Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques

Radwan Jalam et Jean-Hugues Chauchat

Laboratoire ERIC,
Université Lumière Lyon 2,
5, av. Pierre Mendès-France
69767 Bron, France
{rjalam, chauchat}@univ-lyon2.fr

Abstract

Why N-grams constitute an effective tool for the text categorization? How do we pass from the purely formal aspect of the text to its meaning? In order to answer these questions, we seek the specific N-grams to a text subset, then the words that contain those specific N-grams. New key words for this class of texts can be discovered. We present experimentations done on a set of 7.789 documents extracted from the Reuters collection corresponding to the 10 most frequent classes. In order to be close to someone who searches documents of a precise subject, we try to discriminate each class against all the other. Lists of candidates key words are each time obtained. The use of the N-grams appears to be more efficient because the specific stemmings feature in the N-grams; this method is automatic and needs no preliminary linguistic analysis.

Résumé

Pourquoi les n-grammes constituent-ils un outil efficace pour le classement de textes ? Comment passe-t-on de la forme au sens ? Pour répondre à ces questions, nous recherchons les n-grammes spécifiques à un sous-ensemble de textes, puis les mots qui contiennent ces n-grammes spécifiques. On peut ainsi découvrir de nouveaux mots-clés pour cette classe de textes. Nous présentons des expérimentations sur une collection de 7 789 dépêches extraites de la collection Reuters correspondant aux 10 classes parmi les plus fréquentes. Pour nous rapprocher de la situation de celui qui recherche des documents sur un sujet précis, nous cherchons à discriminer chaque classe contre toutes les autres. On obtient à chaque fois des listes de “candidats mots clés”. On compare les résultats avec la simple recherche de mots spécifiques à une classe. L'utilisation des n-grammes apparaît plus performante car les racines spécifiques apparaissent à travers les n-grammes ; elle est automatique, sans analyse linguistique préalable.

1. Introduction

De nombreux travaux ont montré l'efficacité des n-grammes comme méthode de représentation des textes pour leur classification : recherche d'une partition en groupes homogènes, ou pour leur catégorisation : attribution d'un texte à une, ou plusieurs, catégorie(s) parmi une liste pré-déterminée (Damashek, 1995; Dunning, 1994; Miller et al., 1999; Teytaud and Jalam, 2001; Cavnar and Trenkl, 1994).

Un premier objectif de ce travail est de montrer pourquoi la représentation en n-grammes est efficace. On rétablit le lien entre l'aspect purement formel du texte et son sens ; on passe des n-grammes, caractéristiques d'une classe de textes, aux mots contenant ces n-grammes dans ces textes.

Un deuxième aspect du travail est la recherche automatique d'une liste de mots statistiquement caractéristiques, c'est à dire de candidats “mots-clefs” dans laquelle l'utilisateur pourra choisir ceux qu'il retiendra. Notre travail se situe donc en amont de celui de (Lelu and Hallab, 2000) qui ont proposé une méthode interactive pour sélectionner des mots ou groupes de mots dans le même objectif.

La section 2 décrit les étapes de la recherche des mots statistiquement caractéristiques ; la section suivante décrit deux applications sur de grands corpus de données réelles ; la dernière section conclue et propose des pistes pour la poursuite de ce travail. Tout d'abord, nous rappelons le principe du codage en n-grammes, puis ses qualités.

1.1. Le codage en n-grammes

Un n-gramme est une séquence de n caractères consécutifs. Pour un document quelconque, l'ensemble des n-grammes (en général n prend les valeurs 2 ou 3) qu'on peut générer est le résultat qu'on obtient en déplaçant une fenêtre de n cases sur le corps de texte. Ce déplacement se fait par étapes, une étape correspond à un caractère. Ensuite on compte les fréquences des n-grammes trouvés. Par exemple la phrase "*La nourrice nourrit le nourrisson*" se représente par [la_=1, a_n=1, _no=3, nou=3, our=3, urr=3, rri=3, ric=1, ice=1, _ce=1, e_n=2, rit=1, it_=1, t_l=1, _le=1, le_=1, ris=1, iss=1, sso=1, son=1]. Dans le présent papier, nous représentons les n-grammes en utilisant le caractère "_" à la place des blancs, pour faciliter la lecture.

