Chapitre 5 Représentation des textes

Ce chapitre montre comment les textes sont transformés en vecteur de nombres pour être utilisés par les approches mettant en œuvre des apprentissages numériques. En général, les représentations n'utilisent pas d'information grammaticale ni d'analyse syntaxique des mots : seule la présence ou l'absence de certains mots est porteuse d'informations.

Nous présentons dans ce chapitre une méthode originale de sélection de descripteurs en deux étapes, qui présente plusieurs avantages. Elle est entièrement automatique et ne nécessite pas de ressources externes (comme une liste de mots les plus fréquents dans une langue donnée) et elle est couplée avec un critère d'arrêt pour trouver le "bon" nombre de descripteurs.

5.1 La représentation en sac de mots

La représentation des textes la plus simple a été introduite dans le cadre du modèle vectoriel présenté au chapitre 2, et porte le nom de "sac de mots". Les textes sont transformés simplement en vecteurs dont chaque composante représente un terme. Dans un premier temps, les termes sont les mots qui constituent un texte. Dans les langues comme le français ou l'anglais, les mots sont séparés par des espaces ou des signes de ponctuations ; ces derniers, tout comme les chiffres, sont supprimés de la représentation. On peut choisir de conserver les majuscules pour aider, par exemple, à la reconnaissance de noms propres, mais il faut alors résoudre le problème des débuts de phrase.

Les composantes du vecteur sont une fonction de l'occurrence des mots dans le texte.

À titre d'exemple, nous présentons, sur la Figure 5.1, une dépêche de l'Agence France Presse qui fournit des informations sur des prises de participations entre des entreprises. La transformation de ce texte en vecteur est présentée sous le texte. À partir de ces informations, un filtre doit détecter que cette dépêche est pertinente pour le thème des participations.

Cette représentation des textes exclut toute analyse grammaticale et toute notion de distance entre les mots : c'est pourquoi cette représentation est appelée "sac de mots".

```
Marionnaud: Union et Etudes Investissement franchit 5% des droits de vote
 PARIS, 31 juil (AFP) - La société Union et Etudes Investissement (caisse
Nationale de Crédit Agricole) a franchi en hausse le seuil de 5% des droits
de vote du groupement français de parfumerie Marionnaud et détient
désormais 292.157 actions, soit 8,09% du capital et 5,05% des droits de
vote, a indiqué vendredi le Conseil des Marchés Financiers.
Ce franchissement de seuil résulte de l'acquisition de 11.460 actions,
précise le CMF.
```

a	2	détient	1	le	3
acquisition	1	en	1	marchés	1
actions	2	et	4	marionnaud	2
agricole	1	études	2	nationale	1
caisse	1	financiers	1	parfumerie	1
capital	1	franchi	1	précise	1
ce	1	franchissement	1	résulte	1
cmf	1	franchit	1	seuil	2
conseil	1	français	1	société	1
crédit	1	groupement	1	soit	1
de	9	hausse	1	union	2
des	4	indiqué	1	vendredi	1
droits	3	investissement	2	vote	3
du	2	l	1		
désormais	1	la	1		

Figure 5.1 : *Exemple d'un texte et de son vecteur associé. Les composantes du vecteur sont simplement les occurrences des mots du texte.*

Représentation des textes avec des racines lexicales

Dans la description du modèle précédent, chaque flexion d'un mot est considérée comme un descripteur différent ; en particulier, les différentes formes d'un verbe sont autant de mots. Par exemple, dans le texte de la Figure 5.1, les mots *franchi* et *franchit* sont considérés comme des descripteurs différents alors qu'il s'agit de deux formes conjuguées du même verbe qui ont *a priori* le même sens.

Pour remédier à ce problème, il est possible de considérer uniquement la racine des mots plutôt que les mots entiers (on parle de *stem* en anglais). Plusieurs algorithmes ont été proposés pour substituer les mots par leur racine ; l'un des plus connus pour la langue anglaise est l'algorithme de Porter [Porter, 1980][1]. Nous n'avons pas utilisé un tel algorithme sur les corpus en langue française.

