Catégorisation automatique de textes

Agathe Mollé

Introduction

Nous nous intéressons à la tâche de catégorisation (classification) automatique de textes. Cette tâche consiste à employer de l'apprentissage supervisé afin de déterminer à quelle catégorie (classe) appartient un texte.

Ce rapport présente une comparaison des principales méthodes de catégorisation, ainsi qu'une comparaison de sélection de caractéristiques.

Corpus

Nous utilisons les données du corpus Reuters, comportant 10 788 documents, soit 1,3 millions de mots. Ces documents sont étiquetés selon 90 catégories, chaque document possédant en général plusieurs catégories. Le découpage des ensembles d'entraînement (70%) et de test (30%) est déjà effectué.

Méthodes évaluées

Nous cherchons à comparer certaines méthodes de classification ainsi que des méthodes de sélection de caractéristiques.

Catégorisation Nous nous intéressons aux méthodes de catégorisation suivantes :

- Classification naïve de Bayes (Multinomiale, Bernoulli)
- Séparateur à vaste marge (SVM)
- K plus proches voisins
- Algorithme de Rocchio
- Réseau de neurones : le perceptron

Sélection de caractéristiques En ce qui concerne la sélection de caractéristiques, nous testons deux méthodes :

- Fréquence de document pondérée (tf-idf)
- Manque d'indépendance entre un mot et une catégorie (χ^2)

Implémentation

Les tests ont été effectués à l'aide d'un script python. Celui-ci utilise les bibliothèques nltk et sklearn. Pour exécuter le script, il suffit de lancer la commande :

```
python categ.py
```

Le code est disponible sur Github.

Mesures

Pour évaluer la pertinence des classifieurs, nous utilisons les mesures classiques d'apprentissage supervisé, à savoir la précision, le rappel ainsi que la F-mesure.

Résultats

Voici les résultats obtenus pour les différents classifieurs étudiés :

