inalco
Institut national
des langues
et civilisations orientales

XiKO

Étude sur l'apport de la sélection des caractéristiques dans la classification multi-classe des textes

Master 2 Ingénierie Multilingue
Auteur : Yuming ZHAI

Directeur de mémoire : Damien NOUVEL
Encadrant de stage : Gaël PATIN

1. **Contexte et objectif d’étude**
2. État de l’art
3. Méthodes de sélection comparées
4. Algorithmes d’apprentissage appliqués
5. Corpus et algorithme
6. Expériences et résultats
7. Conclusion et perspectives

Contexte d’étude

- travail réalisé chez XiKO, spécialisée dans l’étude des Big Data conversationnels
- le service KOVERI™ exploite des Big Data pour améliorer les performances des campagnes publicitaires
  - augmenter la capacité de ciblage pour les annonceurs
  - améliorer la monétisation des emplacements sur les pages web pour les éditeurs
- livrer des données analysées sémantiquement correspondant aux URLs :
  - **segments contextuels**
  - **segments comportementaux**

Objectif d’étude

Améliorer la performance du classifieur XiKO multi-classe pour attribuer aussi précisément que possible des **segments contextuels** aux pages web.

- classification **multi-classe** des textes

Difficulté majeure de la classification des textes : grande dimensionnalité des caractéristiques (nombre de features).

- étude sur l’apport de la sélection de caractéristiques

1. Contexte et objectif d’étude
2. **État de l’art**
3. Méthodes de sélection comparées
4. Algorithmes d’apprentissage appliqués
5. Corpus et algorithme
6. Expériences et résultats
7. Conclusion et perspectives

État de l’art

- plusieurs algorithmes d’apprentissage utilisés dans la littérature
- dans les travaux de (Yang and Pedersen, 1997), le gain d’information et le CHI-deux ont été prouvés être les méthodes les plus efficaces
- un classifieur SVM peut éliminer le besoin de la sélection des caractéristiques (Joachims, 1998)
- approche de l’apprentissage non supervisé, à travers une représentation vectorielle des textes (Lin et al., 2015), (Lilleberg et al., 2015), (Eensoo et al., 2015)

Positionnement et hypothèse du travail

Positionnement par rapport aux travaux précédents :

- utiliser notre **propre corpus** dans le contexte applicatif
- choisir les caractéristiques **spécifiques à chaque classe**
- se concentrer sur l'approche de **l'apprentissage supervisé**

Hypothèses sur les avantages de la sélection de caractéristiques :

- pallier le problème du sur-apprentissage
- diminuer le temps d'exécution / occupation de mémoire
- faciliter l'interprétation des caractéristiques sélectionnées

1. Contexte et objectif d’étude
2. État de l’art
3. **Méthodes de sélection comparées**
4. Algorithmes d’apprentissage appliqués
5. Corpus et algorithme
6. Expériences et résultats
7. Conclusion et perspectives

Méthodes de sélection comparées

Spécificité lexicale : distinguer les formes spécifiques à une classe

TF-IDF : estimer la pertinence de discrimination d'un mot dans un document (dans **une classe** dans notre cas) relativement à un corpus

Information mutuelle : mesurer la dépendance statistique entre un mot et une classe

Différence proportionnelle catégorique (CPD) : mesurer à quel point une caractéristique contribue à différencier une classe des autres classes

- chaque caractéristique reçoit un **score** par rapport à chaque classe
- filtrer les caractéristiques par un **seuil prédéfini**

1. Contexte et objectif d’étude
2. État de l’art
3. Méthodes de sélection comparées
4. **Algorithmes d’apprentissage appliqués**
5. Corpus et algorithme
6. Expériences et résultats
7. Conclusion et perspectives

