Advanced Information Retrieval Systems Assignment #1

2023-05-08

Takashi Yukawa

Assignment #1
レポート（その1）課題

Assignment #1 (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
  - it can be downloaded from ILIAS with file name "IRSys23_Docs.csv"

Assignment #1 (2)

- Document Set

10,7,10,2,9,6,8,7,5,5,5,6	d_1
3,4,1,3,6,3,5,2,1,6	d_2
7,8,10,5,7,5,2,8	d_3
3,7,5,2,3,7,4,9,8,9,8,4,10,1,4,8,9,5	d_4
:	
7,8,1,5,5,8,2,2,4,6,8,5,4,8,7,9,7,3,7	d_{100}

Assignment #1(3)

- Generate (compute) Document-Term Matrix
  - Strictly follow the lecture materials for the Lectures #2 and #3
  - Use 10 for base of logarithm
- Submission Format
  - 100 rows (the number of documents) x 50 columns (the number of distinct index terms)
  - Each value must be rounded to 4 digit after the decimal point

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

Assignment #1 Submission

- Accept the report with ILIAS only
- Upload at “Assignments” > "Assignment #1 submission”
  - File name for the document-term matrix must be D<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: D123456.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
  - checkDocTermMatrix.py
  - Check your submission files with these programs
    - Programs are written in Python language

Deadline

- Deadline: May 22nd 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