INSA Rennes **2 juin 2021**

3^e année INFO **durée : 1 heure**

DEVOIR SURVEILLÉ DE TALEO

Traitement Automatique du Langage Écrit et Oral

L'épreuve comporte 4 pages. Les questions sont indépendantes. Le barème est donné à titre indicatif ; il pourra être modifié *a posteriori*, dans les limites du raisonnable. Les documents autorisés sont ceux qui ont été précisés sur Moodle. **Pour toutes les questions posées, une réponse synthétique mais précise et justifiée est demandée**. Des calculatrices (sans connexion Internet) sont autorisées.

1 Recherche d'information : évaluation de systèmes - 3,5 points

On s'intéresse à l'évaluation de systèmes de recherche d'information. Pour la question qui nous intéresse ici, les jugements de pertinence indiquent qu'il y a 8 documents pertinents. Les réponses fournies pour cette question par le système que l'on cherche à évaluer sont présentées ci-dessous en une liste ordonnée dans laquelle une croix X marque un document pertinent et un tiret - marque un document non pertinent pour la requête (en clair, le 1er document retourné est pertinent, le 2^e ne l'est pas, le 3^e l'est, etc.) :

Rang	Pertinence du document	Rang	Pertinence du document
1	X	11	-
2	-	12	-
3	X	13	-
4	-	14	-
5	-	15	X
6	-	16	-
7	-	17	-
8	X	18	-
9	X	19	-
10	-	20	X

Question 1.1 : (1 pt) Définitions : qu'est-ce que P(10), la précision à 10 ? Qu'est-ce que R(10), le rappel à 10 ?

Question 1.2 : (1 pt) Quelles sont ici les valeurs de ces 2 mesures ?

Question 1.3 : (1,5 pt) Quelle est la valeur de la MAP ou précision moyenne non interpolée ? Donnez le détail du calcul en plus de son résultat final (une réponse, même correcte, sans le détail du calcul ne sera pas prise en compte).

2 Recherche d'information : représentation de documents - 4 points

On se positionne dans le cadre du modèle vectoriel (VSM, *vector space model)* pour ce qui est de la représentation de documents et on s'intéresse à comparer entre elles les représentations obtenues. La mesure de proximité sémantique utilisée est la distance de Manhattan qui s'exprime de la façon suivante :

$manh(\vec{d_k}, \vec{d_j}) = \sum_{i=1}^{n} |d_{k,i} - d_{j,i}|$

où $\vec{d_k}$ et $\vec{d_j}$ sont les vecteurs associés aux documents k et j respectivement, $d_{k,i}$ (resp. $d_{j,i}$) correspond à la fréquence (c.-à-d. le nombre d'occurrences) du terme d'indexation i dans le document k (resp. j), n est le nombre de termes d'indexation retenus, et $|X|$ représente la valeur absolue de X.

On considère les 4 documents suivants dans lesquels (pour simplifier l'exercice) seules les occurrences des 5 mots retenus comme termes d'indexation sont fournies, tous les autres mots n'étant pas mentionnés :

Document	**Mots**
Doc1	cétacé, mer, bateau, bateau, cétacé, bateau, mer, bateau, bateau
Doc2	cétacés, mer, mer, océan
Doc3	cétacé, océan, cétacé, océan, cétacé
Doc4	cétacés, cétacés, cétacés

Question 2.1 : (1 pt) Construisez la matrice termes-documents (termes d'indexation en lignes classés par ordre alphabétique, documents en colonnes) en supposant que les termes ne sont pas racinisés (c.-à-.d. aucune lemmatisation ou aucun *stemming* appliqué.e).

Question 2.2 : (2 pts) Construisez la matrice documents-documents dont vous ne remplirez que la partie supérieure, en appliquant la distance de Manhattan entre les documents en ligne k et colonne j.

Question 2.3 : (1 pt) Si on considère que le document Doc4 est une question adressée à une base documentaire formée des 3 premiers documents, dans quel ordre décroissant de pertinence les documents seraient-ils proposés ?

3 Questions diverses, liées au cours - 4,5 points

Pour chaque question ci-dessous, une réponse courte mais précise est attendue.

Question 3.1 : (1,5 pt) Dans plusieurs tâches d'annotation de séquences, la norme IOB (ou BIO) est appliquée. À quoi correspond cette annotation, c.-à-d. à quoi correspondent les lettres I, O et B ?

Annotez, en suivant cette norme, la phrase suivante pour une tâche de détection d'entités nommées où l'on s'intéresse à la mise au point d'un apprentissage supervisé des noms de personnes (PERS), de lieux (LOC), et d'organisation (ORG) : *James Quincey, PDG de Coca-Cola, a annoncé hier à Los Angeles son souhait de racheter l'entreprise Minute Maid avant dimanche.* Pour ce faire, recopiez la phrase et précisez l'annotation sous chaque mot.