Algorithmes d’apprentissage automatique appliqués

Bayésien Naïf Multinomial

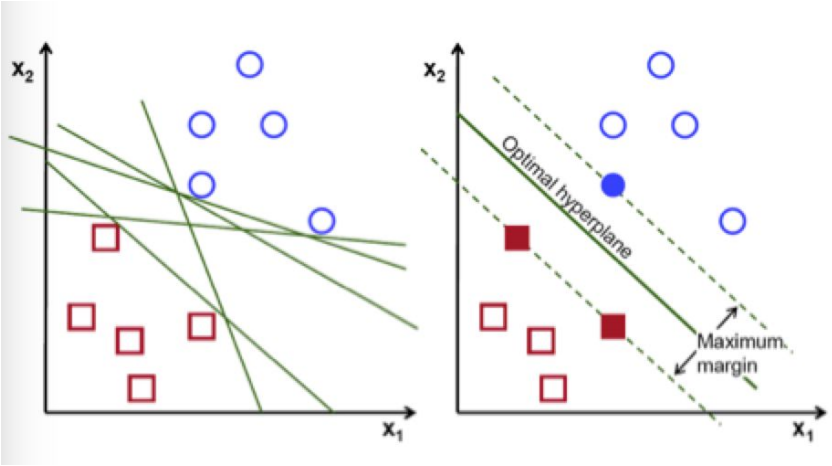
- algorithme principal utilisé
- implémenté dans Weka (Hall et al., 2009)
- modèle probabiliste

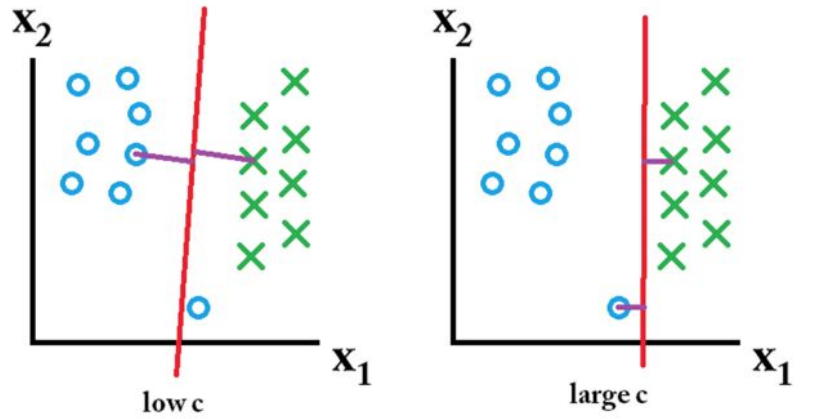
Deux suppositions pour calculer la vraisemblance d’un document :

- hypothèse de « sac de mots »
- « hypothèse bayésienne naïve »

Machines à Vecteurs de Support

- un algorithme qui cherche à trouver un hyperplan avec une marge maximale
- assez robuste pour contrôler le sur-apprentissage grâce aux paramètres de régularisation
- nous utilisons la librairie LIBSVM dans les expériences (Chang and Lin, 2011)

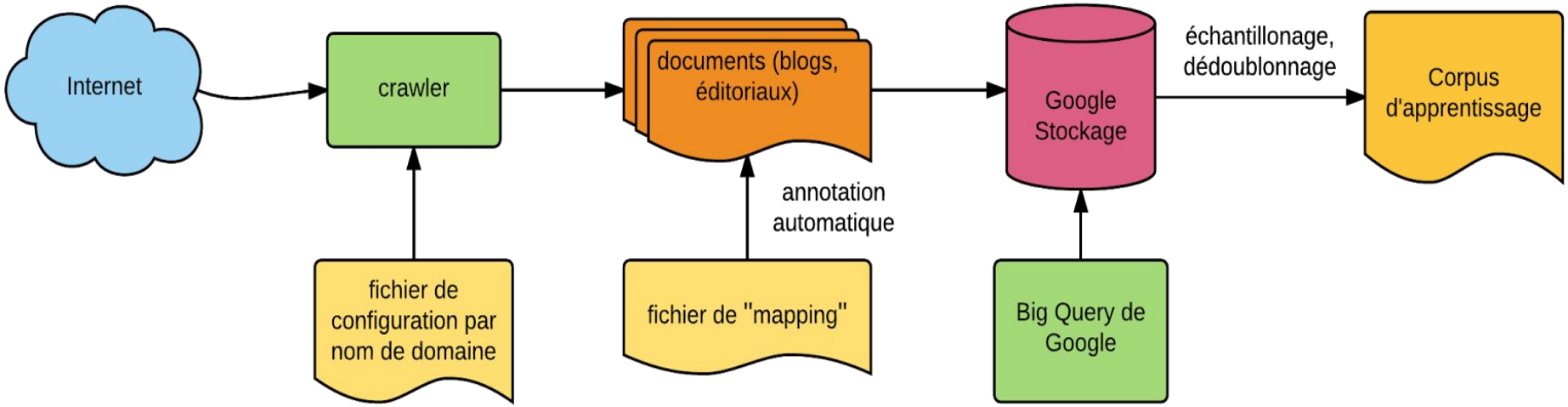




1. Contexte et objectif d’étude
2. État de l’art
3. Méthodes de sélection comparées
4. Algorithmes d’apprentissage appliqués
5. **Corpus et algorithme**
6. Expériences et résultats
7. Conclusion et perspectives

