8 Information Retrieval

Exercise VIII.1

① Describe the main principles of the standard vector space model for semantics.

② Consider the following document:

D = "`the exports from Switzerland to the USA are increasing in 2006`"

Propose a possible indexing set for this document. Justify your answer.

③ What is the similarity between the above document D and

D' = "`Swiss exports have increase this year`"

Justify your answer.

④ Briefly describe the important limitation(s) of the standard vector space approach.

Explain how more sophisticated techniques such as the *Distributional Semantics* can be used to circumvent this/these limitation(s).

⑤ Give some concrete examples of NLP applications that might benefit from the semantic vectorial representations.

⑥ Using the standard vector space model, does the indexing set you considered in question Q2 allow to discriminate between D and this other document:

D" = "`the exports from the USA to Switzerland are increasing in 2006`"

If yes: how? If not, why?

⑦ Would a parser be available, how could it be used to provide a (partial) solution to the problem?

Exercise VIII.2

① What is the cosine similarity?

② Consider the following two documents:

```
D1: Dog eat dog. Eat cat too!
D2: Eat home, it's raining cats and dogs.
```

What would be their cosine similarity in a typical information retrieval setup? Explain all the steps.

③ In a standard cosine-based tf-idf information retrieval system, can a query retrieve a document that does not contain any of the query words? Justify your answer.

④ Other measures than the cosine are possible. For instance, Jaccard similarity computes the ratio between the number of words in common and the total number of occurring words (i.e. "intersection over union").

(a) Can this measure be used on the same document representations as the one used for the cosine similarity can be? Justify your answer.

(b) Using the boolean representation for documents, find an example of three distinct documents d_1, d_2 and d_3 such that d_1 is closest to d_2 using cosine similarity, whereas d_2 and d_3 have the same similarity with respect to d_1 using Jaccard similarity.

Conclude on the comparison between these two similarity measures.

⑤ An information retrieval system with high precision and low recall can be useful for:

(a) Retrieving all relevant documents from a database of legal cases.

(b) Retrieving some interesting documents for a given a topic from the web.

(c) Retrieving a large set of interesting documents for a given a topic from the web.

(d) Checking the existence of a document in a very large document collection.

Choose all possible useful situations in the above list (maybe several). Justify your answer; in particular, define the notions of precision and recall.

Exercise VIII.3

① Official NLP evaluations (especially for task such as Information Retrieval or Information Extraction) are often carried out in the form of "evaluation campaigns".

Precisely describe the various steps of such an evaluation campaign.

For each of the steps, clearly indicate the main goals.

② In an IR evaluation campaign, the following "referential" ("golden truth") has been produced by a set of human judges:

```
q1: d01 d02 d03 d04
q2: d05 d06
q3: d07 d08 d09 d10 d11
q4: d12 d13 d14 d15
```

where the list of document references `dj` associated with a query reference `qi` defines the set of documents considered to be relevant for the query by the human judges.

Is such a referential easy to produce?

Indicate the various problems that might arise when one tries to produce it.

③ Consider two Information Retrieval systems S_1 and S_2 that produced the following outputs for the 4 reference queries q1, q2, q3, q4:

```
S1:                                          | referential:
 q1: d01 d02 d03 d04 dXX dXX dXX dXX         |  q1: d01 d02 d03 d04
 q2: d06 dXX dXX dXX dXX                     |  q2: d05 d06
 q3: dXX d07 d09 d11 dXX dXX dXX dXX dXX     |  q3: d07 d08 d09 d10 d11
 q4: d12 dXX dXX d14 d15 dXX dXX dXX dXX     |  q4: d12 d13 d14 d15

S2::                                         | referential:
 q1: dXX dXX dXX dXX d04                     |  q1: d01 d02 d03 d04
 q2: dXX dXX d05 d06                         |  q2: d05 d06
 q3: dXX dXX d07 d08 d09                     |  q3: d07 d08 d09 d10 d11
 q4: dXX d13 dXX d15                         |  q4: d12 d13 d14 d15
```

where dXX refer to document references that do not appear in the referential. To make the answer easier, we copied the referential on the right.

For each of the two systems, compute the mean Precision and Recall measures (provide the results as fractions). Explain all the steps of your computation.

④ Explain how it is possible to compute Precision at different Recalls.

⑤ How is it possible to compute the average Precision/Recall curves? Explain in detail the various steps of the computation.

