  
Answer: 1
```

Figura 9: Pregunta de ejemplo: guión.

- ¿Quién compuso la banda sonora de El Padrino?

```
Please, enter a question: ¿Quién compuso la banda sonora de El Padrino?  
Answer: Nino Rota
```

Figura 10: Pregunta de ejemplo: música.

- ¿Quiénes son los 3 actores principales de Psicosis?

```
Please, enter a question: ¿Quiénes son los 3 actores principales de Psicosis?  
Answer: Anthony Perkins, Janet Leigh, John Gavin
```

Figura 11: Pregunta de ejemplo: reparto.

- ¿Quién compuso la banda sonora de El Padrino?

```
Please, enter a question: ¿Quién compuso la banda sonora de El Padrino?  
Answer: Nino Rota
```

Figura 12: Pregunta de ejemplo: música.

- ¿Cuál es la productora de El Padrino?

```
Please, enter a question: ¿Cuál es la productora de El Padrino?  
Answer: Paramount Pictures, Alfran Productions
```

Figura 13: Pregunta de ejemplo: productora.

- ¿Es Psicosis de género drama?

```
Please, enter a question: ¿Es Psicosis de género drama?  
Answer: No
```

Figura 14: Pregunta de ejemplo: género.

- ¿A qué saga pertenece El Padrino?


```
Please, enter a question: ¿A qué saga pertenece El Padrino?  
Answer: Trilogía El Padrino | Adaptaciones de Mario Puzo
```

Figura 15: Pregunta de ejemplo: saga.

- ¿Cuál es el resumen de la película Psicosis?

```
Please, enter a question: ¿Cuál es el resumen de la película Psicosis?  
Answer: Marion Crane, una joven secretaria, tras cometer el robo de un dinero en su empresa, huye de la ciudad y, después de conducir durante horas, decide descansar en un pequeño y apartado motel de carretera regentado por un tímido joven llamado Norman Bates, que vive en la casa de al lado con su madre. (FILMAFFINITY)
```

Figura 16: Pregunta de ejemplo: sinopsis.

- ¿El Padrino tiene más valoración de un 7?

```
Please, enter a question: ¿El Padrino tiene más valoración de un 7?  
Answer: Sí
```

Figura 17: Pregunta de ejemplo: valoración.

- ¿Quiénes son los autores de las críticas de Psicosis?

```
Please, enter a question: ¿Quiénes son los autores de las críticas de Psicosis?  
Answer: Javier Ocaña: Cinemanía, Fernando Morales: Diario El País, Roger Ebert: rogerebert.com, Lucero Solórzano: Diario Excelsior, Variety, Bosley Crowther: The New York Times, Mark Monahan: Telegraph, CA Lejeune: The Guardian, New York Daily News, Bill Weber: Slant, David Parkinson: Empire, J. Hoberman: Village Voice, Diego Brodersen: Diario Página 12
```

Figura 18: Pregunta de ejemplo: críticas.

- ¿Quién dirigió La Vida es Bella?

```
Please, enter a question: ¿Quién dirigió La Vida es Bella?  
Answer: We don't have answers for that film. Please introduce a correct film.
```

Figura 19: Pregunta de ejemplo: película desconocida.

- ¿Cuántas entradas se vendieron de El Padrino?

