UNIVERSITÉ
DE MONTPELLIER

WEKA
The University
of Waikato

Classification de documents

Extraction de connaissances à partir de données

Geoffrey DUMAS - Olivier SAINT-PAUL - Quentin PHILBERT - Thibaut CASTANIE

23 avril 2015

Sommaire

1. Introduction
2. Constitution du corpus
3. Format ARFF
4. Phase de pré-traitement
5. Résultats et analyses
6. Conclusion

1.Introduction

Sujet : Classifier des documents et en faire l'analyse

Outil : WEKA

3 types de textes:

- Brut
- Lemmatisé
- Lemmatisé avec analyse morpho-syntaxique

2. Constitution du corpus

- Corpus issu du site L'Equipe.fr

- Classification en trois catégories :
    - Basket
    - Tennis
    - Rugby

3.Format ARFF

```
@relation sport

@attribute document_content string
@attribute document_class {basketball,tennis,rugby}

@data
"Marin Cilic remporte, à l'US Open, son premier titre du Grand Chelem
Dans une finale entre novices [...]",tennis
"La France revient dans la course grâce à Mladenovic Grâce à un
excellent match remporté (6-4, 6-3) contre Sara Errani, Kristina
Mladenovic permet à l'équipe de France de revenir à deux points à un
contre l'Italie [...]",tennis
```

L'Equipe.fr

http://www.lequipe.fr/

Les Spurs plus forts que les Lakers

Après avoir hiberné durant l'hiver, les San Antonio Spurs sont montés en puissance jusqu'à décrocher un record : celui du meilleur ratio victoires-défaites de l'histoire de la saison régulière NBA, devant les Los Angeles Lakers.
Même si leur jeu tout en contrôle et en collectif est très européanisé, c'est bien sur le championnat nord-américian que les San Antonio Spurs ont posé la main. Les Texans sont devenus la nouvelle référence en peu de temps ...

François Trinh-Duc veut «mourir les armes à la main»

François Trinh-Duc : «Personnellement, je me suis senti bien dans le match même s'il y avait beaucoup plus de rythme que contre le LOU. Je n'ai pas buté car c'est difficile de demander à Ben Lucas de laisser sa place. Il est plutôt performant dans l'exercice. Et puis, je prends les choses les unes après les autres même si je m'entraîne énormément pour ça aussi. Pour le match, je suis fier que l'on n'ait pas lâché alors qu'on perdait de 14 points. On a été cherché des choses très profond, des trucs qu'on n'avait plus vu chez nous depuis longtemps. Mais au final, même si on a fait match nul, par rapport à la qualification, ce sera compliqué. Pourtant, on veut mourir les armes à la main.»

Lyon met la pression sur le PSG

Lyon restait sur treize matches consécutifs avec au moins un but marqué sur sa pelouse. Mercredi, les spectateurs de Gerland ont craint que cette série ne prenne fin. Si l'OL a nettement dominé sa rencontre en retard contre Bastia (2-0), il a patienté jusqu'au dernier quart d'heure pour faire la différence. Mohamed Yattara, tout juste entré en jeu, a d'abord repris un centre précis du revenant Clément Grenier (77e). Au terme d'une chevauchée, le véloce Clinton Njie a ensuite servi Alexandre Lacazette (85e), efficace devant Alphonse Areola. Le meilleur buteur de la L1 est devenu le premier joueur lyonnais depuis André Guy (1968/1969) à marquer 25 buts ou plus lors d'une seule saison.

1

- basketball
  - basketball1.txt
  - basketball2.txt
  - basketball3.txt
  - ...
- tennis
  - tennis1.txt
  - tennis2.txt
  - tennis3.txt
  - ...
- rugby
  - rugby1.txt
  - rugby2.txt
  - rugby3.txt
  - ...

2 JAVA

CORPUS.ARFF

```
@relation sport

@attribute document_content string
@attribute document_class {basketball,tennis,rugby}

@data
"Les Spurs plus forts que les Lakers Après avoir hiberné durant l'hiver, les San Antonio Spurs sont montés en puissance jusqu'à décrocher un record : celui du meilleur ratio victoires-défaites de l'histoire de la saison régulière NBA ...", basketball
```

Résumé

4.Phase de pré-traitement

Stopwords :

- Textes en Français
- Creation de notre propre fichier

Filtre appliqué : StringToWordVector

TF-IDF

4.Phase de pré-traitement

Lemmatisation :

a	avoir
reconnu	reconnaître
l’	le
élève	élève
de	de
Michael	Michael
Chang	Chang

4.Phase de pré-traitement

Analyse morpho-syntaxique (TreeTagger):

a	VER :pres	avoir
reconnu	VER :pper	reconnaître
l’	DET :ART	le
élève	NOM	élève
de	PRP	de
Michael	NAM	Michael
Chang	NAM	Chang

