Practical

**Aim :** Document Similarity





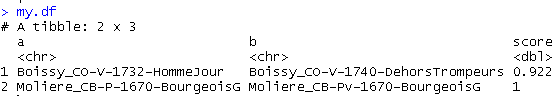






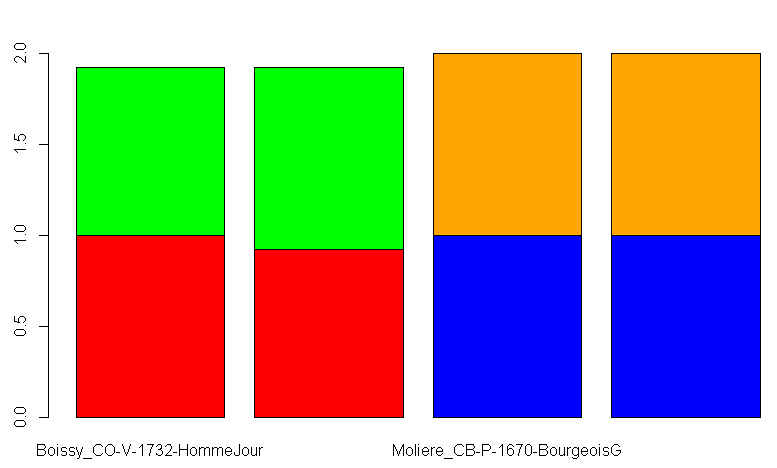












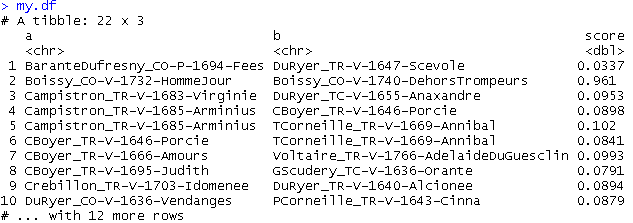






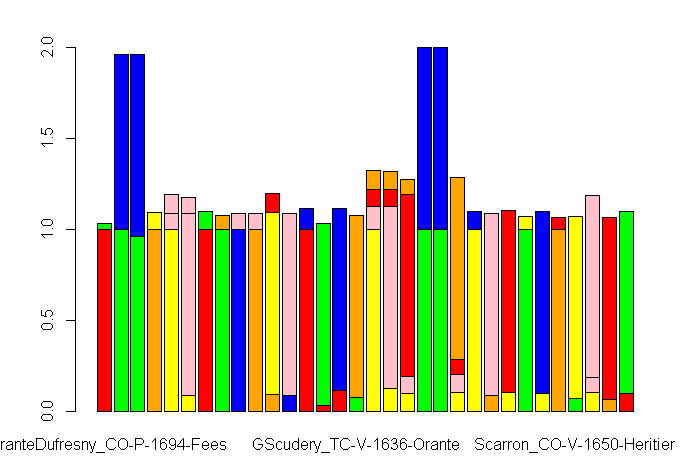












**Description:**

**textreuse-package:** Tools for measuring similarity among documents and detecting passages which have been reused. Implements shingled n-gram, skip n-gram, and other tokenizers; similarity/dissimilarity functions; pairwise comparisons; minhash and locality sensitive hashing algorithms; and a version of the Smith-Waterman local alignment algorithm suitable for natural language.

**minhash-generator:** A minhash value is calculated by hashing the strings in a character vector to integers and then selecting the minimum value. Repeated minhash values are generated by using different hash functions: these different hash functions are created by using performing a bitwise XOR operation ([bitwXor](http://127.0.0.1:21987/help/library/textreuse/help/bitwXor)) with a vector of random integers. Since it is vital that the same random integers be used for each document, this function generates another function which will always use the same integers. The returned function is intended to be passed to the hash\_funcparameter of [TextReuseTextDocument](http://127.0.0.1:21987/help/library/textreuse/help/TextReuseTextDocument).

**as.matrix:** as.matrix attempts to turn its argument into a matrix.

**barplot:** Creates a bar plot with vertical or horizontal bars.