- You can use any personal notes or books you want.
- You may look at electronic slides with a pdf viewer (different from your web navigator), provided you do not use much your keyboard (you can search strings in your pdf, though). It is strictly forbidden to use a connection to internet, and the wifi reception should therefore be switched of. You may consult your phone for getting time.
- Please answer on the provided notebook in either French or English. Answers should be concise but precise.

---

1. **Language Model**
   (a) Which of the following models **always** consult the lower embedded model?
      (1) Katz
      (2) Jelinek-Mercer
   (b) Which of those modèles are giving the same probability mass to any unknown word?
      (1) Add-one
      (2) Jelinek Mercer
      (3) Maximum de vraisemblance
      (4) Good Turing
   (c) Consider those 2 discret probabilistic models defined on the same domain:

	a	b	c	d
p_1	1/2	1/6	1/6	1/6
p_2	1/4	1/4	1/6	1/3

      (1) Which of p_1 and p_2 has the largest entropy ?
      (2) Express the cross-entropy $H(p_1, p_2)$ as a function of those probabilities (you do not have to do the computation).
      (3) Is $H(p_2, p_1)$ higher or lower than the entropie of p_2 ?
   (d) What is the relation between language models and entropy ?

2. **Analogies**

   We are only concerned in this question with formal analogies as defined by Stroppa & Yvon.

   (a) Identify among those relations, those that are formal analogies:
      (1) aaaa : bb :: abab : baba
      (2) postier: osti: pier : i
      (3) aba : baa : bab : aba
      (4) run : ran :: aratun : arutun
      (5) run : ran :: arutun : aratun
   (b) Solve the following analogical equations:
      (1) [aaaa: ? :: abab : baba]
      (2) [abab: aaaa :: bbbb : ?]

3. **Tagging**

   (a) Indicate the sequence of *IOB* tags that represents this sentence of 9 words (words are space delimited):

   (Le grand homme)$_{np}$ boit (de l' eau forte)$_{NP}$.

   (b) How do we call the action of computing such a tagging ?

   (c) What is the tagset used in the CoNLL benchmark for named-entity recognition?

4. **Grammars**

   Consider the grammar G where S is the axiom:

$$\begin{aligned} &\mathrm{S} \rightarrow \mathrm{AS} \mid \epsilon \\ &\mathrm{A} \rightarrow 0\mathrm{A}1 \mid \mathrm{A}1 \mid 01 \end{aligned}$$

   (a) Is G left recursive ? Justify.

   (b) Is G a regular grammar ? Justify.

   (c) Is G LL1? Justify.

   (d) What is the language described by non terminal A?

   (e) What is the language described by G?

   (f) Can we describe a language of type n with a grammar of type m, where $n > m$? Justify.

   (g) Can we describe a language of type n with a grammar of type m, where $n < m$? Justify.

   (h) Build the analysis table of the Earley algorithm when analyzing the sentence: *00111011*. Each item considered should be mentioned.

   (i) Write a grammar which describes the language of all strings over the alphabet $\{a, b, c\}$ that do not contain the sequence `abc`. `ab`, `abac` are examples of strings belonging to the language.

   (j) Is the language described in the previous question regular? Justify.

   (k) Explain in one sentence the principle of CYK.

5. **Markov Models**

   Consider M, a markov model with 3 states that generates symbols over the alphabet $A = \{a, b, c, d\}$. The probability of starting in state i is $1/3$ ($\pi_i = 1/3, \forall\, i\,[1, 3]$).

   (a) If all states can emit any symbol over A, explain how many state sequences can explain an observation of 4 symbols. Justify.

   (b) What exactly is computing the Viterbi algorithm?

   (c) How many sums are computed by the Viterbi algorithm when working on the sequence *abbbcd*? Justify.

(d) Adjust the parameters of M so to favour strings over A that start by a, and that contain symbols in increasing (lexicographic) order. *abd* should for instance be preferred by your model over *adb* or *bac*.

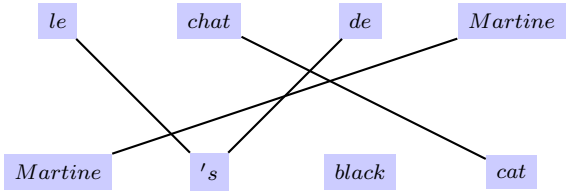
6. **IBM Models**

(a) What is the so-called IBM alignment constraint?

(b) What kind of data structure can represent such an alignment?

(c) Why was this constraint introduced? How are unaligned words dealt with?

(d) With what model has the following IBM alignment been obtained: with $p(e|f)$ or with $p(f|e)$, where e and f stands for English and French words respectively? Justify.



(e) Express (but do not compute) the probability of this alignment, as given by an IBM model 2.

(f) You have two parallel French-English documents. Explain a way to get synonyms of English words using those documents and IBM models.

7. **Edit Distance**

We assume that insertion, deletion and substitution (to another symbol) are edit operations with unity cost.

(a) Can edit distance between two strings be larger than the largest string? Justify.

(b) Represent the edit distance table for strings ABCD and ACBD. How many minimal cost alignments are there? Indicate them.

(c) What is the hamming distance among the strings of the previous question?

(d) You want to introduce a new edit operation that allows the swapping of two adjacent symbols, so that we can transform `corss` into `cross` in one such operation.

What condition on the cost of this operation should be observed for the operation to be useful?

(e) Write an algorithm capable of computing an edit-distance with this swapping operation added.