Corpus et algorithme

Génération du corpus



Annoter automatiquement les documents en s'appuyant sur des fil d'ariane ou URL :

*<map territory="maman.*conception" **topic**="xiko:topic.health_well_being-pregnancy_birth"/>*
<map url="./agriculture-alimentation/Le-Vin.*" **topic**="xiko:topic.food_drink-alcohol-wine"/>*

Statistiques du corpus

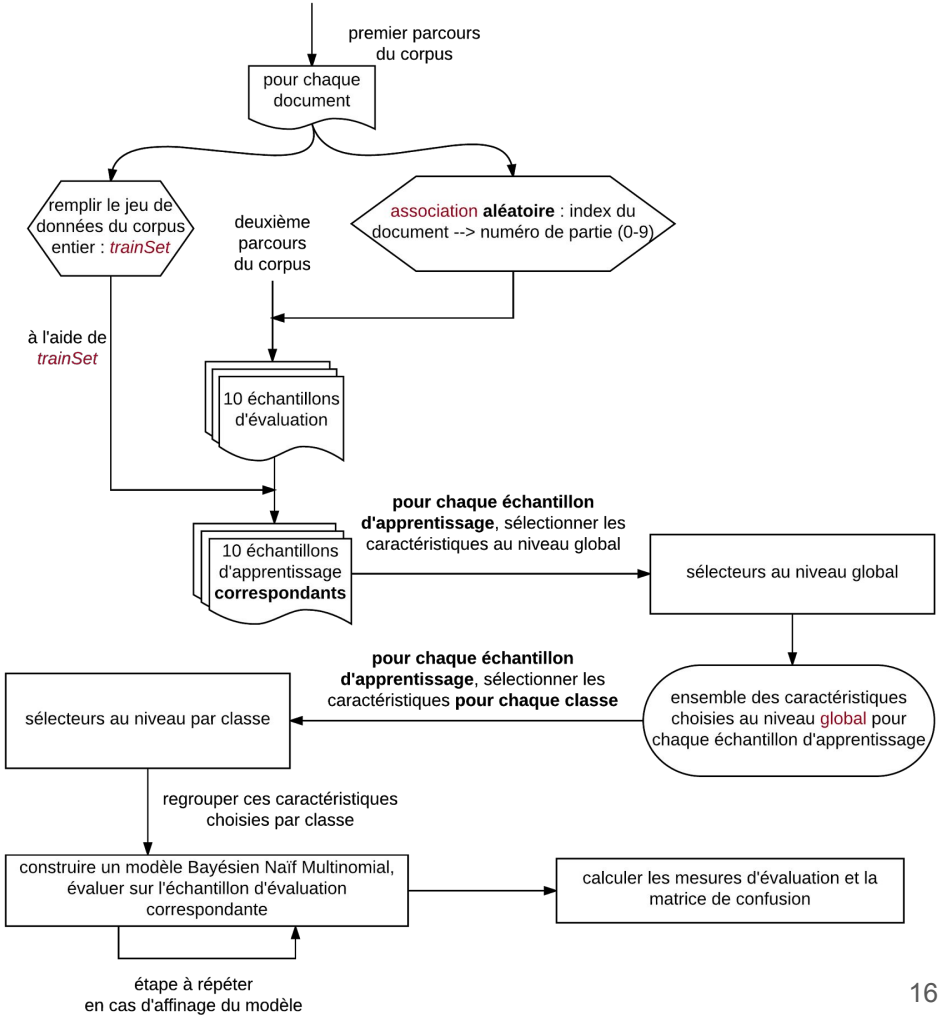
Nom de la catégorie	Pourcentage de la catégorie	Nom de la catégorie	Pourcentage de la catégorie
sports	23,70%	*home & garden*	1,99%
culture & entertainment	14,99%	*family & parenting*	1,63%
health & well-being	7,26%	*pets*	1,48%
business	5,71%	*beauty*	1,31%
politics	5,60%	*legal issue*	1,30%
technology	4,93%	*real estate*	1,24%
society	4,91%	*education*	0,92%
news	4,72%	*weddings*	0,64%
science	3,77%	*outing*	0,51%
environment	2,96%	*automotive*	0,50%
travel	2,44%	*spirituality*	0,42%
personal finance	2,26%	*dating & love & couple*	0,35%
cooking	2,10%	*careers*	0,29%
style & fashion	2,07%		

