Advanced Information Retrieval Systems Assignment #2

2023-05-22

Takashi Yukawa

Assignment #2
レポート（その2）課題

Assignment #2 (1)

- Assume Index Terms
  - $\{t_1, t_2, t_3, \ldots, t_{50}\}$
- Assume documents
  - each document include some index terms
  - d_1: $\{t_{10}, t_7, t_{10}, t_2, t_9, t_6, t_8, t_7, t_5, t_5, t_5, t_6\}$
  - For ease, we write t_n as n
    - with this notation, the document d_1 can be expressed as 10,7,10,2,9,6,8,7,5,5,5,6.
- Assume a document set
  - including 100 documents

Assignment #2 (2)

- Document-Term Matrix is given
  - IRSys23_DocVec.csv
- Format for Doc-Term Matrix
  - 100 rows (the number of documents) x 50 columns (the number of distinct index terms)
  - Each value is rounded to 4 digit after the decimal point

$$\begin{array}{ccccc} & t_1 & t_2 & & t_{50} \\ d_1 & 0.1234, & 0.2938, & \ldots\ \ldots & 0.3485 \\ d_2 & 0.5678, & \ldots\ \ldots & \ldots & \ldots\ \ldots. \\ d_3 & 0.2938, & \ldots\ \ldots & \ldots & \ldots\ \ldots. \\ & : & : & : & : \\ d_{100} & 0.7364, & \ldots\ \ldots & \ldots & 0.8374 \end{array}$$

Assignment #2 (3)

- IDF values are also given
  - IRSys23_IDF.csv
- Format
  - 1 row x 50 columns (the number of distinct index terms)
  - Each value is rounded to 4 digit after the decimal point

t_1 t_2 t_{50}

0.1234, 0.2938, … … 0.3485

Assignment #2 (4)

- Queries are also given
  - IRSys23_Qs.csv
- We have 4 queries
- Format
  - 4 rows (each row corresponds each query

q_1	15,8,25,6
q_2	7,16
q_3	9,43,9
q_4	42,16,17,12

Assignment #2 (5)

- Compute similarity values for each document with each query
- Format
  - 4 rows (each row corresponds each query) x 100 columns (each column corresponds each document)
  - Each value is rounded to 4 digit after the decimal point

	d_1	d_2		d_{100}
q_1	0.1234,	0.2938,	… …	0.3485
q_2	0.5678,	… … …		… ….
q_3	0.2938,	… … …		… ….
q_4	0.7364,	… … …		… ….

Assignment #2 Submission

- Accept the report with ILIAS only
- Upload at "Assignments" > "Assignment #2 submission"
  - File name for the document-term matrix must be S<student_id_6digits>.csv
  - Use only ASCII characters for the file name; do not use Zenkaku characters
- Filename examples
  - Assume your student ID 12345678
  - The document-term matrix: S123456.csv

Submission Format

- Follow the instruction in the assignment **strictly**
- **A submission file with illegal format is not evaluated (i.e. 0 point)**
- Format checker programs are provided
  - checkSimilarityMatrix.py
  - Check your submission files with these programs
    - Programs are written in Python language

"checkSimilarityMatrixpy.sec" on ILIAS. Rename it to "checkSimilarityMatrix.py".

Deadline

- Deadline: June 5th 13:00
- Submission will be closed at the deadline with the ILIAS server's clock
  - do not assume that the clock on the server is quite accurate