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 frequency, binarized, and TF-IDF vector representations of columns *Title* and *Content summary* independently and *Title concatenated with Content summary* using the following features:
     a) Unigrams
     b) Bigrams

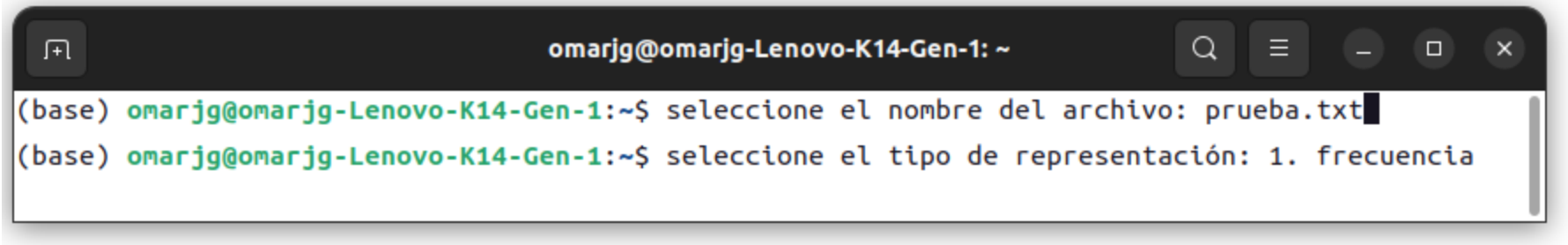
Specifications

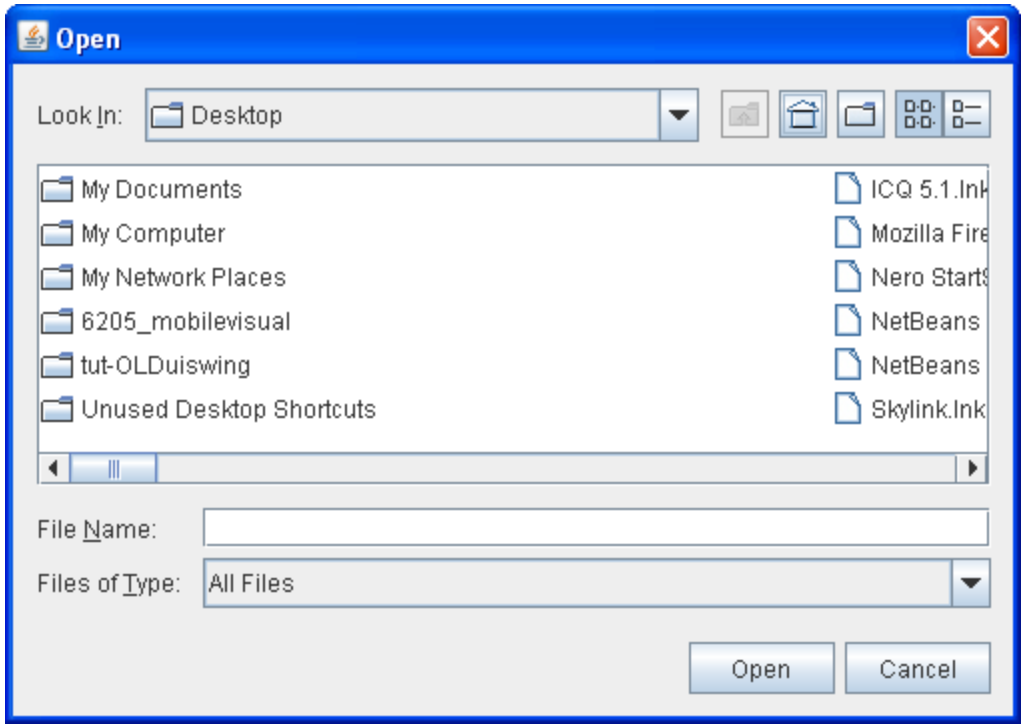
- Select a *test text* document as input and indicate
  a) The type of vector representation
  b) The features to be extracted
  c) The comparative element of the news corpus
- Do the following with this document:
  a) Apply the same normalization process performed with the news corpus
  b) Generate the indicated vector representation
  c) Extract the specified features
  d) Apply the cosine similarity algorithm to determine the similarity between the input document and the rest of the documents in the news corpus using the comparative element
  e) 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 allow the user to select a test file by indicating the file name (and path) or using a browse button in a graphical interface
  - The user must select the type of vector representation, features to be extracted and the comparative element
  - The program should display the 10 most similar documents

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 Start
NetBeans
NetBeans
Skylink.lnk
File Name:
Files of Type:
All Files
Open
Cancel

Evidence

- Source code
- Document in PDF with the following table

Test document <test_num>	**<test_text>**			
Corpus document	*Vector representation*	*Extracted features*	*Comparison element*	*Similarity value*

- Where:
  - <test_num>: number of the test file (1, 2, 3, ...)
  - <test_text>: content of the test file
- The document must include the names of the team's members
- All the members must upload the evidence

Evidence

Test document 1	Al sector energético, una de cada cuatro empresas atraídas por nearshoring			
Corpus document	*Vector representation*	*Extracted features*	*Comparison element*	*Similarity value*
55	TF-IDF	Unigrams	Title	0.65
55	TF-IDF	Bigrams	Content	0.60
60	Frequency	Unigrams	Title	0.55
60	Binarized	Bigrams	Title + Content	0.50
120	TF-IDF	Bigrams	Content	0.48
120	TF-IDF	Unigrams	Title + Content	0.4
120	TF-IDF	Unigrams	Title	0.38
100	Frequency	Unigrams	Content	0.30
100	Binarized	Bigrams	Title	0.25
45	TF-IDF	Unigrams	Title + Content	0.18