[1] Il est possible de trouver des implémentations rapides et efficaces de cet algorithme sur le web : http://www.muscat.com/~martin/stem.html

La Figure 5.2 est un exemple de texte du corpus Reuters dans sa version d'origine et la Figure 5.3 présente le même texte dont tous les mots ont été remplacés par leur racine grâce à l'algorithme de Porter.

```
TEXAS COMMERCE BANCSHARES &lt;TCB> FILES PLAN
Texas Commerce Bancshares Inc's Texas
Commerce Bank-Houston said it filed an application with the Comptroller of
the Currency in an effort to create the largest banking network in Harris
County.
    The bank said the network would link 31 banks having 13.5 billion dlrs
in assets and 7.5 billion dlrs in deposits.
```

Figure 5.2 : *Exemple de texte du corpus Reuters.*

```
TEXA COMMERC BANCSHAR &LT;TCB> FILE PLAN
Texa Commerc Bancshar Inc's Texa
Commerc Bank-Houston said it file an applic with the Comptrol of the
Currenc in an effort to creat the largest bank network in Harri Counti.
    The bank said the network would link 31 bank have 13.5 billion dlr in
asset and 7.5 billion dlr in deposit.
```

Figure 5.3 : *Texte précédent dont les mots ont été remplacés par leur racine.*

Il existe néanmoins d'autres algorithmes que celui de Porter pour déterminer les racines lexicales ; une comparaison entre différents algorithmes a été menée dans [Hull, 1996].

Représentation des textes avec des lemmes

La lemmatisation consiste à utiliser l'analyse grammaticale afin de remplacer les verbes par leur forme infinitive et les noms par leur forme au singulier. La lemmatisation est donc plus compliquée à mettre en œuvre que la recherche de racines, puisqu'elle nécessite une analyse grammaticale des textes. Un algorithme efficace, nommé *TreeTagger*[1] [Schmidt, 1994], a été développé pour les langues anglaise, française, allemande et italienne. Cet algorithme utilise des arbres de décision pour effectuer l'analyse grammaticale, puis des fichiers de paramètres spécifiques à chaque langue.

La Figure 5.4 et la Figure 5.5 montrent deux exemples, l'un en français et l'autre en anglais, de textes dont les mots d'origine ont été remplacés par leur lemme (l'algorithme remplace les nombres par le signe @card@).

[1] Les publications relatives à cet algorithme ainsi que les codes source sont disponibles sur le site : http://www.ims.uni-stuttgart.de/projekte/corplex/DecisionTreeTagger.html

```
TEXAS COMMERCE BANCSHARES LT TCB> FILE PLAN
Texas Commerce Bancshares inc's Texas
Commerce Bank Houston say it file an application with the
comptroller of the currency in an effort to create the large
banking network in Harris County
    The bank say the network would link bank have
@card@ billion dlrs in asset and @card@ billion dlrs in deposit
```

Figure 5.4 : *Texte de la Figure 5.2 dont les mots ont été remplacés par leur lemme.*

```
Marionnaud: Union et Etudes Investissement franchir @card@ de+le droit de
vote
le société union et Etudes investissement (caisse National de Crédit
Agricole avoir franchir en hausse le seuil de @card@ de+le droit de vote
de+le groupement français de parfumerie Marionnaud et détenir désormais
@card@ action être @card@ de+le capital et @card@ de+le droit de vote avoir
indiquer vendredi le conseil de+le marché financier.
Ce franchissement de seuil résulter de l'acquisition de @card@ action,
préciser le CMF.
```

Figure 5.5 : *Texte de la Figure 5.1 dont les mots ont été remplacés par leur lemme.*

La substitution des mots par leur racine ou leur lemme réduit l'espace des descripteurs et permet de représenter par un même descripteur des mots qui ont le même sens. Par exemple, le remplacement des mots *bank*, *banks*, *banking* dans le texte de la Figure 5.3 par l'unique racine *bank* semble être avantageux tout comme le remplacement des formes conjuguées *franchit* et *franchi* par le lemme *franchir* dans le texte de la Figure 5.5.