1.2. L'intérêt du codage en n-grammes

Les techniques basées sur les n-grammes présentent plusieurs avantages :

- ☞ comparativement à d'autres techniques, les n-grammes **capturent automatiquement les racines des mots les plus fréquents** (Grefenstette, 1995). On n'a pas besoin de l'étape de recherche des racines lexicales (nourrir, nourri, nourrit, nourrissez, nourrissant, ... , nourriture, ... , nourrice, ...).
- ☞ elles opèrent **indépendamment des langues** (Dunning, 1994), contrairement aux systèmes basés sur les mots dans lesquels il faut utiliser des dictionnaires spécifiques (féminin-masculin ; singulier-pluriel ; conjugaisons ; etc.) pour chaque langue. De plus, avec les n-grammes, on n'a pas besoin de segmentation préalable du texte en mots ; ceci est intéressant pour le traitement de langues dans lesquelles les frontières entre mots ne sont pas fortement marquées, comme le chinois, ou encore pour les séquences ADN en génétique.
- ☞ elles sont **tolérantes aux fautes d'orthographes** et **aux déformations** causées lors de l'utilisation des lecteurs optiques. Lorsqu'un document est scanné, la reconnaissance optique est souvent imparfaite. Par exemple, il est possible que le mot "*chapitre*" soit lu comme "*clapitre*". Un système basé sur les mots aura de mal à reconnaître qu'il s'agit du mot "*chapitre*" puisque le mot est mal orthographié. Par contre, un système basé sur les n-grammes est capable de prendre en compte les autres n-grammes comme "*apit*", "*pitr*", etc. (Miller et al., 1999) montre que des systèmes de recherches documentaires basés sur les n-grammes ont gardé leurs performances malgrès des taux de déformations de 30%, situation dans laquelle aucun système basé sur les mots ne peut fonctionner correctement.
- ☞ Enfin, ces techniques **n'ont pas besoin d'éliminer les mots-outils (Stop Words) ni de procéder à la lemmatisation (Stemming**). Ces traitements augmentent la performance des systèmes basés sur les mots. Par contre, pour les systèmes n-grammes, de nombreuses études (Sahami, 1999) ont montré que la performance ne s'améliore pas après l'élimination des "Stop Words" et de "Stemming".

2. Étapes de la recherche des mots caractéristiques

L'idée consiste à extraire les n-grammes caractérisant chaque classe puis à extraire les mots qui contiennent ces n-grammes. Nous avons développé un programme en Java qui recherche et compte les n-grammes des classes de textes, sélectionne les plus caractéristiques de chaque classe puis recherche dans ces textes les mots contenants les n-grammes caractéristiques et finalement élimine les mots parasites.

2.1. Recherche des n-grammes caractéristiques et des mots qui les contiennent

Avant de présenter l'algorithme complet, nous expliquons le principe de la démarche.

Les étapes principales sont les suivantes :

- ☞ recherche de tous les n-grammes de tous les textes de l'ensemble d'apprentissage ;
- ☞ constitution du tableau croisé (classe de textes $\times$n-grammes) ;
- ☞ calcul des contributions de chaque cellule de ce tableau au χ^2 d'indépendance ;
- ☞ pour chaque classe : recherche des n-grammes caractéristiques, c'est à dire ceux qui sont significativement plus fréquents dans les textes de cette classe que dans les autres) ;
- ☞ recherche des mots contenants ces n-grammes.

