- Vous avez le droit à vos notes de cours, livres, etc.
- Vous pouvez visualiser vos notes de cours sur votre ordinateur portable. Cependant, vous devez désactiver la connexion au réseau et ne faire usage d'aucune autre application que votre lecteur pdf. Votre téléphone ne peut être utilisé que pour vous donner l'heure.
- Répondez directement sur le carnet de réponse; les questions appellent le plus souvent à des réponses courtes et précises.

---

1. **Modèle de langue**
   (a) Quelle est la différence la plus importante entre un modèle de repli (*backoff*) et un modèle interpolé ? Nommez un modèle de chaque vu en cours.
   (b) Exprimez en fonction de $\#u$ et $|V|$ le dénominateur de la formule suivante pour que p définisse un modèle probabiliste bigramme sur le vocabulaire V, où $\#x$ est la fréquence de x en corpus (d'entraînement), V est l'ensemble des types de ce corpus et $|V|$ désigne la taille de V (c'est-à-dire le nombre de types). Justifiez.

$$p(v|u) = \frac{\#uv + \frac{1}{e^{\#u}}}{\grave{a}\ compl\acute{e}ter} \quad \forall u, v \in V$$

   (c) Indiquez quelle pourrait être l'intuition d'un tel modèle.
   (d) Indiquez comment utiliser un modèle de langue de façon à restaurer la casse d'un texte en minuscule. Vous cherchez donc à transformer un texte comme "*facebook a transmis à cambridge analytica le profil de 50m de ses utilisateurs*" en le texte "*Facebook a transmis à Cambridge Analytica le profil de 50M de ses utilisateurs*".

2. **Grammaire**

A) Soit la grammaire suivante ou S est l'axiome:

$r_1 \equiv$ 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) Cette grammaire est-elle régulière ? Justifiez.
   (b) Le langage décrit par cette grammaire est-il régulier ? Justifiez.
   (c) Cette grammaire est-elle LL1 ? Justifiez.
   (d) Construisez la table d'analyse de l'algorithme d'Earley pour l'analyse de la phrase: *I drink the can*. Chaque item considéré par l'algorithme doit être indiqué.

B) Soit $\mathcal{L}$ le langage des chaînes sur l'alphabet $\{a, b\}$ qui contiennent la séquence *ab*. Les chaînes *aabb*, *aab* et *baba* appartiennent par exemple à ce langage, au contraire de *ba* et *baaa*.

(a) Écrire une grammaire qui reconnait $\mathcal{L}$ et seulement $\mathcal{L}$.

(b) Selon vous, $\mathcal{L}$ est-il régulier ? Justifiez.

3. **Analogies**

(a) En vos mots, décrire ce qu'est l'apprentissage analogique. Expliquez en quoi il est atypique dans le paysage de l'apprentissage automatique.

(b) En prenant la définition d'analogie formelle de Stroppa & Yvon, identifiez ci-après les analogies de formes:

a) aaaa : bb :: abab : baba
b) run : ran :: fun : fan
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) Résoudre l'équation analogique [tinggal : ketinggalan : duduk: ?].

(d) Considérez le corpus constitué de paires de mots en relation de traduction: `(faillites,bankruptcies), (futilités, trivialities), (faillite, bankruptcy)`. Indiquez comment l'apprentissage analogique peut découvrir la traduction du mot `futilité` non vu dans ce corpus. Vous indiquerez une traduction produite.

4. **Modèle de Markov**

Considérez le modèle markovien H dont les matrices de transition A, d'émission B et de transition initiale π sont données ci-après. On rappelle que dans un tel modèle, chaque état atteint émet une observation (ici parmi 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) Quel est le langage reconnu par H si on admet que seul s_2 est un état final du modèle ?

(b) Quelle est la probabilité de la chaîne *aaaa* donnée par H (s_2 est le seul état final)? Je ne vous demande pas de la calculer, mais de l'exprimer en fonction des probabilités de transition et d'émission.

(c) L'algorithme de Viterbi a une complexité en $o(N^2 \times T)$ où N est le nombre d'états et T est la taille de la séquence d'entrée. Pour un modèle avec de nombreux états, cela peut être prohibitif. Proposez une heuristique raisonnable pour limiter les calculs faits par l'algorithme (on accepte l'idée de perdre l'optimalité) et indiquez la complexité résultante. On ne cherchera pas ici à proposer un nouvel algorithme (nous en avons d'ailleurs discuté un autre en cours).

5. **Modèles IBM**

Considérez les deux phrases suivantes, l'une en français: f ≡ *Jean aime la pêche* et l'autre en anglais: e ≡ *John likes fishing*, et la table de transfert IBM1 suivante où f_0 désigne le mot ajouté côté source pour rendre compte des mots cibles non alignés:

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) Dessinez l'alignement de mots le plus probable entre f et e obtenu par le modèle IBM1 ci-dessus.

(b) Exprimez la probabilité selon le modèle IBM 1 que e soit la traduction de f en fonction des probabilités ci-dessus. Je ne vous demande pas de faire le calcul.

(c) Vous disposez de deux corpus de textes, l'un en anglais E, l'autre en français F. Ces deux corpus ne sont pas reliés et donc pas parallèles. Vous avez également un modèle de mots (type IBM 1) $p(e|f)$. Indiquez comment calculer un modèle $p(f|e)$.

(d) Imaginez que vous entraîniez un modèle de traduction IBM avec comme bitexte des paires de phrases identiques (la ième phrase de la partie source est identique à la ieme phrase de la cible). Quel genre d'information pensez-vous que les distributions lexicales vont capturer ? Justifiez.

6. **Distance d'édition**

On supposera dans cette question les coûts d'insertion, de suppression et de substitution unitaires.

(a) Quelle est la distance d'édition maximale entre deux chaines de taille n et m ? Justifiez.

(b) Quelle est la distance d'édition minimale entre deux chaines de taille n et m ? Justifiez.

(c) Est-il possible que deux chaines de même longueur aient une distance d'édition plus petite que leur distance de hamming ? Justifiez.

(d) Une table d'édition peut représenter plusieurs alignements de plus faible distance. Indiquez une condition nécessaire que doit vérifier au moins une case (i, j) de la table d'édition pour qu'au moins 2 alignements de plus faible distance existent. Justifiez.

(e) Faite la table d'édition pour les chaînes TRUC et TURC.

(f) Indiquez un alignement de coût minimal pour ces deux chaines. S'il en existe plusieurs indiquez en un deuxième.