Néanmoins ces substitutions peuvent augmenter l'ambiguïté des descripteurs en représentant par un même descripteur des mots avec des sens différents (des exemples seront donnés au chapitre 7). Même le simple remplacement de la forme plurielle d'un mot par sa forme singulier peut augmenter l'ambiguïté d'un mot comme dans la Figure 5.5 où *actions* est représenté par le descripteur *action*. Dans un contexte économique, en effet, le mot *actions* se réfère le plus souvent à des actions des entreprises et n'a rien à voir avec le concept *action* employé par exemple dans la phrase : "le domaine d'*action* du gouvernement".

Il n'est pas possible de savoir *a priori* quelle représentation conduira aux meilleures performances et des expériences sont effectuées aux chapitres 7 et 8 afin de déterminer la représentation la mieux adaptée à la mise en œuvre des apprentissages numériques.

Dans toute la suite de ce chapitre, chaque mot est considéré comme un descripteur différent.

5.2 Etude de la fréquence des mots sur un corpus : la loi de Zipf

5.2.1 Enoncé de la loi de Zipf

La distribution de l'occurrence des mots dans un corpus de texte donné n'est pas uniforme : certains mots apparaissent très fréquemment, tandis que d'autres apparaissent très rarement. Les mots les plus fréquents en français sont les mots grammaticaux comme *le, la, les, et, … .* Sur le corpus Reuters, les cinq mots qui apparaissent le plus fréquemment sont : *the, of, to, in, and.*

La distribution de fréquence des mots dans un corpus a été étudiée empiriquement par Zipf [Zipf, 1949] et les résultats de cette analyse sont connus sous le nom de loi de Zipf. Pour énoncer cette loi, Zipf est parti d'un principe général qu'il a ensuite énoncé mathématiquement. Si l'on considère un corpus contenant T textes et que l'on note $TF(m, t)$ l'occurrence d'un mot m dans un texte t, on peut définir $CF(m)$, l'occurrence totale du mot m sur le corpus T :

$$CF(m) = \sum_{t \in T} TF(m,t)$$

Si l'on classe ensuite l'ensemble des mots du corpus par ordre décroissant d'occurrence totale, on obtient pour chaque mot un rang $r(m)$. La loi formulée par Zipf s'écrit alors :

$$CF(m) \cdot r(m) = K_T$$

K_T est une constante qui dépend du corpus. Cette relation peut s'écrire également :

$$Log(r(m)) = Log(K_T) - Log(CF(m))$$

Cette dernière relation, montre que si l'on trace le logarithme de l'occurrence en fonction du logarithme du rang, on doit obtenir une droite de pente –1.

La Figure 5.6 montre la vérification expérimentale de la loi de Zipf sur le corpus Reuters. On compte 51427 termes différents sur ce corpus ; à partir du rang 6243, les termes ont une fréquence totale sur l'ensemble du corpus inférieure à trente.

En fait, il est fréquent, comme le montre la Figure 5.6, que la loi ne soit pas très bien vérifiée pour les hautes fréquences et les basses fréquences, et il existe différentes méthodes pour corriger cette loi dans les domaines où elle n'est plus tout à fait vérifiée [Manning et Schütze, 1999]. Cependant, cette loi reflète bien le comportement général de la distribution des

occurrences : il existe un petit nombre de mots très fréquents, il existe un grand nombre de mots très rares n'apparaissant qu'une fois ou deux sur le corpus et il existe tout un ensemble de mots dont la fréquence d'apparition se situe entre ces deux domaines.

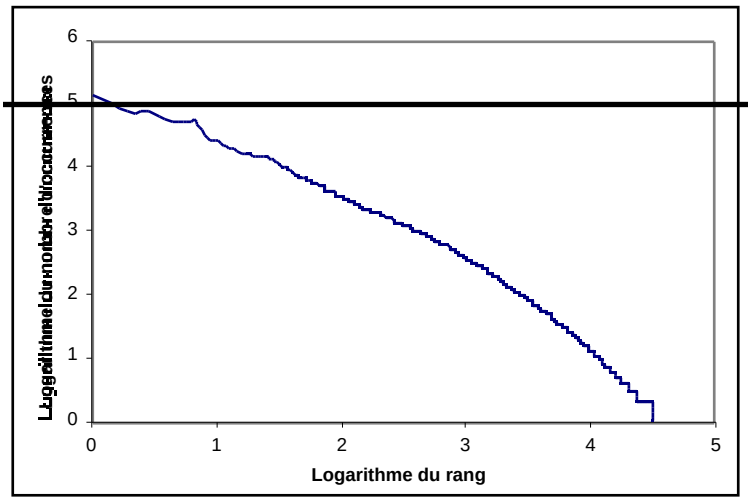

Figure 5.6 : *Vérification expérimentale de la loi de Zipf sur le corpus Reuters.*