Pour caractériser une classe de texte, nous utilisons donc la statistique du χ^2. De nombreux autres indices sont disponibles, à partir de la matrice (N_{ij}) des occurrences des n-grammes i dans les classes de textes j (Yang, 1999; Aas and Eikvil, 1999) ; la statistique du χ^2 est souvent citée parmi les plus efficaces lors des comparaisons empiriques.

En pratique, la méthode proposée fournit une longue liste de mots parmi lesquels certains sont des "parasites", c'est à dire des mots contenants par hasard un des n-grammes caractéristiques de la classe, sans que le mot lui-même soit intéressant. L'objectif suivant est d'affiner la liste des *"candidats mots-clefs"*.

2.2. Filtrage des mots "parasites"

Pour éviter les mots "parasites" nous proposons de reprendre le traitement à l'inverse : pour chaque mot extrait précédemment, nous examinons l'ensemble des n-grammes qu'il contient et vérifions si ces n-grammes sont suffisamment nombreux à faire partie des n-grammes caractéristiques de la classe. Si :

1. la proportion des n-grammes de ce mot présents dans la liste des n-grammes caractéristiques dépasse un certain seuil (seuil1),
2. la fréquence de ce mot dans le texte dépasse également un certain seuil (seuil2),

alors le mot sera considéré comme mot clé candidat. Si un mot se répète plusieurs fois dans le texte, c'est qu'il est un candidat. Si un mot se répète rarement et qu'il a été sélectionné car il contient seulement un ou deux n-grammes en commun avec un autre mot qui se répète souvent c'est que ce mot est parasite.

Exemple : un 3-gramme comme *acq* dans la classe "*Acquisition*" va donner des mots caractérisant la classe, tels *"acquisition"* ou *"acquire*" ; mais il peut également être inclus dans des mots qui n'ont rien de caractéristiques, comme *"Jacques*" *ou "racquets"* qui seront considérés comme parasites car ils sont rares dans cette classe.

2.3. L'algorithme complet

Nous avons développé un programme en Java qui recherche les n-grammes, les compte, sélectionne les plus caractéristiques de chaque texte et recherche les mots correspondants comme indiqué dans l'algorithme 1.

Algorithm 1 La méthode proposée

1. Pour chaque classe j, rechercher tous les n-grammes dans tous les textes de l'ensemble d'apprentissage
2. constituer le tableau croisé (N_{ij}) des occurrences des n-grammes i dans la classe j,
3. calculer les fréquences f_{ij} correspondantes : $f_{ij} = \frac{N_{ij}}{N}$
4. calculer les contributions de (ij) à la statistique du χ^2 : $\chi^2_{ij} = \frac{\left(N_{ij} - \frac{N_{i.} \times N_{.j}}{N}\right)^2}{\frac{N_{i.} \times N_{.j}}{N}} = N \times \frac{(f_{ij} - f_{i.} \times f_{.j})^2}{f_{i.} \times f_{.j}}$
5. calculer le $\chi^2_{ij} \times signe(f_{ij} - f_{i.} \times f_{.j})$
6. trier le tableau des χ^2_{ij} dans l'ordre décroisant
7. <u>pour</u> chaque classe <u>j faire</u>
   (a) déterminer la liste $\{gram_{ij}\}$ des K premiers n-grammes de la classe
   (b) <u>pour</u> chaque $gram_{ij}$ <u>faire</u>
      i. chercher tous les mots (mot_{jk}) tels que $gram_{ij} \subseteq mot_{jk}$
      ii. calculer le nombre $nb_{mots_{jk}}$ des répétitions de mot_{jk} dans la classe
   (c) <u>Pour</u> chaque mot_{jk} <u>faire</u>
      i. extraire les grammes $gram_{mot_{jk}}$ de mot_{jk}, leur total est noté $nbGram_{mot_{jk}}$
      ii. <u>Pour</u> chaque gramme $gram_{mot_{jk}}$ <u>faire</u>
      <u>si</u> $gram_{mot_{jk}} \in \{gram_{ij}\}$ <u>alors</u> $presenceGram_{mot_{jk}} + +$
   (d) <u>si</u> $\frac{presenceGram_{mot_{jk}}}{nbGram_{mot_{jk}}} > seuil_1$ <u>et</u> $nb_{mots_{jk}} > seuil_2$ <u>alors</u> $mot_{jk} \in \{mots\, candidat\, de\, la\, classe_j\}$

