INFORME PŔACTICA 6

Javier Ramos Fernández e Isaac Aiman Salas

Introducción

El objetivo de esta práctica consiste en construir un sistema para **clasificar preguntas de usuarios en StackOverflow**. Para llevarlo a cabo, la realización de esta práctica se divide en tres partes. La primera parte consiste en crear una serie de corpus con las preguntas de cada categoría contenidas en un fichero Excel que se nos proporciona, además de crear el vocabulario a partir de todas las preguntas del corpus. La segunda parte consiste en estimar las probabilidades en el modelo del lenguaje; es decir, estimar las probabilidades de las palabras en cada uno de los corpus creados en la primera parte. Por último, la tercera parte consiste en clasificar las instancias de un corpus especificado dentro de una clase determinada a partir de las probabilidades calculadas en la segunda parte.

Para la consecución de todas las partes de la práctica, hemos optado por implementar su desarrollo con el **lenguaje de programación Ruby**. Al tener la práctica tres partes bien diferenciadas, hemos realizado un programa para realizar cada uno de los procesos (un programa para crear los corpus y otro para crear el vocabulario (primera parte); un programa para crear los ficheros de aprendizaje con las probabilidades (segunda parte) y un programa para clasificar las instancias de un corpus de entrada).

Programa corpus

El programa que construye los corpus de cada una de las clases a las que pertenecen las preguntas que se nos proporcionan utiliza una librería llamada ***roo***. Esta gema de Ruby provee una interfaz que implementa un acceso de lectura para todos los tipos comunes de hojas de cálculo. En este programa se ha utilizado para leer las preguntas del fichero Excel que se nos proporciona para construir los corpus y la construcción de los mismos se realiza a través de la apertura simultánea de tantos ficheros como clases haya. En función de la clase, se escribe cada pregunta del fichero Excel en un fichero u otro. Una vez construido los corpus, se añade de forma ordenada el contenido de cada uno de estos corpus en un fichero que representará el corpus en su totalidad. El orden de las clases se tiene en cuenta debido a la utilidad que presenta para una tarea posterior.

Programa vocabulario

A continuación, para crear el vocabulario del problema utilizamos el fichero generado con todas las preguntas de los corpus y seleccionamos todas las palabras menos aquellas que correspondan con ***stopwords*** (palabras que son muy comunes en los idiomas, en este caso inglés). De esta forma sólo tenemos en cuenta aquellas palabras que aportan significado y nos será útil para el posterior cálculo de la probabilidad de cada una de las palabras del vocabulario.

Programa aprendizaje

Después de generar el vocabulario, para calcular la probabilidad de cada una de las palabras del mismo en cada uno de los corpus, hemos utilizado una tabla *hash* en cada uno de ellos, con el objetivo de guardar como clave cada palabra del vocabulario y su número de apariciones. Una vez hecho esto, se ha utilizado estas apariciones para calcular las probabilidades de cada una de las palabras.

Programa clasificación

En la última parte de la práctica se construye una tabla *hash* por cada uno de los ficheros de aprendizaje, asociando cada palabra con su probabilidad. Esto se realiza para que el acceso a los valores de probabilidad de las distintas palabras sea lo más rápido posible. Una vez hecho esto, se calcula la probabilidad de cada una de las preguntas del corpus a clasificar. Después de clasificar cada una de las preguntas, para saber si se ha clasificado correctamente, se recorre el corpus total desde el principio. En este momento es donde el orden establecido en la primera parte cobra importancia ya que se conoce el número de preguntas de cada clase y las clases están ordenadas de menor a mayor, de modo que esto nos sirve para determinar si una instancia pertenece a la clase correcta.

Porcentaje de acierto para el corpusTodo.txt

El porcentaje de éxito para el *corpusTodo.txt* es: **94.07894736842105%**

Enlace al repositorio en GitHub

Todos los ficheros fuente pueden ser consultados en el siguiente repositorio:
https://github.com/alu0100841565/IAA-ClasificacionDeTextos

Desglose del trabajo realizado por cada miembro

Programa corpus:
Isaac: 30%
Javier: 70%

Programa corpus:
Isaac: 70%
Javier: 30%

Programa corpus:
Isaac: 50%
Javier: 50%

Programa corpus:
Isaac: 50%
Javier: 50%