5.2.2 Conséquences de la loi de Zipf

Comme le nombre de mots présents dans un corpus peut être très grand, les méthodes statistiques cherchent, en général, à réduire le nombre de mots utilisés pour représenter les textes. Nous verrons dans la suite comment s'effectue cette opération, mais les observations du paragraphe précédent permettent d'effectuer une première réduction de la dimension de l'espace de descripteurs.

Il s'agit de supprimer les mots dont on sait *a priori* qu'ils ne seront pas utiles pour les algorithmes d'apprentissage. Cette étape est critique, car les mots supprimés lors de cette étape le sont définitivement et il ne faut donc pas supprimer de mots importants.

La distribution de fréquences décrite par la loi précédente a deux conséquences importantes pour la représentation des textes.

5.2.2.1 Suppression des mots fréquents

Les mots qui apparaissent le plus souvent dans un corpus sont, comme on l'a vu précédemment, les mots grammaticaux ou les mots de liaisons. Ces mots doivent être supprimés de la représentation des textes pour deux raisons :

- d'un point de vue linguistique, ces mots ne comportent que très peu d'informations. La présence ou l'absence de ces mots n'aident pas à deviner le sens d'un texte. Pour cette raison, ils sont communément appelés mots vides (ou *stop words* en anglais).
- d'un point de vue statistique, ces mots se retrouvent sur l'ensemble des textes sans aucune discrimination et ne sont d'aucune aide pour la classification.

Comme le nombre de mots concernés est faible, il est possible de définir une liste de mots qui sont automatiquement supprimés de la représentation.

Par exemple [Sahami, 1998] définit une liste de 570 mots courant en anglais, plus une liste de 100 mots très fréquents sur le web, pour supprimer les mots les plus courants.

Cependant, l'établissement d'une telle liste peut poser des problèmes. D'une part, il n'est pas facile de déterminer le nombre de mots exacts qu'il faut inclure dans cette liste. D'autre part, cette liste est intimement liée à la langue utilisée et n'est donc pas transposable directement à une autre langue.

5.2.2.2 Suppression des mots rares

En général, les auteurs cherchent également à supprimer les mots rares d'un corpus afin de réduire de façon appréciable la dimension des vecteurs utilisés pour représenter les textes, puisque, d'après la loi de Zipf, ces mots rares sont très nombreux. D'un point de vue linguistique, la suppression de ces mots n'est pas nécessairement justifiée : certains mots peuvent être très rares, mais très informatifs. Néanmoins, ces mots ne peuvent pas être utilisés par des méthodes à bases d'apprentissage du fait de leur très faible occurrence.

Une des méthodes communément retenues pour supprimer ces mots consiste à ne considérer que les mots l'occurrence totale est supérieure à un seuil fixé préalablement.

Nous exposons dans le paragraphe suivant une méthode de détermination du vocabulaire spécifique d'un thème permettant de d'écarter automatiquement les mots rares et les mots fréquents sans utiliser de liste de mots prédéfinis. Cette méthode présente en outre l'avantage d'être transposable automatiquement du français à l'anglais, et d'être adaptée à la classification.

5.3 Une méthode automatique de détermination du vocabulaire spécifique

Cette méthode a été développée dans le cadre de l'application Exoweb présentée au chapitre 9.

Lorsque que l'on dispose d'un ensemble de textes pertinents pour un thème donné, on cherche à trouver le vocabulaire spécifique de ce thème, constitués par les termes que partagent ces textes parmi les mots ni trop fréquents, ni trop rares. Cette méthode se décompose en deux étapes : la première étape élimine de chaque texte les mots fréquents, la deuxième étape élimine les mots rares.