3. Un exemple d'application

Notre méthode vise à aider un utilisateur à sélectionner des documents qui l'intéressent pour une tâche donnée, à partir d'un ensemble de documents non structurés, comme le web. Dans un premier temps, l'utilisateur doit constituer son ensemble d'apprentissage, c'est à dire fournir deux sous-ensembles de documents : un sous-ensemble de textes qui l'intéressent, et un sous-ensemble d'autres textes, de même(s) origine(s), qui ne l'intéressent pas pour cette tâche.

Version	Prép par	Nombre de classes	Ens. Apprent	Ens. Test	% Doc étiqutés
Vers 1	CGI	182	21,450	723	80%
Vers 2	Lewis	113	14,704	6,746	42%
Vers 2.2	Yang	113	7,789	3,309	100%
Vers 3	Apte	93	7,789	3,309	100%
Vers 4	PARC	93	9,610	3,662	100%

TAB. 1 – *Les différentes versions de la* collection Reuters

La classe	Acquisition	Earn	Money-fx	Wheat	Trade	Crude	Corn	Grain	Interest	Ship
Nb textes	1629	2841	528	209	362	383	173	427	346	194

TAB. 2 – *La répartition de l'ensemble des textes sur les 10 classes les plus représentées*

Comme ce travail est, par nature, spécifique à chaque utilisateur, nous présentons ici un exemple d'application que chacun peut plus facilement comprendre et contrôler, à savoir la sélection de dépêches d'agence sur tel ou tel sujet.

3.1. Les données indexées de Reuters

Pour cet exemple, nous utilisons un *benchmark* classique : les recueils de dépêches de l'agence de presse Reuters. Le tableau 1 présente les différentes versions de cette collection.

Nous utilisons ici un sous-ensemble de l'ensemble d'apprentissage de la version "Apte" qui contient 7 789 dépêches (Yang, 1999) : les 6 709 dépêches, correspondant aux 10 classes les plus représentées dans la collection d'apprentissage. Le tableau 2 montre la répartition de ces 6 709 dépêches sur les 10 classes.

3.2. Quelques résultats

Le tableau 3 montre le résultat que notre méthode propose pour les classes *Acquisition* et *Crude*. Dans cette expérimentation, la valeur de seuil 1 est égale à 2/3 et la valeur du seuil 2 est égale à 30. La méthode propose une liste d'une centaine de mots clés candidats pour chaque classe mais, faute de place, nous ne présentons que les premiers grammes significatifs avec les mots correspondants.

3.3. Discussion des résultats sur la collection Reuters

Pour chaque classe, la méthode nous propose une liste de candidats-mots-clefs, dont la plupart des mots parasites ont été éliminés, et ces mots sont spécifiques à chaque classe. On voit sur les tableaux que les mots proposés sont raisonnables. Mais, évidemment, ces résultats sont en partie liés aux événements de l'époque où les textes sont sélectionnés. La règle du "*toutes choses égales par ailleurs*" s'applique ici, comme ailleurs.

La méthode est complètement indépendante de la langue, dans la mesure où on peut ne retirer aucun séparateur : ni blanc, ni signe de ponctuation.

Les essais réalisés en utilisant - soit les 1+2+3-grammes, - soit les 4-grammes n'ont pas apporté d'amélioration ; les résultats sont quasi semblables. Sur ce point, nous retrouvons les conclusions de nombreux auteurs notamment dans (Cavnar and Trenkl, 1994; Lelu and Hallab, 2000; Fürnkranz, 1998).

