ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE – LAUSANNE
POLITECNICO FEDERALE – LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY – LAUSANNE

Faculté Informatique et Communication
Introduction to Natural Language Processing (Ms; CS-431)
Chappelier, J.-C. & Rajman, M.

CS-431 Hands On Evaluation

J.-C. Chappelier M. Rajman

v. 20190912 – 1

QUESTION I [2 pt]

(from Fall 2018 quiz 4)

For a given query Q, an IR system retrieves 50 documents with a precision $P = 0.6$. If one assumes that the total number of relevant documents for Q is $N = 100$, what is the recall R of the system for Q?

Provide the answer in a form of a fraction.

$$R = \frac{50 \times P}{100} = \frac{3}{10}$$

Extra questions (not in the original quiz)**:** Assume the total collection contains 1000 documents. For the above system, for query Q:

- What is the confusion matrix?
- What would be the accuracy?
- Does it make sense?

Confusion matrix is

System \ Reference	Relevant	Irrelevant	
Retrieved	30	20	50
Not Retrieved	70	880	950
	100	900	1000

Thus the accuracy is

$$\text{acc} = \frac{30 + 880}{1000} = 91\%$$

.

Does not make much sense for the evaluated task as not retrieving non-relevant documents is not so interesting. Relevant documents (on one hand) or retrieved documents (on the other) are of real interest. Thus recall and precision measures.

QUESTION II [6 pt]

(from Fall 2018 quiz 2)

To prepare for an evaluation campaign of NLP systems performing sentiment analysis (i.e. NLP systems able to indicate, for any input text, whether it corresponds to a positive or negative feeling), 100 documents have been analyzed by two human annotators, leading to the following confusion matrix:

Annotator 1 \ Annotator 2	Positive	Negative	
Positive	57	3	60
Negative	3	37	40
	60	40	100

① What is the kappa score for the agreement between the annotators? (Provide your answer in the form of a fraction)

$$\kappa = \frac{(57 + 37) - (36 + 16)}{100 - (36 + 16)} = \frac{42}{48} = \frac{7}{8}$$

② Based on the obtained kappa score, can one say that the gathered annotated data is suitable for performing the targeted campaign?

[✔] Yes

[ ] No

QUESTION III [2 pt]

(from Fall 2018 quiz 2)

To validate the reference corpus prepared for the evaluation of NLP systems on a given task, a large corpus has been annotated by two human experts, and the resulting kappa score is slightly

above 0.8. The organizers of the evaluation campaign would like to provide some additional guarantee that the produced annotated corpus is indeed good enough. Indicate which of the following methods is the most adequate one for this purpose:

[ ] Perform a cross-validation to learn the kappa score

[ ] Split the annotated corpus into T consecutive sub-corpora of equal size and perform a statistical test on the kappa score

[✔] Derive from the annotated corpus T random sub-corpora of equal size and perform a statistical test on the kappa score