Examen CC2

Consignes :

- La durée de cet examen est de 1h.
- Les appareils communicants sont interdits. Seul document autorisé : votre mémo personnel :1 feuille A4 recto/verso. La calculette est autorisée (fortement conseillée).
- Le barème est donné à titre indicatif. L'examen est noté sur 22 points.

1. Question de cours (5 points)

Répondre à 3 questions parmi :

a. Quelles sont les causes majeures du développement des techniques de data mining ?
b. Qu'est-ce que le text mining ? Citez quelques exemples d'applications pour chacun des domaines : fouille descriptive ou prédictive.
c. Lorsque vous travaillez sur une tâche de catégorisation de texte, quels sont les différents choix que vous devez faire en fonction des 3 grands temps à mettre en œuvre ?
d. En quoi consiste le domaine de la détection d'opinion ? Votre réponse devra notamment intégrer la définition d'une opinion selon [Liu,2012] et détailler quels sont les différents niveaux de granularité.
e. Qu'est-ce que le langage figuratif (définition et spécificités) ? Citez deux attributs exploitables pour différencier un texte ironique d'un texte non ironique.

Note : Une réponse assez longue et complète est attendue pour chaque question (5points/3questions=15min/3questions=5min/question environ !)

2. Catégorisation de texte par un arbre de décision sémantique (14 points)

Soit le corpus d'apprentissage suivant :

un hôtel dans le onzième arrondissement, LOC
une chambre double s'il_vous_plait, CHAM
le dernier, LIST
du deux au trois aout, DATE
pour trois personnes, CHAM
un hôtel en plein centre, LOC
je voudrais une chambre double, CHAM
le premier hôtel s'il_vous_plait, LIST

On souhaite construire un arbre de décision sémantique sur ce corpus en tenant compte des prétraitements suivants :

1. lemmatisation
2. suppression des mots de la stop-liste suivante : *un, dans, le, il, du, au, pour, en, vouloir*
3. considération des classes de mots :
   TACH={chambre, hotel}, *CHIF*={deux, onzième, trois},
   POLI={s'il_vous_plait, au_revoir, merci}, *RANG* ={dernier, premier}
4. suppression de tous les mots qui n'apparaissent qu'une seule fois

a. Prétraitements (1pt)

Écrire le corpus d'apprentissage après prétraitements. Il s'agit de celui qui va être traité par l'arbre de décision sémantique de type set-membership.

b. Construction de l'arbre (8 pts)

La pureté d'un ensemble se calculera selon l'indice de Gini qui, pour rappel, se calcule ainsi :

$$Gini(X) = 1 - \sum_{k \in Y} p_k^2$$

i. Écrire l'ensemble des tests qui seront effectivement (par rapport au corpus d'apprentissage) candidats à la racine de l'arbre. Pour chacun, on veut connaître : l'expression régulière testée, le nombre global de documents pour chacune des branches issues de ce test.

ii. Pour tous les tests permettant d'obtenir **au moins 2 éléments dans chaque nœuds fils** : indiquer de manière détaillée le gain résultant.

iii. Choisir la racine de l'arbre parmi ces tests.

iv. Terminer la construction de l'arbre en respectant les règles suivantes :

- Pré-élagage : ne permettre que 3 niveaux en tout (niveau 1 = racine ; niveau2= fils de la racine ; niveau3 = petits fils de la racine qui sont uniquement des **feuilles**).
- Pour le nœud validant l'expression régulière de la racine, tous les tests possibles seront calculés.
- Pour le reste, on ne fera les calculs que pour les tests permettant **au moins 2 éléments dans chaque feuille**.

c. Classification (0,5 pt)

Justifiez la classe qui sera attribuée au document suivant:

« *le dernier hôtel dans le onzième merci* ».

d. Analyse du modèle (1 pt)

Quelles remarques pouvez-vous faire à partir de l'analyse de votre *modèle* ?

e. Évaluation (3,5 pts)

Soit X un ensemble de 10 nouveaux documents à classer. Les hypothèses et les références sur X sont indiquées dans le tableau ci-dessous :

	Réf.	**Hyp.**
1	LOC	LOC
2	LOC	LOC
3	CHAM	CHAM
4	CHAM	CHAM
5	CHAM	LIST
6	LIST	LIST
7	LIST	LIST
8	LIST	DATE
9	DATE	DATE
10	DATE	CHAM

i. Calculer les taux d'erreurs en prédiction et généralisation de votre arbre. Les calculs seront détaillés.

ii. Calculer les macro-mesures et micro-mesures afin d'évaluer le pouvoir de généralisation de votre arbre. Les calculs seront détaillés.

iii. Quelles remarques faites-vous sur les informations apportées par les performances micro/macro ? (réponse courte)

iv. Vos résultats vous semblent-ils significatifs ? Justifiez. (réponse courte)

3. Analyse et commentaires de résultats (3 points)

Lors de la campagne d'évaluation DEFT 2017, le LIUM a participé à deux des trois tâches proposées concernant la détection d'opinion dans les tweets.

Voici les résultats obtenus par le LIUM sur la tâche 1 ainsi que la répartition des classes sur le corpus.

Tâche 1	Entrainement	Test
Objectif	1 642 (42,1%)	411 (42%)
Positif	494 (12,6%)	123 (12,65%)
Négatif	1 268 (32,5%)	317 (32,5%)
Mixte	502 (12,8%)	125 (12,85%)

Table1. Répartition des classes de la tâche 1.

Modèle	Macro F-Score
LSTM-NN	0,48
MLP-NN	0,51
Boost	0,54
LogReg	0,54
Rules-Boost	0,59
Stack-Tree	0,61

Table 2. Résultats obtenus par cross validation sur le corpus d'entrainement par l'équipe du LIUM.

Modèle	Macro F-Score
Minimum	0,23
Moyenne	0,47
Mediane	0,52
Rules-Boost	0,53
LogReg	0,53
Stack-Tree	0,54
Maximum	0,64

Table 3. Résultats obtenus sur le corpus de test. Minimum, Moyenne, Mediane et Maximum présentent les résultats obtenus toutes équipes confondues. Les trois autres systèmes (Rules-Boost, LogReg et Stack-Tree) sont les résultats soumis par l'équipe du LIUM.

1. Quels sont les commentaires que vous auriez écrits dans votre article pour présenter les résultats ?
2. Quelles autres informations auriez-vous aimé avoir pour mieux argumenter vos propos ?