```
Please, enter a question: ¿Cuántas entradas se vendieron de El Padrino?  
Answer: We don't have an answer for you question.
```

Figura 20: Pregunta de ejemplo: pregunta desconocida.

Cómo puede usted utilizar nuestro sistema se explica en el manual de usuario que se encuentra en el Anexo [A](#).

8. Conclusiones

Limitación de las preguntas El conjunto de preguntas que se puede realizar sobre una fuente de información de este tipo podría ser casi infinita si se incluyese una combinación de preguntas donde entrasen en juego más de una dimensión, como por ejemplo la pregunta: *¿Participó el director de la película en la misma?* o, siendo un poco más rebuscados, *¿Hay alguna crítica que mencione directamente a alguno de los 3 actores principales y al propio director?*. Dadas estas combinaciones casi infinitas, hemos optado por realizar preguntas de forma distintiva para cada una de las dimensiones, teniendo que reducir alguna de las preguntas que planteamos inicialmente dada su complejidad.

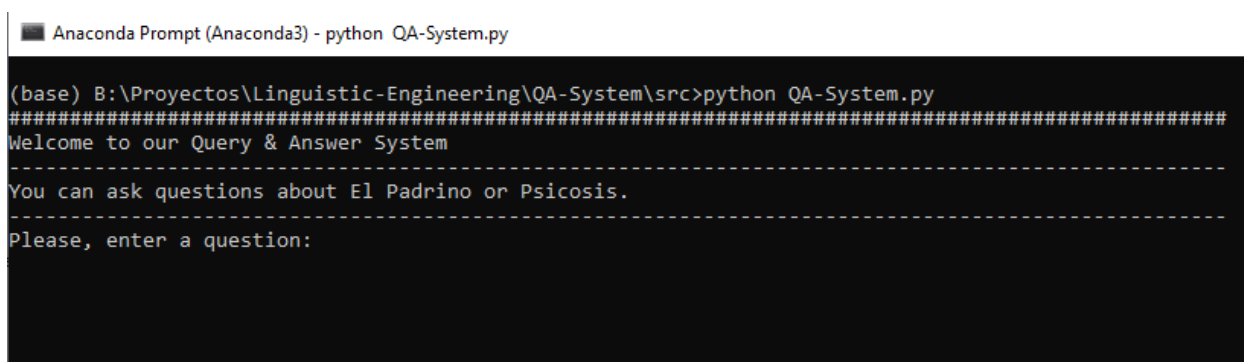
Flexibilidad en las preguntas La estructura libre utilizada en nuestras reglas, que emplean operadores lógicos, no tiene en cuenta el orden en el que las cláusulas son introducidas. Esto causa que, por ejemplo, la pregunta *“¿Cuál es el título original de El Padrino?”* sea tratada de igual forma, y por tanto obtenga la misma respuesta, que si introdujésemos la pregunta *“¿Título original de Cuál tiene El Padrino?”*. Además, únicamente estamos trabajando con determinadas palabras de las preguntas (las correspondientes a la estructura de la regla en cuestión), por lo que mientras éstas aparezcan, el resto del texto introducido puede contener cualquier palabra o carácter, ya que se identificará de igual forma la regla dentro de la pregunta, permitiendo así la aparición de ruido en las preguntas y haciendo más flexibles las preguntas introducidas.

Automatización El sistema desarrollado es capaz de responder correctamente a todas las preguntas expuestas en el Capítulo 4. Sin embargo, una limitación que hemos tenido a la hora de crear las reglas, es que las hemos creado de manera manual. Esto obliga a que, si queremos introducir un conjunto nuevo de preguntas con sus respectivas reglas, el proceso puede ser tedioso y complicado cuando tenemos una cantidad razonable de reglas ya que puede implicar superposición de las mismas, haciendo que el sistema pueda responder de manera incorrecta. El uso de técnicas que, partiendo de un conjunto de preguntas que consideremos que son iguales (pregunta y parafraseo de la misma) pudiese obtener y añadir una regla nueva a nuestro sistema de reglas, facilitaría mucho trabajar en entornos menos controlados.

A. Manual de uso

A continuación, procederemos a detallar los pasos necesarios para poder ejecutar nuestro proyecto haciendo uso de Python3.x.

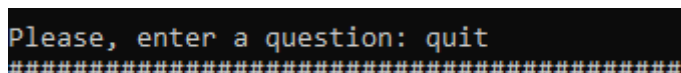
1. Nosotros le recomendamos instalar la versión Python3.7. En este [enlace](#) puede encontrar una guía detallada de cómo instalarlo en Windows.
2. Instalado Python3.7, abra una consola de comandos.
3. Antes de abrir el proyecto, deberá instalar la librería NumPy, para lo cuál puede ejecutar el comando `pip install numpy`:
4. Entre en la carpeta que le hemos entregado.
5. Navegue hasta el directorio `/src`.
6. Ejecute el comando `python QA-System.py`.



```
Anaconda Prompt (Anaconda3) - python QA-System.py
(base) B:\Proyectos\Linguistic-Engineering\QA-System\src>python QA-System.py
#####
Welcome to our Query & Answer System
-----
You can ask questions about El Padrino or Psicosis.
-----
Please, enter a question:
```

Figura 21: Abriendo la aplicación.

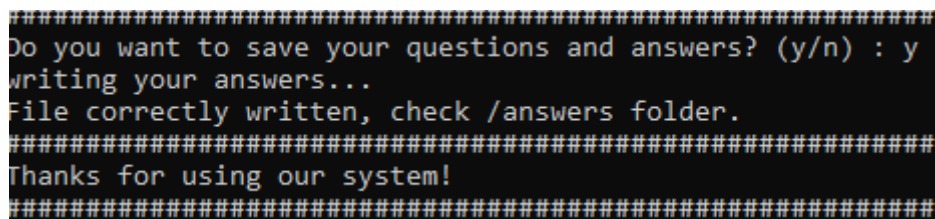
7. Realice las preguntas que desee (No es necesario introducir símbolos de interrogación).
8. Una vez que no quiera preguntar más, deberá introducir *quit*.



```
Please, enter a question: quit
#####
```

Figura 22: Punto en el que no se desea preguntar más.

9. Se le preguntará si desea guardar sus preguntas, introduzca *y* o *n*.



```
#####
Do you want to save your questions and answers? (y/n) : y
writing your answers...
File correctly written, check /answers folder.
#####
Thanks for using our system!
#####
```

Figura 23: Punto en el que puede guardar las preguntas.

10. Si ha guardado las preguntas, podrá visualizarlas en la carpeta */answers*.

Si tiene cualquier problema o duda a la hora de seguir este manual, póngase en contacto con nosotros mediante los siguientes correos electrónicos: *cesar.garcia.cabeza@alumnos.upm.es* o *a.gabarre@alumnos.upm.es*.