- nous ne détectons que les catégories du premier niveau
- langue : français (et un peu d'anglais)
- 136 821 documents, longueur moyenne 370 tokens, écart-type 360 tokens

Algorithme du processus

- sélecteurs au niveau global :
  - **nombre d'occurrence minimal ;**
  - longueur minimale ;
  - ignorer les nombres ;
  - ignorer les mots-outils

- sélecteurs au niveau local :
  - nombre d'occurrence minimum dans une classe ;
  - pourcentage d'occurrences minimum dans une classe ;
  - **seuil de la méthode de sélection ;**
  - nombre de caractéristiques retenues par classe



1. Contexte et objectif d’étude
2. État de l’art
3. Méthodes de sélection comparées
4. Algorithmes d’apprentissage appliqués
5. Corpus et algorithme
6. **Expériences et résultats**
7. Conclusion et perspectives

- classe majoritaire (*sports*) : 23,70%
- micro-moyenne F-mesure : évaluation sur tous les documents (des classes importantes ont plus de poids)
- macro-moyenne F-mesure : évaluation par classe puis faire la moyenne (toutes les classes ont le mêmes poids)

Sans application de méthode de sélection

- ignorer les nombres et les mots-outils
- filtrer les mots avec une faible fréquence

	nombre d'occurrences minimal = 0	nombre d'occurrences minimal = 3	nombre d'occurrences minimal = 10
Nb. caractéristiques en moyenne	320 787	158 216	81 710
Micro-moyenne F1	-	**72,26%**	72,15%
Macro-moyenne F1	-	59,32%	**62,30%**

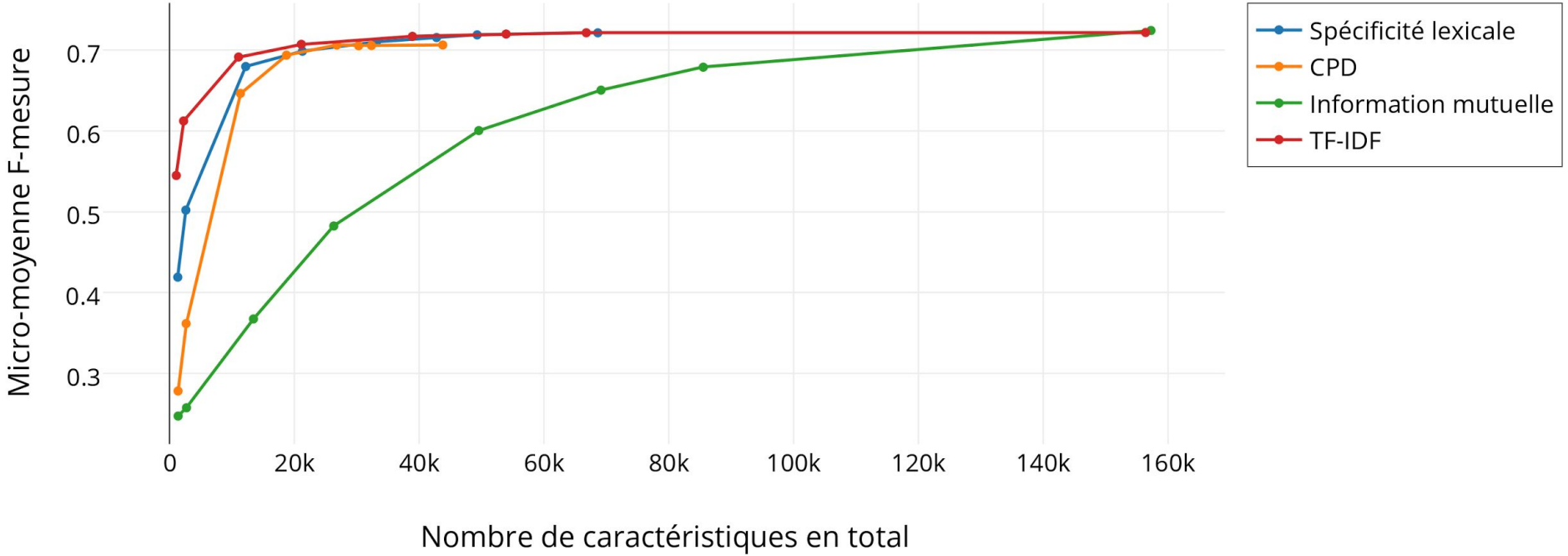
Comparaison entre les méthodes de sélection

