PRÁCTICA 2: ECOSISTEMA HADOOP.

Memoria de la sesión segunda de laboratorio.
SING.

Resumen

Obtención de resultados referido al análisis de palabras de libros de grandes literatos españoles y obtención de conclusiones.

Javier Argente Micó y José Lluch Palop

Índice

Obtención de resultados.

Tras las ejecuciones pertinentes, se obtienen los resultados referidos al número total de palabras, promedio de las mismas y la desviación típica de las palabras.

Estos resultados, los vemos reflejados en la Tabla 1, asociada a los resultados extraídos de ejecuciones individuales para algunos libros de cada uno de los autores facilitados.

En segundo lugar, tenemos la Tabla 2, la cual hace referencia a ejecuciones, también individuales, pero en este caso respecto a todas las obras de un mismo autor. La razón de ser de esta segunda tabla es para conocer la visión de conjunto de un mismo autor y así poder obtener una perspectiva perimétrica.

Autor	Total Palabras	Longitud Media	Desviación Típica
Benito Pérez Galdós	66468	5,03015	3,13223
Benito Pérez Galdós	77848	4,62775	2,62321
Benito Pérez Galdós	85588	4,93136	3,06188
Leopoldo Alas Clarín	309473	4,96734	3,0593
Leopoldo Alas Clarín	92166	4,92968	3,05839
Blasco Ibáñez	93944	5,04289	3,09377
Blasco Ibáñez	84483	5,04425	3,10286
Blasco Ibáñez	59446	5,10894	3,11438
Blasco Ibáñez	207338	5,10688	3,14744
Miguel de Cervantes	383184	4,5956	2,74983
Miguel de Cervantes	58765	4,64812	2,63564

Tabla 1. Resultados de obras individuales.

Autor	Total Palabras Obras	Longitud Media Obras	Desviación Típica Obras
Benito Pérez Galdós	959084	4,89817	3,04859
Leopoldo Alas Clarín	401639	4,9585	3,05499
Blasco Ibáñez	1361980	5,11288	3,1246
Miguel de Cervantes	441949	4,60528	2,73498

Tabla 2. Resultados del conjunto de obras un mismo autor.

Conclusiones tras la obtención de resultados

Analizando el objetivo de la práctica, el cual consiste en dado un texto cualquiera, poder concretar a qué autor pertenece dicho texto, en función de resultados tales como el total de palabras, la longitud media de la palabra o la desviación típica; llegamos a la conclusión de que este objetivo no puede conseguirse con estos datos, ya que dentro de un mismo autor dichos valores oscilan dentro de un rango con diferencias notables de resultados unos respecto de otros, y en el caso de comparar los valores de un texto concreto con los valores medios sacados de analizar un conjunto de textos de un mismo autor, esta comparativa puede dar como resultado que el texto analizado pertenezca a otro autor que no sea el suyo, ya que los valores de un caso concreto puede variar bastante de la media de dicho autor y ser similar a la media de otro autor.

Como hemos nombrado anteriormente, la comparativa de los datos sacados de un caso concreto comparados con la media del autor no nos permitiría sacar con certeza a que autor pertenece el texto, por lo que una manera de mejorar la precisión de esta comparativa sería comparar los datos del texto concreto con los datos individuales de un conjunto de libros de dicho autor, además de con la media de dicho conjunto. Todas las comparativas antes nombradas podrían darnos, como por ejemplo que los datos del texto no se parecen a la media pero que sí tienen similitud con algunos de los datos individuales de parte del conjunto de textos analizados. Comentar que esto no es nada certero, ya que también se podría dar la casuística en el que los datos del texto analizado difieran mucho tanto de la media como de los datos individuales de los libros de dicho autor, caso en el cual se relacionaría, seguramente, con otro autor diferente.

Dicho todo lo anterior, llegamos a la conclusión de que, a partir de los datos que extraemos se puede determinar si un texto pertenece a un determinado autor, siempre y cuando el conjunto de prueba analizado para dicho autor sea semejante al texto concreto que queramos analizar, y por ello determinamos que no se trataría de una prueba fiable para determinar a que autor pertenece un determinado texto, debido a que, aunque la mayoría de textos puedan estar dentro de una determinada media, de forma aproximada, los casos que estén fuera de esta media no serian clasificados correctamente, y por tanto, necesitaríamos otro tipos de datos o mas datos, aparte de los que disponemos, para poder llevar a cabo la clasificación de forma correcta.