5.3.1 Elimination des mots fréquents

Si l'on note, comme précédemment, *TF(m, t)* l'occurrence d'un mot *m* dans un texte *t,* et *CF(m)* l'occurrence de ce mot sur le corpus, on calcule pour chaque mot d'un texte *t* le rapport :

$$R\,(m,t) = \frac{TF\,(m,t)}{CF\,(m)}$$

Les mots du texte sont classés par ordre décroissant de ce rapport. Plus le mot *m* est fréquent dans le corpus, plus le ratio est faible et, inversement, plus un mot est rare, plus le ratio est élevé. Dans le cas limite où un mot n'apparaît qu'une seule fois dans le corpus, ce ratio vaut 1 et le mot est classé en première place.

On supprime la deuxième moitié de la liste des mots de la représentation du texte, si bien que cette représentation ne contient plus les mots fréquents du corpus.

A la fin de cette étape, chaque texte *t* est représenté par un vecteur $\overrightarrow{v(t)}$ avec un codage booléen (1 si le mot est présent et 0 s'il est absent).

La Figure 5.7 montre l'application de cette méthode au texte de la Figure 5.1. On retrouve bien dans les premiers mots sélectionnés (le haut de la colonne de gauche) les mots les plus rares avec en tête le nom propre *marionnaud*. La fin de la liste (le bas de la colonne de droite) montre que les mots grammaticaux sont automatiquement supprimés. Les mots de la colonne de gauche sont conservés, et ceux de la colonne de droite sont supprimés.

marionnaud	désormais
études	union
parfumerie	actions
investissement	précise
franchissement	conseil
financiers	capital
franchit	ce
résulte	français
marchés	soit
groupement	société
franchi	vendredi
caisse	hausse
nationale	indiqué

agricole	la
seuil	et
cmf	des
vote	le
droits	du
détient	de
acquisition	a
crédit	en

Figure 5.7 : *Représentation du texte : la colonne de gauche montre les mots conservés pour la représentation et la colonne de droite montre les mots supprimés.*

5.3.2 Elimination des mots rares

Comme on l'a souligné précédemment, il est également nécessaire de supprimer de la représentation les mots qui apparaissent peu sur le corpus. Pour cela, on calcule la somme des vecteurs sur l'ensemble des textes dont on veut trouver le vocabulaire spécifique. On considère un sous-ensemble *S* du corpus *T* tel que tous les textes de *S* soient relatifs à un même thème. On définit le vocabulaire spécifique d'un thème en considérant la somme vectorielle suivante :

$$\vec{V}_s = \sum_{t \in S} \overrightarrow{v(t)}$$

Si l'on classe maintenant les composantes du vecteur obtenu par ordre décroissant, on obtient une liste de mots classés de telle sorte que les mots rares sont en fin de liste et les mots fréquents en haut de la liste.

5.3.3 Conclusion sur la méthode et exemples de mise en œuvre

La méthode proposée ci-dessus supprime automatiquement les mots rares et les mots fréquents d'un corpus sans utiliser une liste externe de mots liés à une langue particulière.
Cette méthode détermine le vocabulaire commun d'un sous-ensemble de textes d'un corpus ; les mots choisis ont la particularité de figurer dans le milieu de la distribution de la loi de Zipf.

La Figure 5.8 présente les dix premiers mots de quatre listes de vocabulaire spécifique obtenus sur des corpus différents ; ces listes ont été construites à partir de quatre ensemble de textes pertinents, chaque ensemble définissant un thème.
Les deux premiers ensembles ont été obtenus grâce à l'application ExoWeb présentée au chapitre 9 et sont composés de dépêches de l'AFP. Ces ensembles correspondent aux thèmes *participations* qui traite des échanges de participations entre entreprises (1400 dépêches pertinentes) et au thème *inforoute* qui traitent des informations sur les nouvelles technologies

et notamment l'internet (1000 dépêches pertinentes). Le troisième thème, *coffee*, est issu du corpus Reuters (111 documents pertinents). Le dernier thème est le thème 351 du corpus TREC-8 (31 documents pertinents) et traite de la recherche de pétrole au large des îles malouines (*Falkland petroleum exploration*).