- **configuration générale des expériences**
  - ignorer les nombres et les mots-outils
  - mettre tous les mots en minuscule
  - garder les accents
  - nombre d'occurrences minimal : **3**
  - raciniser (optionnel)
  - prendre en compte les spécificités négatives (optionnel)

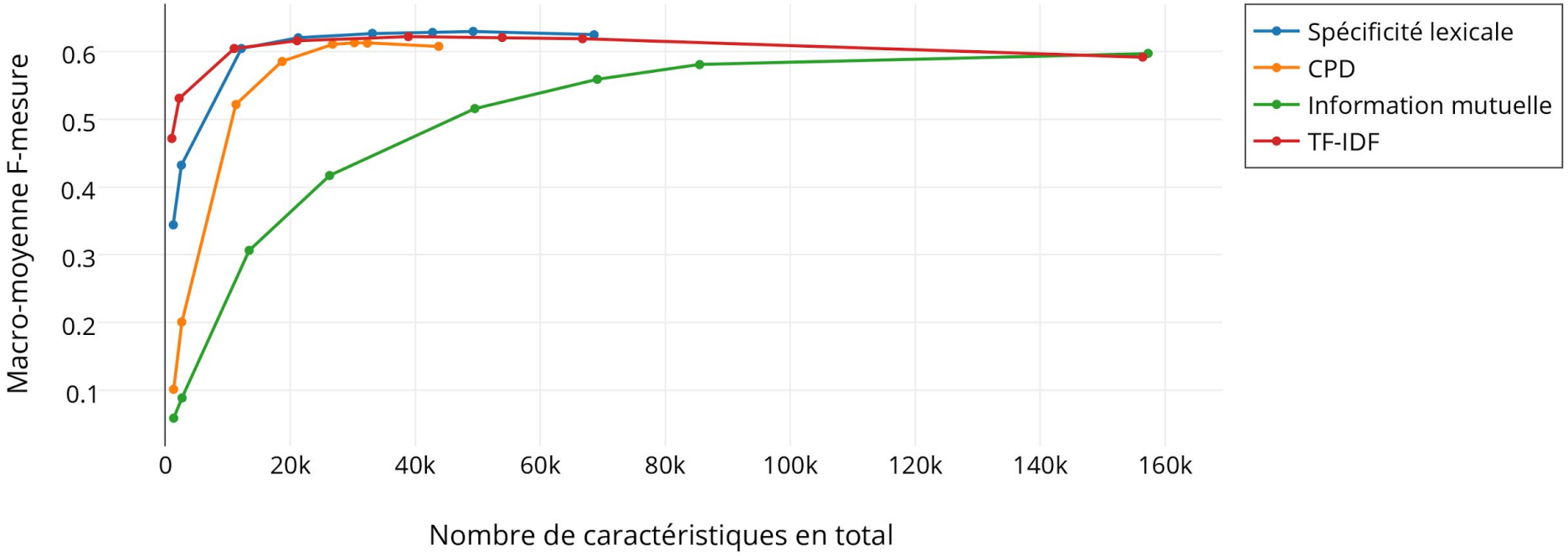
- détermination du meilleur seuil pour chaque méthode de sélection :

spécificité lexicale	CPD	TF-IDF	Information Mutuelle
< 0.00001	> 0.99	> 0.00	> 1.00

- filtrage des N premières features par classe triées selon le critère de sélection : 50, 100, 500, 1000, 2000, 3000, 4000, sans limite (jusqu'au seuil)



- la macro-moyenne F-mesure est toujours inférieure que la micro-moyenne F-mesure



Quand la limite du nombre de caractéristiques retenues par classe est supprimée :

- des catégories difficiles à classer (*careers, dating_love_couple, outing*)

