2 Evaluation

Exercise II.1

① Give some arguments justifying why evaluation is especially important for NLP. In particular, explain the role of evaluation when a corpus-based approach is used.

② Many general evaluation metrics can be considered for various NLP tasks. The simplest one is accuracy.

Give several examples of NLP tasks for which accuracy can be used as an evaluation metric. Justify why.

In general, what property(ies) must an NLP task satisfy in order to be evaluable through accuracy?

③ Consider a Part-of-Speech tagger[1] producing the following output:

```
The/Determiner program/Noun can/Noun deal/Noun with/Preposition three/Number
types/Verb of/Preposition inputs/Noun ./Punctuation
```

(using your own knowledge of general English,) Compute the accuracy of the tagger.

What do you think of the performance of this system with respect to the State of the Art? Is this conclusion reliable?

④ What is the formal relation between accuracy and the error rate? In which case would you recommend to use the one or the other?

⑤ Consider the following "breaking news scanning system":
A company receives a continuous stream of information messages (newswires); each time a new message arrives, its average textual similarity score with respect to the stored collection of previously received messages is computed. If this average similarity is below a given threshold, the message is considered "breaking news" and is automatically distributed to the company personnel.

The company has carried out an evaluation of the system in place, which produced the following average figures:

- one message out of 1000 is considered to be "breaking news" by the system;
- 30% of the claimed "breaking news" messages are evaluated as not new by human judges;
- the system is missing one truly "breaking news" message every 1000 messages processed.

Use the provided figures to compute the accuracy of the system.

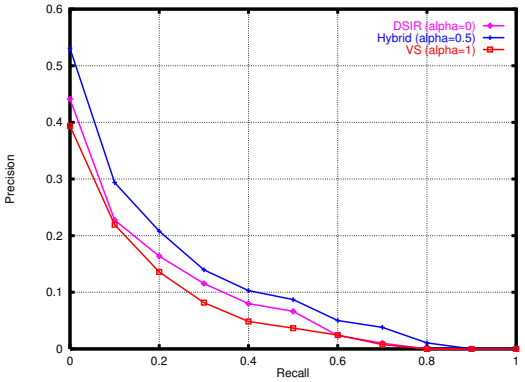
Is accuracy a good metric in this case? Justify your answer, and, possibly, propose some alternative performance score(s) and compute the corresponding value(s).

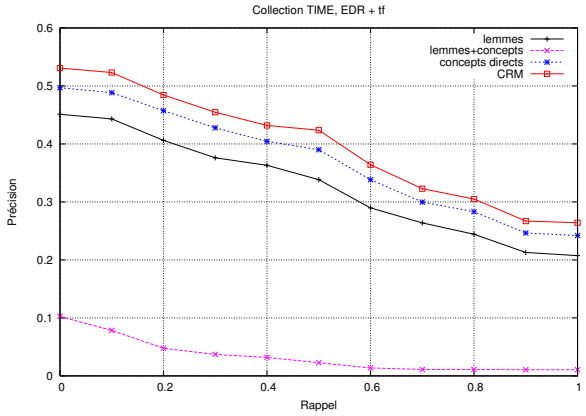
[1] Part-of-Speech tagging, which will be studied in more details later in the semester, consists in adding each word a ("Part-of-Speech") tag corresponding to its syntactic role within the sentence:.

⑥ Another very general evaluation framework concerns this kind of NLP tasks where the goal of the system is to propose a set of outputs among which some might turn to be correct, while other might not (e.g. Information Retrieval (IR)). In this type of situation, the standard evaluation metrics are the Precision and the Recall.

Give the formal definition of Precision and Recall and indicate some examples of NLP tasks (other than IR) that can be evaluated with the Precision/Recall metrics.

⑦ Consider the following Precision/Recall curves





What conclusions can one derive from such curves? Provide a detailed interpretation of the results.

⑧ It is often desirable to be able to express the performance of an NLP system in the form of one single number, which is not the case with Precision/Recall curves.

Indicate what score can be used to convert a Precision/Recall performance into a unique number. Give the formula for the corresponding evaluation metric, and indicate how it can be weighted.

⑨ Give well chosen examples of applications that can be evaluated with the single metric derived from Precision/Recall and illustrate:

- a situation where more weight should be given to Precision;
- a situation where more weight should be given to Recall.

Exercise II.2

You have been hired to *evaluate* an email monitoring system aimed at detecting potential security issues. The targeted goal of the application is to decide whether a given email should be further reviewed or not.

① Give four standard measures usually considered for the evaluation of such a system? Explain their meaning. Briefly discuss their advantages/drawbacks.

② For three of the measures you mentioned in the previous question, what are the corresponding scores for a system providing the following results:

email	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	e_{11}	e_{12}	e_{13}	e_{14}
referential	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_2	C_2	C_2	C_2	C_2	C_2
system	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_1	C_2	C_1	C_2	C_1	C_2	C_2

③ You have been given the results of three different systems that have been evaluated on the same panel of 157 different emails. Here are the classification errors and their standard deviations:

	system 1	system 2	system 3
error	0.079	0.081	0.118
std dev	0.026	0.005	0.004

Which system would you recommend? Why?

④ **Optional** (too advanced for the current version of the course)**:** What should be the minimal size of a test set to ensure, at a 95% confidence level, that a system has an error 0.02 lower (absolute difference) than system 3? Justify your answer.