TD : Construction d'arbres de décision

Exercice : Construire un arbre de décision sémantique.

Soit le corpus textuel suivant qui indique si une personne va faire une ballade ou non en fonction du temps.

Il fait beau et chaud, oui	*Le temps est chaud, non*	*Je sors quand il fait super, oui*
Il fait mauvais, non	*Il fera bon, oui*	*Le temps est maussade, oui*

On souhaite construire un arbre de décision sémantique sur ce corpus en tenant compte des prétraitements suivants :
- lemmatisation
- suppression des mots de la stop-liste suivante : il, et, le, quand, être
- considération des classes de mots : {mauvais, maussade}, {super, beau, chaud, bon}

Pour rappel, le critère de séparation d'un SCT est l'indice de Gini :

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

a. Quel est l'ensemble des symboles disponibles à la racine de l'arbre de décision sémantique de type set-membership ?
b. Combien de test vont être faits pour choisir l'expression régulière composant la racine ?
c. Explicitez cet ensemble.
d. Combien d'expressions régulières vous semblent inutiles ?
e. Quel sera le test placé à la racine de l'arbre ?

Examen de 2018-19 : Catégorisation de texte par un arbre de décision sémantique (14 points)

Soit le corpus d'apprentissage suivant :

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

f. 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.

g. Construction de l'arbre (8 pts)

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

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

a. 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.
b. Pour tous les tests permettant d'obtenir **au moins 2 éléments dans chaque nœuds fils** : indiquer de manière détaillée le gain résultant.
c. Choisir la racine de l'arbre parmi ces tests.
d. 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**.

h. Classification (0,5 pt)

Justifiez la classe qui sera attribuée au document suivant:

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

i. Analyse du modèle (1 pt)

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

j. 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

a. Calculer les taux d'erreurs en prédiction et généralisation de votre arbre. Les calculs seront détaillés.
b. Calculer les macro-mesures et micro-mesures afin d'évaluer le pouvoir de généralisation de votre arbre. Les calculs seront détaillés.
c. Quelles remarques faites-vous sur les informations apportées par les performances micro/macro ? (réponse courte)
d. Vos résultats vous semblent-ils significatifs ? Justifiez. (réponse courte)