Projet 4

Etudes et expérimentation de la reconnaissance des textes avec le SOM

Groupe : LE Viet Man
HO The Nhan

INTRODUCTION

- Etudier la classification par le SOM sur la reconnaissance de textes
- Expérimenter avec les méthodes : SVM et réseau rétro-propagation
- La base de données : 10 groupe de nouvelles
  - graphique, hardware, sports, sciences, politique, etc.

NOTRE METHODE

B.H.ChandraShekar, Dr.G.Shoba, *Classification Of Documents Using Kohonen's Self-Organizing Map*, International Journal of Computer Theory and Engineering, Vol. 1, No. 5, December, 2009

67%

1. Liste d'arrête : a, any, the, of, about,, be, between, each, etc.
   - des mots inutiles ou ordinaires
2. Vocabulaire : des mots dans la base d'apprentissage
   - éliminer des mots apparaissant trop peu dans les documents
   - stemming (éliminer des suffixes) : Porter stemming algorithm
3. Représentation des documents : un vecteur des mots
   - fréquence du terme et fréquence inverse de document (tf-idf)
4. Réduction de dimensions : ACP
5. SOM

EVALUATIONS

- Trois vocabulaires :

1. Premier : 2118 mots
2. Deuxième : 1172 mots
3. Troisième : 545 mots (construit par la main)

PREMIER VOCABULAIRE

Rate	HW	Iteration	Vote	topo	neighbour	seed	Result
0.3	8	1440	FAUX	hexa	8	1	20.6%
0.5	-	-	-	-	-	-	26.9%
0.7	-	-	-	-	-	-	22.5%
-	-	-	VRAI	-	-	-	26.9%
-	15	-	-	-	-	-	34.8%
-	30	-	-	-	-	-	37.5%
-	-	150	-	-	2	-	44.6%
-	-	100	-	-	-	-	45.2%
0.5	**35**	**100**	**VRAI**	**hexa**	**2**	**32**	**54.83%**

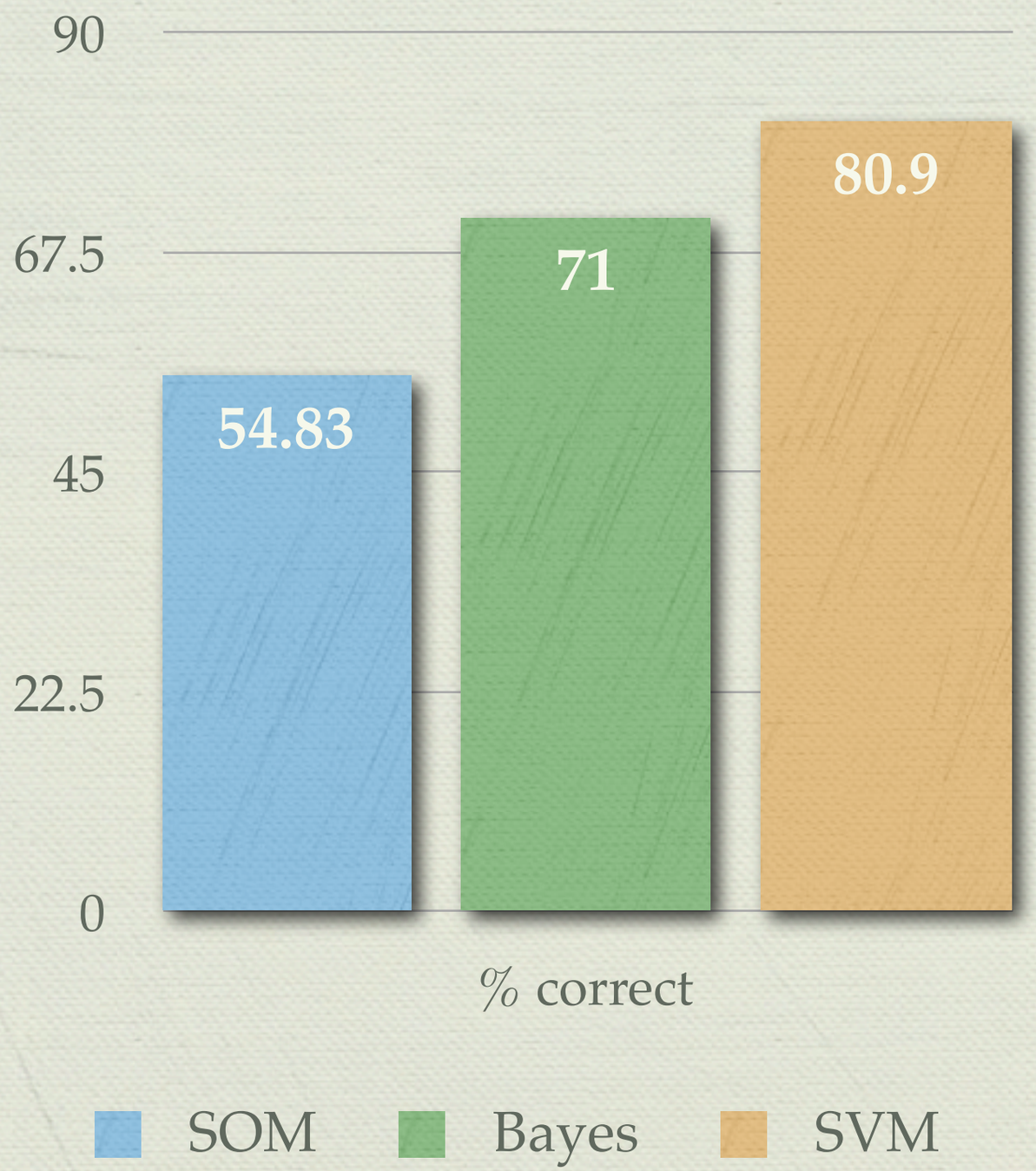
DEUXIEME VOCABULAIRE

Rate	HW	Iteration	Vote	topo	neighbour	seed	Result
0.5	30	100	VRAI	hexa	2	1	46.1%
-	-	-	-	rec	-	-	45.5%
-	-	-	-	-	4	-	28%
-	-	-	-	hexa	8	-	25%
0.3	-	-	-	-	2	-	46%
0.5	35	35	-	-	-	-	48%
-	-	-	-	-	-	16	50.8%
0.5	**35**	**35**	**VRAI**	**hexa**	**2**	**32**	**53%**

TROISIEME VOCABULAIRE

Rate	HW	Iteration	Vote	topo	neighbour	seed	Result
0.3	8	1440	FAUX	hexa	8	1	26.47%
-	30	-	-	-	-	-	39%
0.5	**-**	**-**	**-**	**-**	**2**	**-**	**40.33%**
1	-	-	-	-	-	-	38%
0.7	-	-	-	-	-	-	38.8%
-	40	-	-	-	-	-	37.7%

COMPARAISON DES METHODES



CONCLUSION

- Notre résultats sont un peu similaires avec le résultat de l'article, mais encore trop bas.
- Le meilleur résultat obtenu est 54.83% avec le vocabulaire de 2118 mots
- Raisons :
  - Des faux dans l'algorithme Porter stemming : eye -> ey, apple -> appl
  - Il n'y a pas l'étape ACP
- Mauvaises classes sont hockey, politique, médecin, ibm.hardware

MERCI DE VOTRE ATTENTION