Méthode	Sans méthode de sélection	Spécificité lexicale	CPD	TF-IDF	Information mutuelle
Nb. caractéristiques en moyenne	158 216	**68 618**	43 762	156 404	157 267
Micro-moyenne F1	72,26%	72,14%	70,64%	72,18%	**72,43%**
Macro-moyenne F1	59,32%	**62,46%**	60,76%	59,17%	59,72%
F1 sur des classes difficiles	1,47%, 17,02%, 23,69%	**19,43%,** 44,85%, **33,40%**	17,16%, 45,18%, 32,67%	7,22%, 25,08%, 22,50%	1,94%, 20,31%, 22,70%

- D’autres prétraitements essayés : racinisation, lemmatisation

- les meilleures caractéristiques choisies pour la classe *automotive*

Spécificité lexicale	CPD	TF-IDF	Information mutuelle
parigné	hypersport	crtg	ncap
alltrack	linian	seniorplanet	zaniroli
gle	ncap	smartads	picanto
inrix	zaniroli	formatid	cyl
acda	vendezvotrevoiture	pageid	féminin scritèrespeut
chevrolet	êvêque	split	hélica
talisman	picanto	cookie	bodrogi
stationnements	cyl	undefined	motobot
méhari	kidioui	typeof	fcv
uppsala	niedereimer	sponso	airbumps
logan	boxster	name	barbanti
berlines	verbalisateur	pave	lambroghini
i	tnga	var	fxs
abs	féminin scritèrespeut	sas	laliron
capote	hélica	if	tricorps
carrera	nadjowski	target	lykan
tfsi	jovanka	planet	xje
fortwo	mwt	function	radardroid
contraventions	motospot	radars	airbump

Caractéristiques en n-grammes

- on perd des informations rajoutées par les n-grammes, tel que « collection haute couture »
- filtrer les n-grammes qui contiennent des mots-outils, des nombres ou des symboles
- sélectionner par la spécificité lexicale

	Unigrammes	Bigrammes	Trigrammes
Nb. caractéristiques en moyenne	68 618	78 375	15 145
Micro-moyenne F1	**72,14%**	69,18%	43,55%
Macro-moyenne F1	**62,46%**	55,86%	33,20%

les premières caractéristiques choisies pour la classe *automotive* =>

Bigrammes	Trigrammes
radars mobiles	fiche technique moteur
moteur électrique	boîte double embrayage
combi volkswagen	moteurs diesel truqués
volkswagen golf	quatre roues motrices
dacia logan	crédits photo daimler
automobiles ccfa	d'une voiture neuve
wild rubis	document cookie split
voiture tombe	l'ecologie ségolène royal
rémy josseaume	voiture low cost
mm poids	voitures tour auto
conduite écologique	mercedes classe e
voiture neuve	veyron super sport
double embrayage	bugatti veyron super
délégué interministériel	automobile club association
prévention routière	derniers jours combi
personnes tuées	format pave bas
station service	split for var
design extérieur	radars feux rouges
groupe allemand	renault clio iv
stéphanie thibault	d'un véhicule d'occasion

Résultats de l'algorithme SVM

- utiliser la fonction de noyau linéaire de SVM multi-classe *(one vs all)*
- c faible : plus de tolérance d'erreurs, marge importante, généralise facilement
- c fort : moins de tolérance d'erreurs, modèle plus spécifique, conduit au sur-apprentissage

	LIBSVM, sans application de méthode de sélection		Bayésien Naïf Multinomial, sans application de méthode de sélection	Bayésien Naïf Multinomial + filtrage par la **spécificité lexicale**
Exactitude moyenne	C = 0,001 **C = 0,01** C = 0,1 C = 1 C = 10	70,38% **73,93%** 73,25% 72,11% 71,23%	74,06%	73,56%

- bayésien naïf multinomial a obtenu des bons résultats
- après une réduction de 57% de caractéristiques, le résultat obtenu par la spécificité lexicale est satisfaisant

- Précision, rappel et f-mesure par classe (résultats de la spécificité lexicale)