Deux tableaux complètent cette présentation :

1. Le tableau 4 contient les listes complètes des n-grammes spécifiques des classes *corn* et *crude*, chacune contre les 9 autres classes de dépêches ;
2. Le tableau 5 présente les résultats comparés de notre méthode sur quatre codages possibles des textes :
   (a) calculs des χ^2_{ij} sur le tableau (mots i × classes j) avec un pré-traitement : élimination des ponctuations et espaces,
   (b) calculs des χ^2_{ij} sur le tableau (mots i × classes j) sans pré-traitement : on laisse les ponctuations et les espaces,
   (c) calculs des χ^2_{ij} sur le tableau (n-grammes i × classes j) avec un pré-traitement : élimination des ponctuations et espaces,

La classe Acquisition		La classe Crude	
Les grammes les plus significatif	**Les mots clés extraits**	**Les grammes les plus significatif**	**Les mots clés extraits**
acq cqu qui uis iti sit	acquisition	oil _oi il_ il,	oil oil,
acq cqu qui uir	acquire acquired acquiring acquiring	rud cru ude	crude
sha har are	share	bar arr rel els	barrels
sha har are reh hol lde eho old der	shareholder holders holding hold	cua uad dor ado	ecuador ecuadorean
com omp any pan	companies company;	bpd _bp pd_ pd,	bpd (barrel par jour)
tak ake eov keo	takeover stake take	gas	gas
sto toc ock lde	stockholders	ene erg rgy nergy_	energy
mer erg	merger ; merge	pet etr leu eum ole	petroleum
off ffe fer	offer offers offering offered	plo xpl lor ora	exploration
has pur has	purchase	sau aud udi di_	saudi
usa sai air	USAir	zue ezu nez uel	venezuela
buy	buy buys	bbl	bbl (barrel)
inv nve sto	investment investment investor	pip ipe pel	pipeline
scl	disclosed undisclosed	xxo exx	exxon
cyc ycl	cyclops	ref ner efi	refinery
sac	transaction		
com omp ple let	complete	ara rab iea	arabian arabia
fil	filing	cub bic ubi	cubic
oup	group	_ku kuw uwa wai	kuwait kuwaiti
tst	outstanding	ric ice ces	prices
twa	twa (Trans World Airline)	ope pec	opec (non-opec)

TAB. 3 – *Les premiers grammes significatifs avec les mots correspondants*

(d) calculs des χ^2_{ij} sur le tableau (n-grammes $i \times$ classes j) sans pré-traitement : on laisse les ponctuations et les espaces.

On voit que notre méthode, fondée sur les n-grammes et sans aucun pré-traitement donne d'excellents résultats. Elle est donc complètement indépendante de la langue des textes étudiés.

4. Conclusion et perspectives

Dans ce travail nous exposons une méthode qui aide à comprendre pourquoi les n-grammes donnent de bons résultats. Nous proposons pour cela un algorithme qui extrait des candidats-mots-clés spécifiques à un sous-ensemble de textes. Une application est réalisée sur 6 709 dépêches classées en 10 classes (les classes les plus représentées dans la collection Reuters). La méthode donne des résultats encourageants ; les mots qui ont en commun des grammes significatifs sont sélectionnés. Nous proposons une méthode pour réduire les mots parasites fondée sur la fréquence des mots et la proportion de n-grammes significatifs qu'ils contiennent. Cette méthode s'avère efficace, bien qu'elle travaille sur les fichiers de textes bruts, sans aucune analyse linguistique préalable.

En outre, les résultats de cette approche montrent que le passage des n-grammes sélectionnés vers les mots apporte non seulement les mots statistiquement significatifs (au sens du χ^2) mais également les flexions présentes dans le corpus (et supérieurs au seuil minimal prérequis des fréquences) de ces mots. Nous projetons ainsi d'utiliser cette représentation plus robuste, de part les propriétés des n-grammes, et plus riche, de part les flexions, pour l'aide à la création d'ontologie et également dans le cadre de la problématique de la catégorisation automatique.

