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) à l'internet 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 est donné à titre indicatif seulement.

---

(5) 1. Le coin 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) Vous désirez mettre en place un détecteur de spam. Vous avez pour cela un corpus de messages classifiés comme désirables ou pas. Indiquez comment les modèles de langue peuvent être mis à profit pour réaliser votre détecteur.

(c) Décrivez de manière précise (avec une formule) comment mettre au point un modèle unigrame qui réalise un lissage de type *add-one* si un mot a été vu moins de 10 fois dans le corpus d'entraînement et aucun lissage sinon (estimation par fréquence relative). Votre modèle doit définir une distribution probabiliste. Quelle serait la probabilité associée par votre modèle à un mot non vu dans le corpus d'entraînement? Quelle pourrait être l'intuition derrière un tel modèle ?

(8) 2. Le coin grammaire

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

$r_1 \equiv S \rightarrow GN\ GV$	$r_7 \equiv GV \rightarrow V\ PP\ GN$	$r_{14} \equiv NOM \rightarrow imbécile$
$r_2 \equiv S \rightarrow ADV\ GN\ GV$	$r_8 \equiv GV \rightarrow V$	$r_{15} \equiv NOM \rightarrow idiots$
$r_3 \equiv GN \rightarrow ADJ\ ART\ NOM$	$r_9 \equiv ADV \rightarrow seuls$	$r_{16} \equiv NOM \rightarrow idiot$
$r_4 \equiv GN \rightarrow ART\ NOM$	$r_{10} \equiv ADV \rightarrow seul$	$r_{17} \equiv PP \rightarrow de$
$r_5 \equiv GV \rightarrow V\ GN$	$r_{11} \equiv ADJ \rightarrow seul$	$r_{18} \equiv ART \rightarrow l'$
$r_6 \equiv ART \rightarrow un$	$r_{12} \equiv ART \rightarrow les$	$r_{19} \equiv V \rightarrow aiment$
	$r_{13} \equiv V \rightarrow rit$	$r_{20} \equiv V \rightarrow aime$

(a) Cette grammaire est-elle régulière ? Justifiez.

(b) Cette grammaire est-elle LL1 ? Justifiez (sans démontrer).

(c) Calculez FIRST(S).

(d) Calculez FOLLOW(PP).

(e) Le langage associé est-il régulier ? Justifiez (sans démontrer).

(f) Représenter la table d'analyse CYK pour la phrase:
`seul l' imbécile rit de l' idiot`

(g) Donnez une version probabiliste de cette grammaire[1]. Quelle est l'analyse la plus probable de la phrase précédente dans ce cas?

B) Soit $\mathcal{L}$ le langage des chaînes sur l'alphabet $\{0, 1\}$ d'au moins 2 caractères qui ne contiennent aucune séquence 010.

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

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

(5) 3. Intermède

(a) Trouvez une solution aux équations analogiques suivantes en prenant comme définition de l'analogie formelle celle vue en cours[2]:
[mission : missionnaire : division : ?]
[tinggal : ketinggalan : duduk: ?]

(b) Considérez le corpus constitué de paires de mots en relation de traduction:
`(faillites,bankruptcies), (futilité, 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.

(c) Considérez toutes les séquences sur l'alphabet $\{a, b, c\}$. On suppose que ces séquences sont générées par un processus à trois états: A qui génère seulement des a, B qui génère des a et des b et C qui génère des b et des c. Indiquez un modèle de Markov capable de modéliser ce processus (vous pouvez le dessiner ou l'indiquer sous forme matricielle, toutes les probabilités doivent être mentionnées). Ce modèle est-il visible ou caché? Indiquez la probabilité de la séquence *ccab* (je ne vous demande pas de la calculer mais de l'exprimer en fonction des probabilités élémentaires que vous aurez choisies).

(5) 4. Le coin programmation dynamique

Considérez cette table d'édition entre les formes `MEALEN` et `ETAL`:

		E	T	A	L
	0	1	2	3	4
M	**1**	1	2	3	4
E	2	**1**	**2**	3	4
A	3	2	2	**2**	3
L	4	3	3	3	**2**
E	5	4	4	4	**3**
N	6	5	5	5	**4**

[1] Donnez simplement les probabilités associées à chaque règle r_i.
[2] Vous pouvez répondre sans avoir compris les détails de cette définition!

(a) Quelle est la distance d'édition (levenshtein distance) entre ces deux chaînes ?

(b) Quel est l'alignement induit par le chemin dont les chiffres sont en gras et encadrés ?

(c) Existe-t-il un autre alignement de ces deux chaînes de même distance ? Dans l'affirmative, indiquez-en un.

(d) Quelle est la distance d'édition entre `MEALEN` et `ETA` ? Indiquez un alignement de coût minimal.

(7) 5. Le coin des modèles IBM

Considérez les deux phrases suivantes, l'une en français: f $\equiv$ *Jean aime la pêche* et l'autre en anglais: e $\equiv$ *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 un alignement de mots plausible entre f et e **sous la contrainte IBM**, si le français est la langue source (modèle $p(e|f)$).

(b) Même question si la langue source est l'anglais (modèle $p(f|e)$).

(c) Quel est le score donné par le modèle IBM1 à p(e|f) ? Je ne vous demande pas de calculer cette probabilité exactement, mais plutôt d'exprimer le calcul en fonction des probabilités élémentaires.

(d) On vous donne deux textes *a priori* non traduits l'un de l'autre; l'un en français (F), l'autre en anglais (E). Indiquez comment utiliser ces textes pour dériver du modèle de transfert sus-mentionné un modèle $p(f|e)$. Votre description doit être précise.

(e) De manière générale, combien d'alignements de mots existent entre une phrase source de n mots et une phrase cible de m mots ? Expliquez votre réponse.

(f) Même question dans le cas des alignements de type IBM.

Bonne chance