	Précision	Rappel	F-mesure
tf-idf	.42	.53	.47
χ^2	.66	.68	.67

TABLE 1 – Multinomial Naive Bayes

	Précision	Rappel	F-mesure
tf-idf	.46	.56	.50
χ^2	.67	.64	.65

TABLE 2 – Bernoulli Naive Bayes

	Précision	Rappel	F-mesure
tf-idf	.73	.75	.74
χ^2	.70	.71	.71

TABLE 3 – SVM

	Précision	Rappel	F-mesure
tf-idf	.67	.69	.68
χ^2	.62	.63	.62

TABLE 4 – K plus proches voisins

	Précision	Rappel	F-mesure
tf-idf	.71	.69	.70
χ^2	.69	.60	.64

TABLE 5 – Algorithme de Rocchio

	Précision	Rappel	F-mesure
tf-idf	.73	.70	.72
χ^2	.66	.64	.65

TABLE 6 – Perceptron

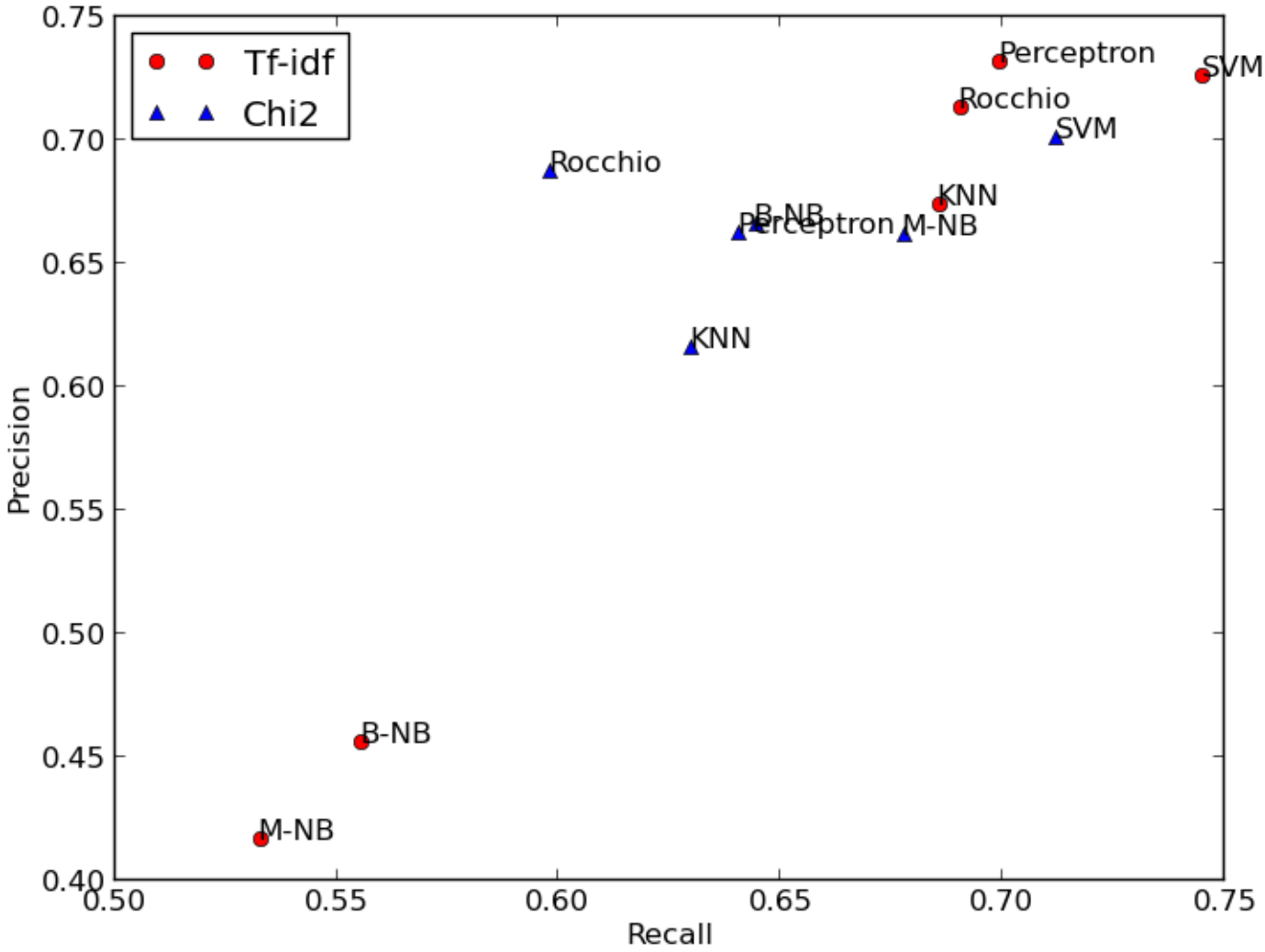

FIGURE 1 – Comparaisons des différentes de méthodes de catégorisation

Discussion

Les classifieurs naïfs de Bayes sont moins efficaces lorsque les caractéristiques s'appuient sur la fréquence des mots dans le document. Ils sont cependant plus intéressants en utilisant la mesure χ^2.

De manière générale, la catégorisation naïve de Bayes et les K plus proches voisins obtiennent des résultats moins intéressants que la méthode des centroïdes, celle du Perceptron ainsi que SVM.

Il aurait été intéressant d'étudier la catégorisation par arbres de décisions, mais nltk ne permet pas de les implémenter correctement, du moins nous n'avons pas su l'exploiter.

En définitive, ces résultats permettent d'obtenir un aperçu des principales méthodes de catégorisation. Cependant, cette étude ne prend pas en compte la multiplicité des catégories attribuées à un même document. En effet, les mesures ne s'intéressent qu'à la classe la plus représentée. Il serait donc intéressant d'entraîner un classifieur par catégorie, soit 90 dans notre cas.