Participation		*Inforoute*		*Coffee*		*351*	
422	détient	603	internet	98	coffee	28	islands
335	capital	229	accès	53	quotas	28	argentina
328	participation	202	site	49	ico	24	argentine
295	actionnaire	181	ligne	40	producers	21	aires
226	actionnaires	175	électronique	39	organization	21	buenos
223	actions	170	web	32	quota	19	sovereignty
192	cession	169	sites	32	producer	17	malvinas
181	parts	159	www	31	bags	17	exploration
177	vote	141	utilisateurs	30	export	16	oil
176	acquisition	135	com	27	colombia	14	islanders

Figure 5.8 : *Exemples des dix premiers mots trouvés sur trois thèmes.*

A ce stade, les résultats ne sont pas évalués quantitativement, mais dans tous les cas, les listes de mots semblent caractéristiques de la description du thème. Il est intéressant de noter que ces exemples correspondent à trois corpus différents (AFP, Reuters et Financial Times), avec des langues différentes et un nombre variable de documents pertinents, et que dans tous ces cas, la méthode semble fiable.

Une évaluation quantitative est effectuée au paragraphe 5.5.3 sur d'autres exemples.

La Figure 5.9 reprend la courbe de la de Zipf présentée précédemment et fait apparaître avec des cercles les dix premiers mots trouvés pour la catégorie *coffee* par la méthode précédente. Les mots trouvés se situent bien dans le milieu de la distribution.

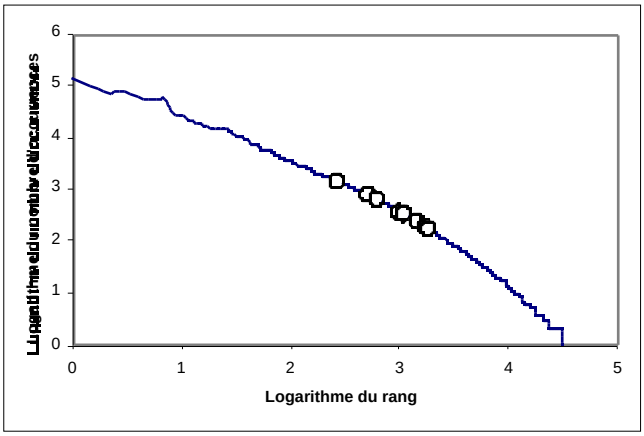

Figure 5.9 : *Répartition des mots spécifiques de la catégorie coffee dans la loi de Zipf.*

5.4 Codage des termes

Une fois les composantes des vecteurs choisies pour représenter un texte, il faut décider comment coder chaque coordonnée du vecteur. Si *TF(m, t)* est l'occurrence d'un terme *m* dans un texte *t*, la composante d'un vecteur est codée *f(TF(m, t))*, où la fonction *f* doit être déterminée.

Dans la représentation de la Figure 5.1, la fonction *f* a été choisie égale à l'identité ; ce qui représente le choix le plus simple.

Il est également possible de choisir la fonction booléenne définie par :

$$f = \begin{cases} 1 \text{ si } TF(m,t) > 0 \\ 0 \text{ si } TF(m,t) = 0 \end{cases}$$

C'est cette fonction qui a été utilisée dans le paragraphe 5.3.2 pour éliminer les mots rares lors de la détermination du vocabulaire spécifique.

Néanmoins, cette fonction est rarement utilisée pour les méthodes statistiques, car ce codage supprime de l'information qui peut être utile : l'apparition du même mot plusieurs fois dans un texte peut constituer un élément de décision important.

5.4.1 Le codage *tf.idf*

Le codage "tf.idf" a été introduit dans le cadre du modèle vectoriel présenté au chapitre 2 et donne parfois son nom à la méthode vectoriel. Le codage utilise une fonction de l'occurrence multipliée par une fonction qui fait intervenir l'inverse du nombre de documents différents dans lequel un terme apparaît. Ce sigle provient de l'anglais et signifie : *term frequency * inverse document frequency*.