Références

Aas K. and Eikvil L. (1999). Text categorization: a survey. Technical report, Norwegian Computing Center.

Cavnar W. and Trenkl J. (1994). N-Gram Based Text Categorization. In *Symposium on Document Analysis and Information Retrieval*, Las Vegas.

Damashek M. (1995). Gauging Similarity with N-Grams: Language-Independent Categorization of Text. *Science*, (267):843–848.

Dunning T. (1994). Statistical Identification of Languages. Technical Report MCCS 94-273, Computing Research Laboratory.

Fürnkranz J. (1998). A Study Using n-gram Features for Text Categorization. Technical Report OEFAI-TR-98-30, Austrian Research Institute for Artificial Intelligence, Austria.

Grefenstette G. (1995). Comparing Two Language Identification Schemes. In *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95)*, Rome, Italy.

Lelu A. and Hallab M. (2000). Consultation "floue" de grandes listes de formes lexicales simples et composées : un outil préparatoire pour l'analyse de grands corpus textuels. In Rajmann M. and Chappelier J. C. editors, *JADT'2000*, volume 1, pages 317–324, Lausanne.

Miller E., Shen D., Liu J., and C.Nicholas (1999). Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System. *Journal of Digital Information*, 1(5).

Sahami M. (1999). *Using Machine Learning to Improve Information Access*. PhD thesis, Computer Science Department, Stanford University.

Teytaud O. and Jalam R. (2001). Kernel-based text categorization. In *IJCNN'01*, Washington, DC, USA.

Yang Y. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1(1/2):67–88.

