Practice III

Document similarity

Specifications

- Form a team of 3 to 4 people
- With the corpus of news generated in practice II perform the following
  1. Load the corpus
  2. Generate the three vector representations reviewed in class (frequency, binarized and tf-idf)
- Select a new text document as input and indicate the type of vector representation. Do the following with this document:
  1. Apply the same normalization process performed with the news corpus
  2. Generate the indicated vector representation
  3. Apply the cosine similarity algorithm to determine the similarity between the input document and the rest of the documents in the news corpus.
  4. Display the 10 most similar documents in descending order

Evidence

- Source code
- Document in PDF with the following table showing the 10 most similar documents of each test

documento_prueba_<num_prueba>	**<contenido>**	
representación_<tipo_de_representación>	documento_corpus_<num_documento>	<valor_de_similitud>

  - <num_prueba>: nombre del archivo de prueba (1, 2, 3, ...)
  - <contenido>: contenido de la noticia de prueba
  - <tipo de representación>: binarizada, frecuencia o tf-idf
  - <num_documento>: número de reglón de la noticia en el corpus (1,2,3, ...)
  - <valor_de_similitud>: valor de sumilitud coseno

- The document must include the names of the team's members
- All the members must upload the evidence