Practice III

Document similarity

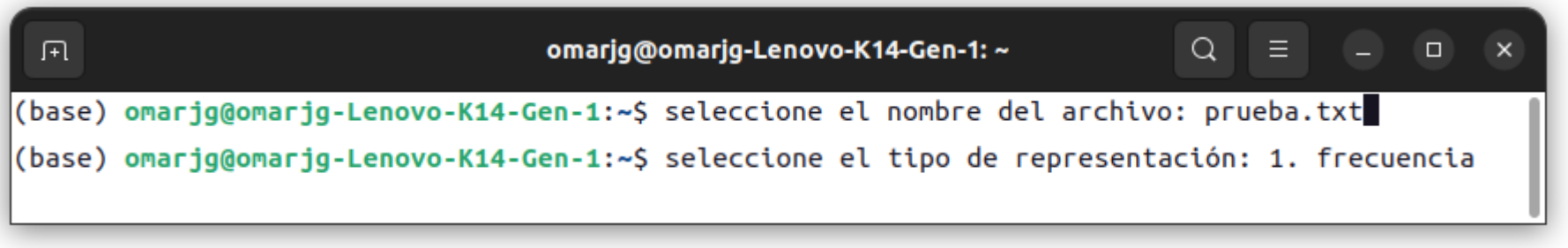
Specifications

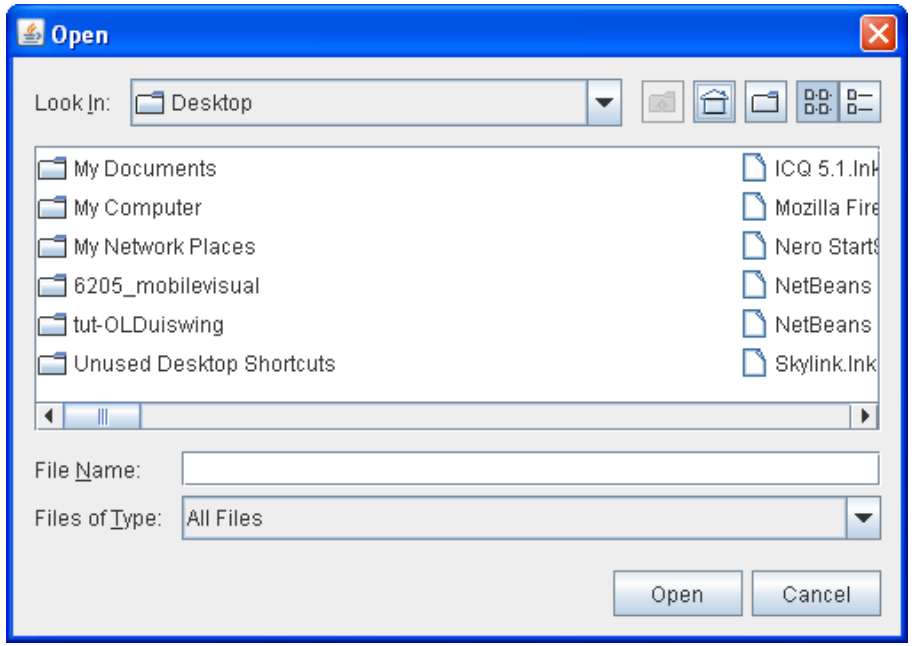
- Form a team of 3 to 4 people
- With the corpus of news generated in practice II perform the following
  1. Load the corpus
  2. Generate the three vector representations reviewed in class (frequency, binarized and tf-idf)
- Select a new text document as input and indicate the type of vector representation. 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

Interface

- An interface with the following specifications needs to be created for the practice
  - The news corpus should be uploaded, and the three distinct text representations must be generated
  - The program must enable the user to select a test file by indicating the file name (and path) or using a browse button in a graphical interface
  - The program should display the 10 most similar documents in descending order

Interface


omarjg@omarjg-Lenovo-K14-Gen-1: ~
(base) omarjg@omarjg-Lenovo-K14-Gen-1:~$ seleccione el nombre del archivo: prueba.txt
(base) omarjg@omarjg-Lenovo-K14-Gen-1:~$ seleccione el tipo de representación: 1. frecuencia


Open
Look In:
Desktop
My Documents
My Computer
My Network Places
6205_mobilevisual
tut-OLDuiswing
Unused Desktop Shortcuts
ICQ 5.1.lnk
Mozilla Fire
Nero Starts
NetBeans
NetBeans
Skylink.lnk
File Name:
Files of Type:
All Files
Open
Cancel

Evidence

- Source code
- Document in PDF with the following table

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

Evidence

- Conclusions of the practice, describing the difficulties encountered, the solutions applied, the results obtained in the different tests and representations generated, as well as suggestions for improvements for future work.
- The document must include the names of the team's members
- All the members must upload the evidence