Les Grams	**Les mots extraits**
[orn] [onn] [rn_] [nne] [aiz] [acr] [ton] [0_t] [gra] [usd] [soy] [oyb] [ybe] [_–] [sda] [ze_] [nes] [arv] [–_] [ghu] [bea] [rai] [cor] [ize] [_us] [hum] [da_] [-_-] [sov] [far] [_so] [arm] [ltu] [ush] [rgh] [gri] [u.s] [.s.] [iet] [_to] [ssr] [ram] [mai] [uga] [por] [ean] [nro] [._c] [mm_] [_mm] [ogr] [vie] [rog] [icu] [huc] [s._] [gnu] [ort] [_u.] [enr] [rtm] [bue] [86/] [xpo] [wee] [_ep] [cul] [6/8] [/87] [cre] [ain] [cer] [nkn] [eat] [ovi] [nup] [aby] [rod] [odu] [pik] [rn,] [kab] [_gr] [rva] [whe] [ric] [ze,] [sug] [epa] [hea] [cro] [sr_] [duc] [liv] [dob] [rn.] [mpo] [rme] [eag] [llm] [fob] [unk] [she] [fre] [-55] [rts] [tot]	[corn] [corn.] [corn,] [tonnes,] [tonnes] [tonne] [tonnes.] [tonne,] [maize] [maize,] [acreage] [acres.] [acres] [washington,] [grain] [program] [program,] [grains] [program.] [usda] [usda's] [usda.] [soybeans] [soybean] [soybeans,] [harvest] [sorghum] [rains] [record] [ussr] [soviet] [soviets] [farmers] [farm] [sources] [agriculture] [agricultural] [bushel] [bushels] [u.s.] [u.s.-ussr] [to] [total] [sugar] [reported] [export] [report] [imports] [exports] [exporters] [import] [reports] [imports.] [enrollment] [u.s. corn] [huckaby] [u.s. agriculture] [department] [1986/87] [week] [between] [certificates] [producers] [unknown] [wheat] [wheat,] [products] [production] [growers] [conservation] [price] [prices] [crop] [french]
[oil] [_oi] [bpd] [rud] [_bp] [il_] [cru] [pd_] [bar] [rel] [cua] [arr] [etr] [uad] [gas] [...] [rgy] [eum] [leu] [xpl] [els] [sau] [dor] [pet] [ado] [zue] [ezu] [0_b] [nez] [uel] [ira] [aud] [ole] [di_] [bbl] [ec_] [pip] [lor] [cks] [l_p] [pd,] [tpu] [plo] [utp] [gy_] [..1] [odu] [cub] [rod] [ude] [kuw] [uwa] [pd.] [n_b] [i_a] [_cr] [pel] [iea] [rre] [bic] [_ir] [ice] [xxo] [exx] [raq] [bia] [ner] [udi] [/bb] [ara] [ipe] [rab] [ene] [pd)] [mex] [obr] [thq] [hqu] [s/b] [uak] [_op] [duc] [al-] [ref] [(bp] [e_o] [ubi] [fie] [ait] [pec] [tro] [fue] [uot] [_ie] [ora] [ls_] [dri] [quo] [fsh] [efi] [eia] [mob] [as_] [exa] [pdv] [vsa] [dvs] [wai] [_ku] [_ga] [f_o] [ric] [um_] [_bb] [ces] [ukm] [dez] [iel] [urk] [aqi] [try] [xic] [uct] [abi] [naz] [rol] [xac] [a_b] [erg] [eik] [_dr] [put] [prt] [qi_] [c_m] [kh_] [ian] [tex] [ikh] [aeg] [ia_] [ia'] [_pd] [aq_] [rs/] [wti] [_km] [noc] [ope] [ubr] [il,]	[oil.] [oil,] [oil] [oilfield] [bpd] [(bpd)] [bpd.] [bpd,] [crude] [crude,] [crudes] [crude.] [oil prices] [oil industry] [oil companies] [oil prices,] [oil prices.] [oil price] [oil and] [oil production] [bpd in] [barrel.] [barrel] [barrels] [barrels.] [barrel,] [barrels,] [reliance] [ecuador,] [ecuador] [ecuador's] [ecuadorean] [petroleum] [petrobras] [petroleos] [petroleum,] [gasoline] [gas] [energy] [exploration] [exploratory] [exploration.] [saudi] [saudis] [venezuela] [venezuela,] [venezuelan] [venezuela's] [fuel] [iraqi] [iraq] [iranian] [iran] [iran's] [saudi arabia] [dlrs/bbl.] [opec] [non-opec] [pipeline] [pipeline,] [stocks] [output] [products] [product] [production] [production.] [producing] [producer] [produce] [producers] [produced] [products.] [products,] [production,] [cubic] [kuwait] [kuwait,] [kuwaiti] [mln barrels] [mln bpd] [iea] [current] [prices] [prices.] [price] [prices,] [exxon] [arabia] [arabian] [arabia's] [arabia,] [refinery] [general] [refineries] [refiners] [including] [arab] [mexico] [earthquake,] [earthquake] [operating] [opec's] [operations] [open] [reduction] [refining] [crude oil] [the oil] [because of] [price of] [fields] [field] [expected] [quota] [quoted] [barrels per] [barrels of] [barrels a] [drill] [drilling] [offshore] [mobil] [was] [has been] [as] [texas] [as a] [texaco] [pdvsa] [of oil] [american] [sources] [industry] [ministry] [a barrel.] [a barrel] [sheikh] [drop] [canadian]

TAB. 4 – *Les listes complètes de n-grammes spécifiques et de "candidats-mots-clefs" pour les classes* Corn *puis* Crude.