As it would be too tedious to compute the average Precision/Recall curves by hand, plot, on a Precision/Recall graph, the Precision and Recall values obtained in subquestion ③ for each of the two systems and for each of the 4 queries.

Based on the resulting curves, what is your relative evaluation of the two systems?

⑥ The Precision/Recall based evaluation of the IR systems S1 and S2 above does not explicitly take into account the order in which the documents have been retrieved by the systems. For this purpose, another metric can be used: the Precision at k (P@k), which corresponds to the fraction of truly relevant documents among the top k documents retrieved by a system.

Compute the average P@k values for k between 1 and 5 for the IR systems S1 and S2 above. What additional insight do these values provide in addition to the Precision/Recall curves?

Based on these results, what is your relative evaluation of the two systems? How does it compare to ③?

⑦ It is often desirable to be able to express the performance of an NLP system in the form of a single number, which is not the case when the Precision/Recall framework is used.

Indicate what scores can be used to convert Precision/Recall measures into a unique number. For each score, give the corresponding formula.

⑧ Give well chosen examples of applications that illustrate:

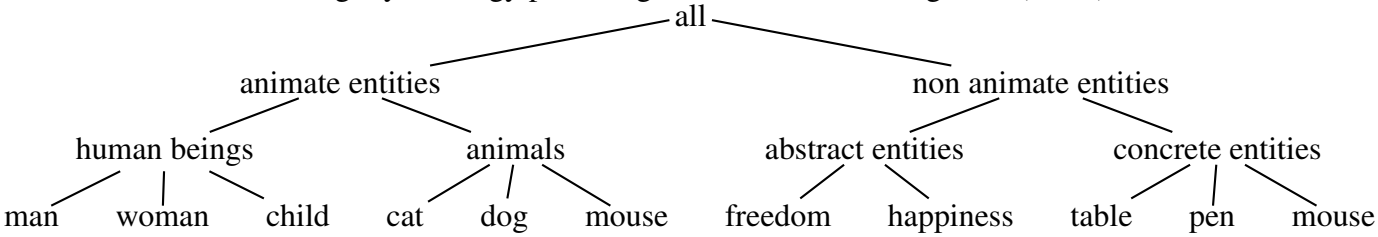
- a situation where more importance should be given to Precision;
- a situation where more importance should be given to Recall.

9 Lexical Semantics

Exercise IX.1

The objective of this question is to illustrate the use of a lexical semantics resource to compute lexical cohesion.

Consider the following toy ontology providing a semantic structuring for a (small) set of nouns:



① What is the semantic relation that has been used to build the ontology?

Cite another semantic relations that could also be useful for building lexical semantics resources.

For this semantic relation, give a short definition and a concrete example.

② The word "*mouse*" appears at two different places in the toy ontology. What does this mean?

What specific problems does it raise when the ontology is used?

How could such problems be solved? (just provide a sketch of explanation.)

③ Consider the following short text:

```
Cats are fighting dogs. There are plenty of pens on the table.
```

What pre-processing should be performed on this text to make it suitable for the use of the available ontology?

④ We want to use lexical cohesion to decide whether the provided text consists of one single topical segment corresponding to both sentences, or of two distinct topical segments, each corresponding to one of the sentences.

Let's define the lexical cohesion of any set of words (in canonical form) as the average lexical distance between all pairs of words present in the set[3]. The lexical distance between any two words is be defined as the length of a shortest path between the two words in the available ontology.

For example, "*freedom*" and "*happiness*" are at distance 2 (length, i.e. number of links, of the path: happiness $\longrightarrow$ abstract entities $\longrightarrow$ freedom), while "*freedom*" and "*dog*" are at distance 6 (length of the path: freedom $\longrightarrow$ abstract entities $\longrightarrow$ non animate entities $\longrightarrow$ all $\longrightarrow$ animate entities $\longrightarrow$ animals $\longrightarrow$ dog)

Compute the lexical distance between all the pairs of words present in the above text and in the provided ontology (there are 6 such pairs).

[3]Here, is actually the *lack of* cohesion that we measure: since it's a distance, the lower the more cohesion and the bigger the less cohesion.

⑤ Compute the lexical cohesion of each of the two sentences, and then the lexical cohesion of the whole text.

Based on the obtained values, what decision should be taken as far as the segmentation of the text into topical segments is concerned?

⑥ Give some examples of NLP tasks for which lexical cohesion might be useful. Explain why.