- You can use any notes, books you want.
- You may as well look at electronic slides with a pdf viewer (not your navigator), provided they are downloaded on your laptop and you do not use much your keyboard. In any case, the wifi connexion should be turned of. You may consult your phone to get time.
- Write down (in English if you wish) your answers on the homework book provided.

---

1. **Language model**
   (a) What is the most important distinction between a *backoff* model and an interpolated one? Name one model seen in class of each category.
   (b) Use $\#u$ and $|V|$ to express the denominator of the following equation, such that p defines a bigram model over the vocabulary V. Here, $\#x$ stands for the frequency of x in the training set, V is the vocabulary set, and $|V|$ stands for its size (the number of types). Justify.

$$p(v|u) = \frac{\#uv + \frac{1}{e^{\#u}}}{to\ complete} \quad \forall u, v \in V$$

   (c) What is the intuition behind this model ?
   (d) You want to truecase a lowercased text. So you want to transform a text such as "*facebook a transmis à cambridge analytica le profil de 50m de ses utilisateurs*" into "*Facebook a transmis à Cambridge Analytica le profil de 50M de ses utilisateurs*". Explain how you could do this with a language model.

2. **Grammaire**
   A) Consider this grammar, where S is the axiom:

$r_1 \equiv$ Ph $\rightarrow$ Suj Verb Comp	$r_5 \equiv$ GNom $\rightarrow$ Art Adj Nom	$r_9 \equiv$ Verb $\rightarrow$ *is* | *can* | *drink*
$r_2 \equiv$ Ph $\rightarrow$ Verb Comp	$r_6 \equiv$ GNom $\rightarrow$ Art Nom	$r_{10} \equiv$ Nom $\rightarrow$ *can* | *bottle*
$r_3 \equiv$ Suj $\rightarrow$ GNom	$r_7 \equiv$ Comp $\rightarrow$ GNom	$r_{11} \equiv$ Pro $\rightarrow$ *it* | *I*
$r_4 \equiv$ Suj $\rightarrow$ Pro	$r_8 \equiv$ Adj $\rightarrow$ *empty* | *full*	$r_{12} \equiv$ Art $\rightarrow$ *the*

   (a) Is this a regular grammar ? Justify.
   (b) Is the language described by this grammar regular ? Justify.
   (c) Is this grammar LL1 ? Justify.
   (d) Construct the Earley table for the analysis of the sentence *I drink the can*. Each item considered by the algorithm seen in class should be mentioned.

B) Let $\mathcal{L}$ be the language of strings over the alphabet $\{a, b\}$ that contain the sequence *ab*. Strings *aabb*, *aab* and *baba* belogn to this language, on the contrary to *ba* or *baaa*.

(a) Write a grammar which recognizes $\mathcal{L}$ and only $\mathcal{L}$.

(b) Do you think $\mathcal{L}$ is regular? Justify.

3. **Analogies**

(a) Use your words to describe what analogical learning is. Explain how atypic it is in machine learning.

(b) Assuming the definition of analogy proposed by Stroppa & Yvon, identify formal analogies in what follows:
   a) aaaa : bb :: abab : baba
   c) run : ran :: drink: drunk
   d) run : ran :: aratun : arutun
   e) il mange la pomme : la pomme est mangée :: il avale la poire : la poire est avalée

(c) Resolve the analogical equation: [tinggal : ketinggalan : duduk: ?].

(d) Consider this corpus of pairs of translated forms:
`(faillites,bankruptcies), (futilité, trivialities), (faillite, bankruptcy)`.
Explain how analogical learning might identify the translation of the word `futilité` unseen in this corpus. You shoudl identify the translation produced.

4. **Hidden Markov model**

Consider the model H specified by a transition matrix A, an emission matrix B and probabilities of initial transitions π. We assume that each time the model enter a state, its emits an observation (here among a, b et c):

$$A = \left[\begin{array}{c|ccc} & s_1 & s_2 & s_3 \\ \hline s_1 & 0.1 & 0.3 & 0.6 \\ s_2 & 0.5 & 0.0 & 0.5 \\ s_3 & 0.8 & 0.2 & 0.0 \end{array}\right], B = \left[\begin{array}{c|ccc} & a & b & c \\ \hline s_1 & 1.0 & 0.0 & 0.0 \\ s_2 & 0.4 & 0.6 & 0.0 \\ s_3 & 0.0 & 0.0 & 1.0 \end{array}\right], \pi = [1.0, 0.0, 0.0]$$

(a) Assuming s_2 is the only final state of the mode, what is the language that H recognize?

(b) Express the probability of the string *aaaa* in terms of emission and transition probabilities. Recall s_2 is the only final state.

(c) The Viterbi algorithm has a complexity in $o(N^2 \times T)$ where N is the number of states, and T is the size of the input sequence. For a model with any states, this may be prohibitive to compute. Propose a sensible heuristic to reduce the computations, and express the resulting complexity. We accept here to loose the optimality of the initial algorithm. You should not propose a new algorithm (by the way, we have seen one in class).

5. **IBM Models**

Consider the French sentence f ≡ *Jean aime la pêche* and the English one: e ≡ *John likes fishing*, as well as the lexical table of an IBM model 1, where f_0 indicates the source added to account for words in the target language without any sourvce association.

Jean	John *(0.8)* Jean *(0.2)*
aime	likes *(0.7)* loves *(0.15)* hates *(0.15)*
la	the *(0.9)* fishing *(0.1)*
pêche	fishing *(0.95)* John *(0.05)*
f_0	the *(0.6)* it *(0.2)* Jean *(0.1)* likes *(0.1)*

(a) Draw the most likely alignment obtained by IBM model 1 for those two sentences.

(b) Express the probability according to this IBM1 model 1, that sentence e is the translation of f, as a function of the probabilities provided. You do not have to do the computation.

(c) You have two corpora of texts, one in English E, the other in French F. Those corpora are not related, and are therefore likely not parallel. You also have a lexical transfer table: $p(e|f)$. Explain how to compute $p(f|e)$.

(d) What do you think will happen if you train an IBM model with a bitext where the target part is identical to the source one ? Justify.

6. **Edit Distance**

Assume insertion, deletion and substitution costs to be 1.

(a) What is the maximal edit-distance between two strings of length n and m respectively ? Justify.

(b) What is the minimal édit distance between two strings of length n and m respectively ? Justify.

(c) Is it possible that two strings of similar length have an edit distance smaller than their hamming distance ? Justify.

(d) An edition table may encode several alignments of minimal cost. What is a necessary condition that one cell (i, j) in this table must verify for two alignments to exist ? Justify.

(e) Give the edition table of strings TRUC et TURC.

(f) Provides a minimal cost alignment between those two strings. If there are more than one alignment, provide a second one.