Ultima duda de la primera parte. Para reducir el vocabulario basta con el AttributeSelection CfsSubsetEval y BestFirst que vimos en las prácticas para quitar los que no correlan con la clase y los que si correlan entre si y después un Remove para adaptar los datos del fichero de test???

[weka.sourceforge.io/doc.dev/weka/filters/unsupervised/attribute/FixedDictionaryStringToWordVector.html](https://weka.sourceforge.io/doc.dev/weka/filters/unsupervised/attribute/FixedDictionaryStringToWordVector.html)

* [weka.sourceforge.io/doc.dev/weka/filters/unsupervised/attribute/FixedDictionaryStringToWordVector.html](https://weka.sourceforge.io/doc.dev/weka/filters/unsupervised/attribute/FixedDictionaryStringToWordVector.html)
* **10:40**

Converts String attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings. The set of words (attributes) is taken from a user-supplied dictionary, either in plain text form or as a serialized java object.

Yo tenía tres dudas de la primera parte

No he conseguido que me salga lo del FixedDictionaryStringToWordVector. Ni siquiera me detecta el import del paquete weka.filters.unsupervised.attribute.FixedDictionaryStringToWordVector;.

He estado intentando poner un Tokenicer que es algo importante para el filtrado, he probado varios de por internet pero algunos hacían cosas muy raras (poner las palabras de dos en dos) y otros no hacían nada. En ninguno de los casos me han considerado actriz y actor como la misma palabra.

Y finalmente tras hacer el AtributeSelection CfsSubsetEval y BestFirst se quedan 70 atributos que no se filtran... son demasiados pocos? Muchas instancias se quedan totalmente vacías y la mayoría solo con una. Evidentemente clasificar instancias que no tienen ningún atributo no creo q valla a dar buenos resultados.

Al hacer el AtributeSelection CfsSubsetEval y BestFirst se quedan 70 atributos. son demasiados pocos? Muchas instancias se quedan totalmente vacías y la mayoría solo con una. clasificar instancias que no tienen ningún atributo no creo q valla a dar buenos resultados.

Un tokenicer dejaría mejores resultados al considerar iguales como actor y acrtess que tienen el mismo significado pero son palabras distintas.

He estado intentando poner un Tokenicer que es algo importante para el filtrado, he probado varios de por internet pero algunos hacían cosas muy raras (poner las palabras de dos en dos) y otros no hacían nada. En ninguno de los casos me han considerado actriz y actor como la misma palabra.