IFN647 Tutorial (Week 7): Evaluation

****

Task 1. Understand the definitions of Precision, recall and F measure. You may discuss these measures with your classmate or your tutor if you have any questions.

A is set of relevant documents (e.g., a benchmark, or relevance judgments)**,**

B is set of retrieved documents (e.g., the output of an IR model, a binary output)

	Relevant	Non-Relevant
Retrieved	$A \cap B$	$\overline{A} \cap B$
Not Retrieved	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$

$$Recall = \frac{|A \cap B|}{|A|}$$

$$Precision = \frac{|A \cap B|}{|B|}$$

F-Measure (F1),

$$F = \frac{1}{\frac{1}{2}(\frac{1}{R}+\frac{1}{P})} = \frac{2RP}{(R+P)}$$

where R is Recall, P is Precision.

The IR model can also return a list of ranked documents (in descending order by the document weight or score). At each position *n*, we have the following definition of precision (recall) at position *n*:

$$Recall@n = \frac{|A \cap B_n|}{|A|}$$

$$Precision@n = \frac{|A \cap B_n|}{n}$$

where B_n is the set of ***top-n*** documents in the output of the IR model.

Task 2: read a topic-doc-assignment file (e.g., relevance_judgments.txt, the benchmark) and a retrieved topic-doc-assignment file (e.g., binary_output.txt, the output of an IR model for query *R105*); calculate the IR model’s Recall, Precision, and F-Measure (F1).

- Please download two topic-doc-assignment files and save them is a folder (e.g., “rel_data”). The two files are in format of “topic documentID Relevance_judgment (or Relevence_value (1 or 0))”. For the file “relevance_judgments.txt”, *relevance_judgment* = “1” indicates relevant and “0” means non-relevant. For the file “binary_output.txt”, it provides a set of retrieved documents whose *relevance-value* are labeled as ‘1’ by the IR model.
- Define a function *rel_setting*(*inputpath*), which reads the two topic-doc-assignment files in the folder *inputpath*, and returns a pair of dictionaries {*documentID*: *Relevance_judgment*, …} and {*documentID*: *Relevance_value*, …} for all documents in “relevance_judgments.txt”, and “binary_output.txt”, respectively.
- Define a main function to call function *rel_setting*(), calculate Recall, Precision, and F-measure and display the result.

Example of output

```
The number of relevant documents: 16
The number of retrieved documents: 14
The number of retrieved documents that are relevant: 11
recall = 0.6875
precision = 0.7857142857142857
F-Measure = 0.7333333333333334
```

Task 3: Read the ranked output file (“ranked_output.txt”, the ranked output of the IR model for query *R105*), and calculate its average precision.

- Please download the ranked output file and save it in the same folder you created for Task 2.
- Extend function *rel_setting*(inputpath) to return 3 dictionaries (two for task 2 and one for task 3). The dictionary for task 3 should have the format of {*rankingNo*:

documentID, …}, and it only includes top-10 documents (ranked in descending order), where *rankingNo* = 1, 2, …, 10.

- Extend the main function to calculate Recall and Precision at the rank positions where a relevant document was retrieved, and then calculate the average precision, and print out the result.

Example of output

```
At position 1docID: 2493, precision= 1.0
At position 2docID: 3008, precision= 1.0
At position 3docID: 5225, precision= 1.0
At position 4docID: 5226, precision= 1.0
At position 5docID: 15744, precision= 1.0
At position 7docID: 40239, precision= 0.8571428571428571
At position 8docID: 48148, precision= 0.875
At position 9docID: 49633, precision= 0.8888888888888888
At position 10docID: 51493, precision= 0.9
The average precision = 0.9467813051146384
```