Consignes

- 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 wifi et ne faire usage d'aucune autre application que votre lecteur pdf.
- Répondez directement sur le carnet de réponse.
- Les questions appellent le plus souvent à des réponses courtes et précises.
- Le barème (entre parenthèse) est donné à titre indicatif seulement.

---

(5) 1. **Le coin modèle de langue**

(a) Nommez une technique de lissage vue en cours qui donne la même probabilité à tous les événements non vus lors de l'entraînement et indiquez clairement la masse de probabilité associée.

(b) Quelle est l'idée maîtresse de la technique de lissage *Absolute discounting* ?

(c) Exprimez en fonction de λ, δ, $|u|$, $|V|$ et $|T|$ (la taille comptée en mots du corpus d'entraînement T) le dénominateur de la formule suivante pour que p définisse un modèle probabiliste bigramme sur le vocabulaire V. Justifiez.

$$p(v|u) = \frac{|uv| + \lambda|v| + \delta}{\textit{à compléter}} \quad \forall u, v \in V$$

où $|\bullet|$ est la fréquence de $\bullet$ en corpus (d'entraînement), V est l'ensemble des types de ce corpus et $|V|$ désigne le nombre de types dans T.

(d) Exprimez la masse de probabilité $p(v|u)$ associée par ce modèle à un mot v lorsque u n'a pas été vu à l'entraînement.

(3) 2. **Le coin présentations**

(a) Brièvement, qu'est-ce que FrameNet ?

(b) En quoi consiste la tâche d'étiquetage sémantique présentée par William Lechelle ?

(c) Dans sa présentation, Joseph Le Roux nous a parlé d'analyse en dépendance projective. De quoi s'agit-il ?

(10) 3. **Le coin grammaire**

Soient G_1 et G_2 deux grammaires d'axiome S et de symboles terminaux {a,b,c,d,e,f}.

G_1 :

S → AC | BD | f	E → S	
B → aE | c D	D → d S	
A → c E | b E	C → c S	

G_2 :

S → A S c S | B S d S | f
A → c | b
B → a | cd

(a) G_1 est-elle LL1 ? Justifiez.

(b) Calculez FIRST(S) pour G_1.

(c) La chaîne *abcfcfcfdf* appartient-elle au langage décrit par la grammaire G_1 ? Démontrez-le.

(d) Existe-t-il une chaîne du langage décrit par la grammaire G_1 qui contient exactement 3 symboles ? Justifiez.

(e) La grammaire G_1 est-elle en forme normale de Chomsky ? Dans la négative, donnez-en une forme CNF.

(f) Faire la table d'analyse CYK pour la chaîne `cdcfcfdf` et la grammaire G_1 (ou sa forme CNF selon votre réponse à la question précédente)

(g) Tracez l'arbre d'analyse résultant.

(h) Les grammaires G_1 et G_2 décrivent-elles le même langage ? Justifiez.

(3) 4. **Analogies**

Trouvez une solution aux équations analogiques suivantes en prenant comme définition de l'**analogie formelle**

celle de Stroppa & Yvon vue en cours [1] :

(a) [dream : undreamable : : believe : ?]

(b) [KoFib : KuFob : : QoRi' : ?]

(c) [Atomkraftwerken : Atomkriegen : : Kraftwerks : ?]

(4) 5. **Le coin HMM**

Considérez un modèle markovien de 3 états (s_1,s_2,s_3) générant des symboles dans {a,b,c,d} et défini comme suit :

$$A = \left[\begin{array}{c|ccc} & s_1 & s_2 & s_3 \\ \hline s_1 & 0.3 & 0.7 & 0.0 \\ s_2 & 0.0 & 0.4 & 0.6 \\ s_3 & 0.0 & 0.0 & 1.0 \end{array}\right], B = \left[\begin{array}{c|cccc} & a & b & c & d \\ \hline s_1 & 1.0 & 0.0 & 0.0 & 0.0 \\ s_2 & 0.0 & 0.1 & 0.9 & 0.0 \\ s_3 & 0.0 & 0.0 & 0.4 & 0.6 \end{array}\right], \pi = [1.0, 0.0, 0.0]$$

(a) Quel est le langage reconnu par ce modèle ?

(b) Exprimez la probabilité donnée par le modèle à la chaîne *abcd* ? (je ne vous demande pas de calculer cette probabilité).

(4) 6. **Information mutuelle**

Considérez le corpus suivant : `abbaabcbacbaaabc`

(a) Calculez l'information mutuelle ponctuelle entre a et b, entre a et c.

(b) Sur la base du calcul précédant, quel est la paire de mot la plus "collocative" ?

(c) Indiquez une limitation de l'information mutuelle ponctuelle dans le cas d'événements peu fréquents.

(d) Étant donné ce corpus, quel est parmi a,b,c, le symbole qu'il serait le plus surprenant de voir ? Justifiez.

1. Vous pouvez répondre sans avoir compris les détails de cette définition !

(4) 7. **Distance d'édition**

Considérez la table d'édition suivante :

		C	H	I	E	N	S
	0	1	2	3	4	5	6
N	1	1	2	3	4	4	5
I	2	2	2	2	3	4	5
C	3	2	3	3	3	4	5
H	4	3	2	3	4	4	5
E	5	4	3	3	3	4	5

(a) Quelle est la distance d'édition entre *chiens* et *niche* ?

(b) Indiquez deux alignements de ces chaînes menant à la distance minimale.

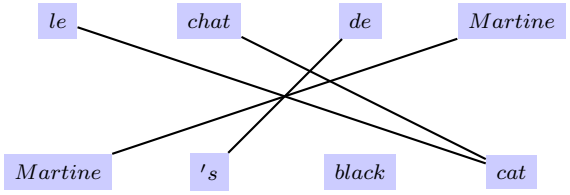
(c) Quelle est la distance d'édition entre *chiens* et *niches* ? (vous n'avez pas besoin de calculer une nouvelle table !)

(5) 8. **Le coin IBM**

(a) Expliquez clairement ce que désigne $a(i|j, I, J)$ dans le cas des modèles IBM2 ?

(b) L'ensemble des modèles IBM vus en cours partagent un ensemble de distributions : lesquelles ?

(c) Considérez cet alignement IBM. Avec quel modèle cet alignement a-t-il été obtenu : $p(e|f)$ ou $p(f|e)$ (où e et f désignent respectivement des mots anglais et français) ?



(d) Vous disposez d'un modèle $p(e|f)$ pour une paire de langues (E, F) ainsi que d'un corpus de textes de la langue E et un corpus de textes de la langue F ; ces deux corpus ne sont à priori pas en relation de traduction. Indiquez comment obtenir un modèle $p(f|e)$.

(4) 9. **Le coin programmation**

Considérez le fichier `capitale.txt` qui contient des analogies (une par ligne) de la forme (capitale$_1$ pays$_1$ capitale$_2$ pays$_2$), dont voici un extrait :

```
Athens Greece Baghdad Iraq
Athens Greece Bangkok Thailand
Baghdad Iraq Bangkok Thailand
```

(a) Sachant qu'une même paire capitale/pays peut se retrouver plusieurs fois dans des analogies de ce fichier, écrivez un programme (de préférence une ligne de commande) qui permet de lister les différentes paires capitale/pays présente dans le fichier `capitale.txt`

(b) Écrivez une commande (ou un programme) capable de lister les pays différents pour lesquels une capitale est mentionnée dans le fichier `capitale.txt`.

Bonne chance