5.Résultats et analyses

A. Textes bruts

5.Résultats et analyses

a. Naïve Bayes

a	b	c	classé dans
20	0	0	a = basketball
0	20	0	b = tennis
0	0	22	c = rugby

100%

5.Résultats et analyses

b. Machine à support de vecteurs

a	b	c	classé dans
19	0	1	a = basketball
0	20	0	b = tennis
0	0	22	c = rugby

98,3871%

5.Résultats et analyses

c. K plus proche voisin (K=1)

a	b	c	classé dans
19	1	0	a = basketball
2	18	0	b = tennis
20	1	1	c = rugby

61,2903%

5.Résultats et analyses

c. K plus proche voisin (K=4)

a	b	c	classé dans
16	4	0	a = basketball
0	20	0	b = tennis
0	18	4	c = rugby

64,5161%

5.Résultats et analyses

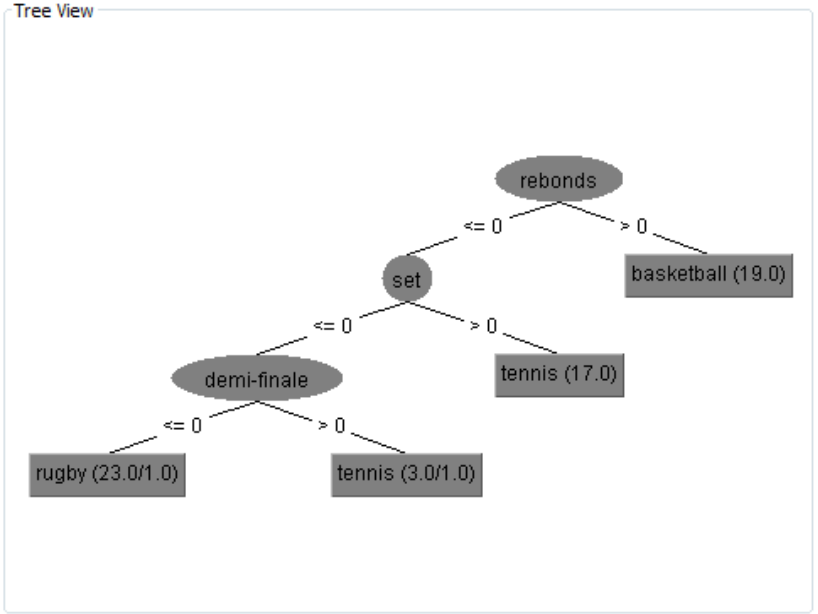
d. Arbre de décision

a	b	c	classé dans
19	0	1	a = basketball
0	16	4	b = tennis
0	1	21	c = rugby

90,3226%

5.Résultats et analyses

d. Arbre de décision



90,3226%

5.Résultats et analyses

B. Textes Lemmatisés

Naïve Bayes : **100%**

SMO : **98,3871%**

iBk k=1 : **59,6774%**

j48 : **91,9335%**

iBk k=4 : **59,6774%**

5.Résultats et analyses

C. Textes lemmatisés avec analyse morpho-syntaxique (noms et verbes)

Naïve Bayes : **100%**

SMO : **98,3871%**

iBk k=1 : **61,2903%**

j48 : **91,9335%**

iBk k=4 : **56,4516%**

5.Résultats et analyses

C. Textes lemmatisés avec analyse morpho-syntaxique (noms uniquement)

Naïve Bayes : **100%**

SMO : **96,7742%**

iBk k=1 : **54,8387%**

j48 : **88,7097%**

iBk k=4 : **53,2258%**

5.Résultats et analyses

C. Textes lemmatisés avec analyse morpho-syntaxique (adjectifs uniquement)

Naïve Bayes : **87,0968%**

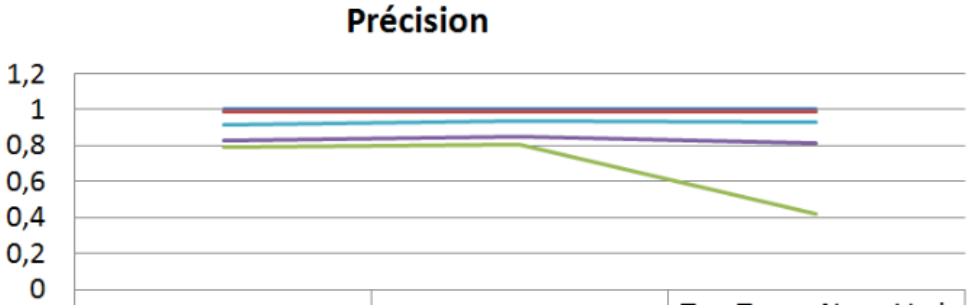
SMO : **75,8065%**

iBk k=1 : **56,4516%**

j48 : **56,4516%**

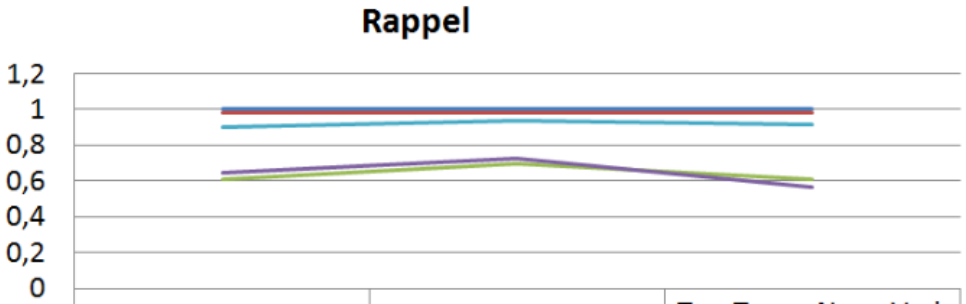
iBk k=4 : **45,1613%**

5.Résultats et analyses



	Brut	Lemmatisé	TreeTaggerNomsVerbes
Bayes	1	1	1
SMO	0,985	0,985	0,985
IBK KNN=1	0,795	0,806	0,422
IBK KNN=4	0,831	0,852	0,815
J48	0,913	0,941	0,934

5.Résultats et analyses



	Brut	Lemmatisé	TreeTaggerNomsVerbes
Bayes	1	1	1
SMO	0,984	0,984	0,984
IBK KNN=1	0,613	0,694	0,613
IBK KNN=4	0,645	0,726	0,565
J48	0,903	0,935	0,919

6. Conclusion