# Catégorisation des questions/réponses

### XU Yizhou JIANG Chunyang



27 mars 2019

- 1 Introduction
- 2 Données
- 3 Ingénierie de caractéristiques
- 4 Sélection de classifieurs
- **5** Paramétrage de classifieurs
- **6** Évaluation

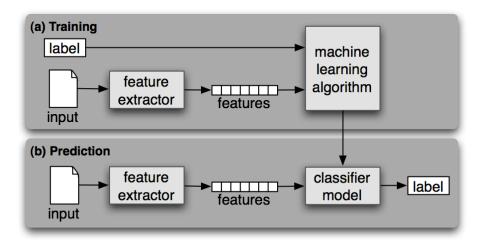


Figure – classification supervisée

(source: http://www.nltk.org/)

- 1 Introduction
- 2 Données
- 3 Ingénierie de caractéristiques
- 4 Sélection de classifieurs
- **6** Paramétrage de classifieurs
- **6** Évaluation

- Introduction
- 2 Données
- 3 Ingénierie de caractéristiques
- 4 Sélection de classifieurs
- **5** Paramétrage de classifieurs
- **6** Évaluation

# Jeu de données

- multi-classe (8)
- $\Rightarrow immobilier (imm)$
- $\Rightarrow travail (trv)$
- $\Rightarrow$  entreprise (ent)
- ⇒ personne et famille (per)
- ⇒ finances, fiscalité et assurance (fin)
- ⇒ société (soc)
- $\Rightarrow$  justice (jus)
- ⇒ internet, téléphonie et prop. intellectuelle (int)

### Jeu de données

# • déséquilibre

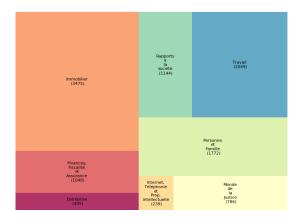


Figure – Distribution de classes

# Jeu de données

- Taille: 12308 doc(questions/réponses)
- Répartition :
  - $\triangleright$  train 80%(9848)
  - $\triangleright$  test 20%(2460)

- Introduction
- 2 Données
- 3 Ingénierie de caractéristiques
- 4 Sélection de classifieurs
- **5** Paramétrage de classifieurs
- **6** Évaluation

# Caractéristiques

- Caractéristiques lexicales (sac de mots)
- Caractéristiques sémantiques (plongement)
- Caractéristiques d'apprentissage en profondeur
- Caractéristiques à base de règles

# Caractéristiques Sac de X

### Mot (Token ou Unigram)

« notre », « syndic », « indique », « cette »

#### Lemma + POS

- « notre/DET :pos », « syndic/NOM »,
- « indiquer/VER :pres », « ce/DET :dem »

### N-Gram+Lemma (Bigram, Trigram)

« notre syndic », « syndic indiquer », « indiquer ce »

# Représentation

- présence/absence
- nombres d'occurrences (fréquence)
- tf-idf
- $\chi^2$
- information mutuelle
- ...

# Filtrage et Nettoyage

 ${\bf Outil:} \ \ sklearn. feature\_extraction. text. \ Tfidf Vectorizer$ 

- mots vides et mots fréquents (corpus-specific stop words)
- $\Rightarrow \max_{d} = 0.7$ 
  - mots rares (cut-off)
- $\Rightarrow \min_{df=5}$
- accents
- ⇒ strip\_accents='ascii'

Premature optimization is the root of all evil.

— Donald Knuth

- Introduction
- 2 Données
- 3 Ingénierie de caractéristiques
- 4 Sélection de classifieurs
- **5** Paramétrage de classifieurs
- **6** Évaluation

# Classifieurs

- Baseline (BL)
- Random Forest (RF)
- Gradient Boosting (GB)
- Logistic Regression (LR)
- Naive Bayes (NB)
- Support Vector Machines (SVM)

### scikit-learn

- $\Rightarrow$  DummyClassifier
- $\Rightarrow$  RandomForestClassifier
- $\Rightarrow$  GradientBoostingClassifier
- $\Rightarrow$  LogisticRegression
- $\Rightarrow$  ComplementNB
- $\Rightarrow$  LinearSVC

- Introduction
- 2 Données
- 3 Ingénierie de caractéristiques
- 4 Sélection de classifieurs
- **5** Paramétrage de classifieurs
- **6** Évaluation

# Paramétrage

- Outil: sklearn.model\_selection.GridSearchCV
- Exemple :

#### **Random Forest**

- \* n\_estimators :  $20,30,40,...,200 \Rightarrow 130$
- \*  $max_depth : 10,20,30,...,100 \Rightarrow 60$
- \* min\_samples\_split : 2, 5,  $10 \Rightarrow 10$
- \* min\_samples\_leaf : 1, 2,  $4 \Rightarrow 2$

- Introduction
- 2 Données
- 3 Ingénierie de caractéristiques
- 4 Sélection de classifieurs
- **5** Paramétrage de classifieurs
- **6** Évaluation

# Évaluation

	$\mathbf{BL}$	$\mathbf{RF}$	$\mathbf{G}\mathbf{B}$	$\mathbf{L}\mathbf{R}$	NB	SVM
Token	0.18	0.72	0.77	0.77	0.76	0.78
Lemma+POS	0.18	0.72	0.76	0.78	0.76	0.78
N-Gram	0.19	0.72	0.75	0.76	0.69	0.79

Table - Micro F1

ent	52.6% 50/95	6.3% 6	7.4% 7	6.3% 6	1.1%	6.3% 6	11.6% 11	8.4% 8	- 600
- fin	1.6% 4	57.1% 145/254	12.2% 31	0.4% 1	3.9% 10	10.6% 27	8.3% 21	5.9% 15	000
inm.	0.9% 7	1.6% 12	91.3% 685/750		1.5% 11	2.4% 18	1.3% 10	0.9% 7	- 450
	1.9%		3.8%	58.5% 31/53	5.7% 3	7.5% 4	22.6% 12		
Actual jus int	1.0%	6.5% 13	9.5% 19	1.0%	49.8% 100/201	11.4% 23	17.4% 35	3.5% 7	- 300
ber		4.4% 17	3.9% 15	1.0% 4	3.6% 14	83.2% 323/388	2.1% 8	1.8% 7	
SOC	3.0%	8.2% 22	9.7% 26	1.9% 5	8.2% 22	5.2% 14	58.0% 156/269	5.9% 16	- 150
소 -	1.1% 5	1.1% 5	0.4%	0.9% 4	0.9% 4	0.7% 3	2.0%	92.9% 418/450	
	ent	fin	imm	int Pred	jus icted	per	soc	trv	- 0

Figure – SVM - matrice de confusion