La technique	Les mots extraits
extraction à partir de mots complets avec élimination des ponctuations et espaces	[oil] [bpd] [crude] [opec] [barrels] [barrel] [ecuador] [energy] [exploration] [petroleum] [prices] [gasoline] [gas] [refinery] [saudi] [saudis] [pipeline] [production]
extraction à partir de mots complets sans élimination de ponctuation et espaces	[oil.] [oil,] [oil] [crude] [opec] [opec's] [non-opec] [barrels] [barrels.] [barrels,] [bpd] [(bpd)] [bpd.] [bpd,] [energy] [petroleum] [ecuador,] [ecuador] [ecuador's] [exploration] [gasoline] [gas] [refinery] [saudi] [saudis] [prices] [prices.] [prices,] [barrel.] [barrel] [barrel,] [cubic] [production] [production,] [output] [stocks] [drilling] [pipeline] [pipeline,] [today] [day] [days] [yesterday] [iea] [arabia] [arabian] [natural] [venezuela] [venezuelan] [texaco] [petrobras] [api] [herrington] [mobil] [exxon] [offshore] [iranian] [feet] [15.8] [quota] [refining] [reserves] [kuwait] [wells] [fuel] [fields] [industry] [field] [iraqi] [minister] [spot] [demand] [price] [lukman] [santos] [producing] [iraq] [shell] [sources] [texas] [rigs] [research] [sea] [iran] [greece] [gulf]
extraction à partir de n-grams complets avec élimination de ponctuation et espaces	[oil] [bpd] [bp] [crude] [crudes] [oil prices] [oil industry] [oil stocks] [oil companies] [oil minister] [oil company] [oil price] [oil and] [oil production] [bpd in] [barrel] [barrels] [ecuador] [ecuadorean] [petroleum] [petrobras] [petroleos] [petro-canada] [gasoline] [gas] [energy] [exploration] [exploratory] [levels] [saudi] [saudis] [venezuela] [venezuelan] [fuel] [iraq] [iranian] [iran] [000 barrels] [000 bpd] [saudi arabia] [bbl] [pipeline] [stocks] [output] [products] [product] [production] [producing] [producer] [produce] [producers] [produced] [cubic] [kuwait] [kuwaiti] [iea] [mln barrels] [mln bpd] [current]
extraction à partir de n-grams complets sans élimination de ponctuation et espaces	[oil.] [oil,] [oil] [oilfield] [bpd] [(bpd)] [bpd.] [bpd,] [crude] [crude,] [crudes] [crude.] [oil prices] [oil industry] [oil companies] [oil prices,] [oil prices.] [oil price] [oil and] [oil production] [bpd in] [barrel.] [barrel] [barrels] [barrels.] [barrel,] [barrels,] [reliance] [ecuador,] [ecuador] [ecuador's] [ecuadorean] [petroleum] [petrobras] [petroleos] [petroleum,] [gasoline] [gas] [energy] [exploration] [exploratory] [exploration.] [saudi] [saudis] [venezuela] [venezuela,] [venezuelan] [venezuela's] [fuel] [iraqi] [iraq] [iranian] [iran] [iran's] [saudi arabia] [dlrs/bbl.] [opec] [non-opec] [pipeline] [pipeline,] [stocks] [output] [products] [product] [production] [production.] [producing] [producer] [produce] [producers] [produced] [products.] [products,] [production,] [cubic] [kuwait] [kuwait,] [kuwaiti] [mln barrels] [mln bpd] [iea] [current] [prices] [prices.] [price] [prices,] [exxon] [arabia] [arabian] [arabia's] [arabia,] [refinery] [general] [refineries] [refiners] [including] [arab] [mexico] [earthquake,] [earthquake] [operating] [opec's] [operations] [open] [reduction] [refining] [crude oil] [the oil] [because of] [price of] [fields] [field] [expected] [quota] [quoted] [barrels per] [barrels of] [barrels a] [drill] [drilling] [offshore] [mobil] [was] [has been] [as] [texas] [as a] [texaco] [pdvsa] [of oil] [american] [sources] [industry] [ministry] [a barrel.] [a barrel] [sheikh] [drop] [canadian]

TAB. 5 – *Comparaisons de quatre techniques sur la classe "Crude".*