Table of Contents

Informe Inteligencia Artificial Avanzada

> Descripción de la implementación, por **Joaquín Sanchiz** y **Ángel Hamilton.**

Lenguaje

El lenguaje que hemos utilizado en esta práctica ha sido el lenguaje Java. Hemos seleccionado este lenguaje por razones claras: ha sido el lenguaje de programación que más hemos trabajado a lo largo de este cuatrimestre, por lo que tenemos bastante agilidad y facilidad de implementación con el; el manejo de cadenas de caracteres, algo que se utiliza constantemente en esta práctica, resulta muy cómodo y nos ha ahorrado muchos problemas a la hora de trabajar con ellas, sobre todo con temas de buffers de entrada/salida; el manejo de excepciones y la facilidad de transformar objetos nos ha resuelto bastante la práctica, por ejemplo a la hora de ordenar las tablas hash.

Estructura de los programas

Primer programa

En el primer programa, la generación de los corpus fue llevado a cabo por un programa compuesto por un único objeto: Vocabulario. Este objeto tiene un constructor que recibe como entrada un fichero de preguntas, va analizando cada pregunta rompiendo las líneas de entrada (por espacios) en tokens. Dichos tokens se insertan posteriormente en una tabla hash, que recogerá el vocabulario con el número de apariciones de cada palabra. Vocabulario también tiene un método de escritura del vocabulario en fichero, de tal manera que la salida del vocabulario queda ordenada por el número de apariciones de la palabra. Mediante estos métodos hemos generado los **corpus** correspondientes, como salida del programa.

Segundo programa

El segundo programa funcionaba de la siguiente manera: Primero con los métodos previamente explicados, se construye un objeto Vocabulario a partir de los corpus de entrada, y posteriormente sobre ese vocabulario, se aplica un método que representa el **suavizado laplaciano**.

Este método recibe como entrada un vocabulario con todas las palabras y el fichero sobre el que se escribirán los resultados. En un bucle se recorren todas las palabras del vocabulario y se comparan con las del corpus, viendo si dicha palabra del vocabulario aparece en el Vocabulario generado a partir del corpus. Con el número de palabras total en el vocabulario y el número de apariciones de la palabra analizada en nuestro Vocabulario, ya podemos realizar el logaritmo neperiano del suavizado laplaciano de la siguiente manera:

$${log((nApariciones+1) / (npalabras + nPalabrasVocabulario + 1))}$$

Con dicha ecuación, obtenemos el logaritmo y la salida del método para cada palabra es:

Palabra: <palabra_actual> Frec: <nApariciones> LogProb: <logaritmoSuavizado>

Tercer programa

El tercer y último entregable incluye una nueva clase “**Clasificación**” que será la encargada de analizar los textos y clasificarlos entre los 20 corpus posibles. Para esto,el programa se vale de un método “aprender” que almacena en una tabla las palabras de un corpus y el logaritmo de la probabilidad de que aparezcan en dicho corpus. Cada una de las veinte tablas las almacena en un Array.

Para analizar los textos, corpus por corpus accede a su tabla y calcula la probabilidad de que estas palabras aparezcan en ese corpus. Elige la mayor de todas ellas y ese es el corpus que le asigna. Estas frases con sus corpus asignados se escriben en un fichero y se guardan en una tabla. Finalmente, repasamos cada coprus individualmente y comprobamos si cada frase fue asignada correctamente para comprobar el porcentaje de éxito.

De un total de 19000 frases, el programa clasifica correctamente 16569 de ellas y 2431 las clasifica incorrectamente obteniendo un porcentaje de éxito de un 87.20526315789473 %.

Participación de cada miembro

Este proyecto ha sido desarrollado a partes iguales en su totalidad. En el primer programa, uno se encargó del método de lectura de ficheros y otro en el de escritura. En el segundo programa, uno desarrolló el método del suavizado laplaciano y el otro se encargó de la depuración y de construir los ficheros. Durante la tercera parte, Ángel realizó la mayor parte del código y colaboró un poco con el desarrollo del informe mientras que Joaquín escribió el grueso del informe y se encargó de la depuración del programa.