TD : Construction d'arbres de décision

2. 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 ?

Phase 1 : Lemmatisation
Il faire beau et chaud
Il faire mauvais
Le temps être chaud
Il faire bon
Il sortir quand il faire super
Le temps être maussade

Phase 2 : stop-liste
faire beau chaud
faire mauvais
temps chaud
faire bon
sortir faire super
temps maussade

Phase 3 : classe
Avec P={super, beau, chaud, bon} et N={mauvais, maussade}
1 : faire P P oui
2 : faire N non
3 : temps P non
4 : faire P oui

5 : sortir faire P oui
6 : temps N oui

Vocab ? faire, P, N, temps, sortir ***ET*** +

Rappel de P0 = w, w+, +w, +w+

faire	faire+	+faire	+faire+
P	P+	+P	+P+
N	N+	+N	+N+
temps	temps+	+temps	+temps+
sortir	sortir+	+sortir	+sortir+

Gini de départ ?
2 non et 4 oui → Gini(X)= 1- (2/3)^2 – (1/3)^2 = 0,44
Quel test mettre à la racine?

Faire +
Oui : 2 oui et 1 non → p_oui=3/6=1/2 et Gini(oui)=Gini(X)
Non : 2 oui et 1 non → p_non=3/6=1/2
Gain=Gini(X)-p_oui*Gini(Faire+=oui)-p_non*Gini(Faire+=non)
= 0,44-0,5*0,44-0,5*0,44=0

+Faire+ (= +P+ = sortir+)
Oui : 1 oui et 0 non → p_oui=1/6 et Gini(oui)=0
Non : 3 oui et 2 non → p_non=5/6 et Gini(non)=1-(3/5)^2-(2/5)^2=0,48
Gain=0,44-(5/6)*0,48=0,04

+P = +N = temps+
Oui : 3 oui et 1 non
Non : 1 oui et 1 non
Gain = 0,0233

Niveau0 : RACINE = sortir +
Sortir +
Oui : 1 oui
Non : 3 oui et 2 non ← C'est elle qui n'est pas terrible !!
Ici le Gini est de 0,48

Corpus de la branche non :
1 : faire P P oui
2 : faire N non
3 : temps P non
4 : faire P oui
6 : temps N oui

faire	faire+	+faire	+faire+
P	P+	+P	+P+
N	N+	+N	+N+
temps	temps+	+temps	+temps+

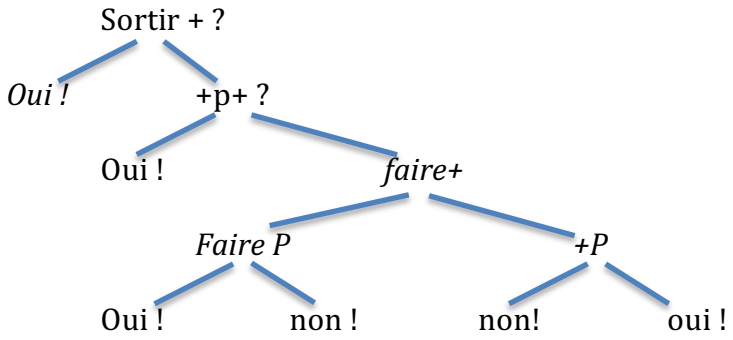
Niveau 1, branche non

faire+ (=+N=temps+=+P)
oui : 2 oui et 1 non : gini(oui)=1-(2/3)^2-(1/3)^2=0,4444=4/9
non : 1 oui et 1 non : gini (non) = 0,5 ! max
Gain = 0,48- (3/5)*4/9– (2/5)*0,5
Gain=0,0133 (au lieu de 0,016 car j'avais arrondi à 0,44…)

+P+
Gain=0,08 GAIN MAX !

~~~~~Encore des calculs~~~~~

Un des arbres possible au final :


~~~~~