Question 3.2 : (2 pts) Parmi les ambiguïtés sémantiques, on distingue la polysémie et l'homonymie. Expliquez la différence entre les deux et citez pour chaque cas un exemple en français non présent

(que ce soit en français ou en anglais) dans les transparents du cours (en explicitant les sens des mots à chaque fois).

Question 3.3 : (1 pt) L'algorithme d'analyse syntaxique CKY est reconnu comme étant efficace en temps. Quel est le principe-clé qui lui permet d'avoir une telle propriété par rapport à des algorithmes d'analyse fondés sur une stratégie *bottom-up* ou *top-down* standard ?

4 Modèles de langue - 8 points

On considère un modèle de type bigramme sur un vocabulaire à trois mots a, b et c, modèle défini par les tables de probabilités suivantes, où ϵ désigne l'historique vide (début de phrase) :

h	w	p
ϵ	a	0.5
ϵ	b	0.5
ϵ	c	0

h	w	p
a	a	0.8
a	b	0.1
a	c	0.1

h	w	p
b	a	0.2
b	b	0.3
b	c	0.5

h	w	p
c	a	0
c	b	0.4
c	c	0.6

Question 4.1 : (1,5 pt) Donnez la probabilité de la phrase $a\ b\ a\ b\ a\ c\ c\ a$ avec ce modèle.

Question 4.2 : (4 pts) Sur un corpus d'apprentissage de 1 000 textes, on observe les comptes suivants pour des échantillons issus de ce modèle :

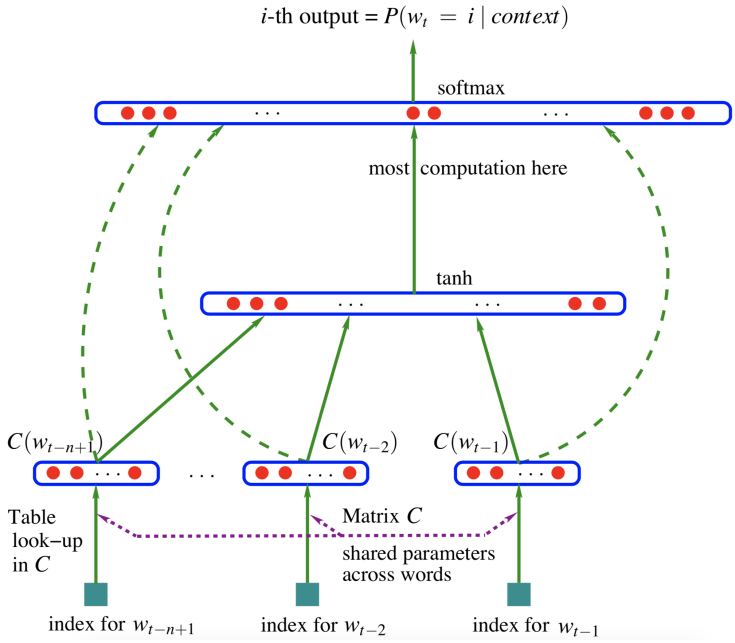
h	a	b	c
ϵ	496	504	0
a	10 054	1 210	1 242
b	2 241	3 368	5 721
c	0	6 531	9 827

En d'autres termes, le mot a a été vu 496 fois en début de phrase, 10 054 fois à la suite du mot a, 2 241 fois à la suite du mot b, *etc*.

— Donnez l'estimation des probabilités bigrammes $P[a|a]$ et $P[a|c]$ en utilisant une technique de lissage additif avec une constante de 2. Comparez les valeurs obtenues aux valeurs correspondantes dans le modèle dont sont issus les textes.

— Une autre technique de lissage possible est celle dite de *absolute discounting*. Expliquez les différences fondamentales entre le lissage additif et *absolute discounting* et discutez l'influence du paramètre de lissage δ sur la qualité des modèles de langue.

Question 4.3 : (2,5 pts) Nous avons vu en cours le modèle neuronal rappelé ci-dessous pour réaliser la modélisation du langage. Précisez les entrées et sorties de ce modèle. Les hypothèses sous-jacentes diffèrent-elles de celle du modèle n-gramme ?


i-th output = $P(w_t = i \mid context)$
softmax
most computation here
tanh
$C(w_{t-n+1})$
$C(w_{t-2})$
$C(w_{t-1})$
Table look-up in C
Matrix C
shared parameters across words
index for w_{t-n+1}
index for w_{t-2}
index for w_{t-1}