

Colecciones de prueba de Recuperación de información

Colecciones clásicas de RI: http://ir.dcs.gla.ac.uk/resources/test_collections/

Proyecto Gutenberg: <https://www.gutenberg.org/> (hay que descargarse los libros a mano)

Colección del Parlamento Europeo: <http://www.statmt.org/europarl/>

Reuters collection: <http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html>

Reuters RCV1 y RCV2 multilingüe

<https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual,+Multiview+Text+Categorization+Test+collection>

Obras de Shakespeare en XML: <https://www.ibiblio.org/xml/examples/shakespeare/>

Colección NSF: <https://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html>

Colección 2010 IMDB: <https://inex.mmci.uni-saarland.de/protected/dc/2010-datacentric-documents-beautified.tar.gz>

Colecciones de tuits para análisis de sentimiento: <https://blog.cambridgespark.com/50-free-machine-learning-datasets-sentiment-analysis-b9388f79c124>

Un par de colecciones extraídas de la Wikipedia: <http://www.search-engines-book.com/collections/>