1	Category name	Expected	TruePositive	FalsePositive	Precision	Recall	F-Mesure
2	xiko:topic.sports	32428	29374	1520	95.08 %	90.58 %	92.78 %
3	xiko:topic.home_garden	2726	2227	448	83.25 %	81.69 %	82.47 %
4	xiko:topic.cooking	2876	2481	761	76.53 %	86.27 %	81.10 %
5	xiko:topic.culture_entertainment	20513	15842	4083	79.51 %	77.23 %	78.35 %
6	xiko:topic.politics	7666	5891	1852	76.08 %	76.85 %	76.46 %
7	xiko:topic.technology	6747	5095	1681	75.19 %	75.52 %	75.35 %
8	xiko:topic.personal_finance	3096	2449	1073	69.53 %	79.10 %	74.01 %
9	xiko:topic.automotive	682	543	282	65.82 %	79.62 %	72.06 %
10	xiko:topic.weddings	880	530	84	86.32 %	60.23 %	70.95 %
11	xiko:topic.travel	3332	2548	1578	61.75 %	76.47 %	68.33 %
12	xiko:topic.real_estate	1698	1146	567	66.90 %	67.49 %	67.19 %
13	xiko:topic.pets	2020	1578	1154	57.76 %	78.12 %	66.41 %
14	xiko:topic.beauty	1787	1192	627	65.53 %	66.70 %	66.11 %
15	xiko:topic.health_well_being	9931	5571	1541	78.33 %	56.10 %	65.38 %
16	xiko:topic.science	5158	2902	980	74.76 %	56.26 %	64.20 %
17	xiko:topic.education	1260	977	873	52.81 %	77.54 %	62.83 %
18	xiko:topic.style_fashion	2833	1892	1367	58.05 %	66.78 %	62.11 %
19	xiko:topic.spirituality	574	302	131	69.75 %	52.61 %	59.98 %
20	xiko:topic.business	7814	4993	3949	55.84 %	63.90 %	59.60 %
21	xiko:topic.family_parenting	2233	1165	615	65.45 %	52.17 %	58.06 %
22	xiko:topic.environment	4049	2375	1773	57.26 %	58.66 %	57.95 %
23	xiko:topic.news	6456	4157	4717	46.84 %	64.39 %	54.23 %
24	xiko:topic.dating_love_couple	474	231	325	41.55 %	48.73 %	44.85 %
25	xiko:topic.legal_issue	1785	903	1802	33.38 %	50.59 %	40.22 %
26	xiko:topic.outing	691	243	521	31.81 %	35.17 %	33.40 %
27	xiko:topic.society	6716	2042	3698	35.57 %	30.41 %	32.79 %
28	xiko:topic.careers	396	55	115	32.35 %	13.89 %	19.43 %

1. Contexte et objectif d’étude
2. État de l’art
3. Méthodes de sélection comparées
4. Algorithmes d’apprentissage appliqués
5. Corpus et algorithme
6. Expériences et résultats
7. **Conclusion et perspectives**

Conclusion et perspectives

- évaluer quatre méthodes de sélection de caractéristiques par une validation croisée qui intègre la sélection sur chaque échantillon d'apprentissage
- le meilleur compromis entre le nombre de caractéristiques et les performances de la classification est obtenu par la spécificité lexicale
- différents prétraitements n'améliorent pas voire baissent les résultats

- ❖ implémenter d'autres méthodes de sélection
- ❖ essayer de choisir les caractéristiques globalement
- ❖ essayer d'autres algorithmes d'apprentissage
- ❖ approche de l'apprentissage non supervisé

Merci pour votre attention, des questions ?

Effet de racinisation et lemmatisation

Racinisation par SnowBall

- réduction de 19% ~ 31% de caractéristiques, résultats similaires en utilisant moins de temps

Lemmatisation par TreeTagger

- une baisse d'un ordre de grandeur de nombre de caractéristiques, résultats assez dégradés

Corpus lemmatisé	Corpus brut	
taillez	pâté	303
lavez	beurrés	91
sauce	cornets	disparue
saumon	mouillettes	disparue
tiédir	gla	disparue
réservez	whiskies	disparue
asperge	enlevez	disparue
tailler	grappes	disparue
cube	épongez	disparue
rôtir	beurrée	disparue
marinade	genviron	disparue
soigneusement	wine	disparue
amande	spéculoos	20
passez	perbellini	disparue
levure	gms	disparue
crabe	jayer	disparue
marron	xérès	disparue
tomate	cuisson	25
spéculoos	taboulé	479
gousse	passard	disparue

Les 20 premières caractéristiques choisies pour la classe *cooking*