Il est admis que l'occurrence d'un terme est une information importante, mais cependant, on considère généralement que la fonction *f* ne doit pas être l'identité. Si, par exemple, un mot apparaît deux fois dans un texte, son importance n'est pas nécessairement deux fois plus grande que s'il n'apparaissait qu'une seule fois. Pour cette raison, si l'occurrence n'est pas nulle, la fonction souvent utilisée est de la forme :

$$1 + Log\left(TF(m,t)\right)$$

En fait, les mots-clefs ont tendance à apparaître plusieurs fois dans un document pertinent : ainsi, deux mots peuvent avoir la même d'occurrence sur le corpus, mais l'un des deux peut apparaître dans beaucoup plus de documents que l'autre. Pour compenser cet effet, le modèle vectoriel utilise souvent le codage appelé *tf.idf* qui consiste à choisir une fonction telle que :

$$f = \begin{cases} (1 + Log(TF(m,t))).Log\dfrac{N}{DF(m)} & si\ TF(m,t) > 0 \\ 0 & si\ TF(m,t) = 0 \end{cases}$$

DF(m) est le nombre de documents différents où le terme *m* apparaît au moins une fois et *N* est le nombre de documents contenus dans le corpus. Ainsi, si un terme apparaît dans tous les documents, son poids a une valeur nulle puisqu'il n'apporte aucune information.

Des variations de ce codage ont été proposées dans [Salton et Buckley, 1990].

5.4.2 Un codage efficace : le codage Lnu

Les différents textes qui composent un corpus ont des tailles différentes dont il faut tenir compte dans le codage des termes.

Selon Singhal [Singhal, 1996], il existe deux phénomènes à considérer dans les textes longs par rapport aux textes courts :

- les mots présents tendent à avoir des fréquences plus élevées,
- les textes longs sont plus susceptibles de contenir des mots-clefs différents.

Il propose le codage Lnu, défini à la Figure 5.10, pour tenir compte de ces deux phénomènes ; ce codage tient également compte des remarques faites lors paragraphe précédent et est inspiré du codage tf.idf.

$$\text{Lnu} = L * u$$

$$L = \frac{1 + Log\left(TF(m,t)\right)}{1 + Log\left(\overline{TF}(m)\right)}$$

$$u = \frac{1}{0.8 + 0.2\,\dfrac{U(t)}{\bar{U}}}$$

Figure 5.10 : *Définition des termes du codage Lnu.*

$\overline{TF}(m)$: Fréquence moyenne dans le texte *t.*

$U(t)$: Nombre de termes uniques dans le texte *t.*

$\bar{U}$: Nombre moyen de termes sur l'ensemble des textes du corpus.

Ce codage est utilisé avec succès dans [Singhal, 1998] et [Ng *et al*., 2000].

5.5 Sélection de descripteurs pour la catégorisation de textes

5.5.1 Nécessité de la sélection de descripteurs

Les considérations précédentes ont permis de définir des critères pour représenter les textes en vecteur, mais il faut ensuite choisir plus spécifiquement les descripteurs qui vont être utilisés comme vecteurs d'entrées des modèles obtenus par apprentissage.

Comme dans tout le reste de cette étude, le filtrage est considéré comme un problème de classification à deux classes ; chaque thème nécessite une représentation différente des textes et un choix différent de descripteurs. Comme pour d'autres problèmes de classification, certains descripteurs peuvent être non discriminants pour la tâche que l'on cherche à résoudre. Par exemple, pour un problème de ciblage de clientèle pour un magasin, la couleur des yeux des clients a peu de chance d'être une caractéristique très pertinente pour construire un modèle.

Les entrées non discriminantes doivent être supprimées pour deux raisons différentes :

- Pour les modèles tels que les réseaux de neurones, le nombre de poids du réseau croît linéairement avec le nombre de descripteurs utilisés en entrée du modèle. Donc plus le nombre d'entrées est grand, plus le nombre de paramètres à déterminer est élevé : il faut alors disposer d'une base d'exemples plus grande afin d'avoir une bonne estimation de ces paramètres, même si les méthodes de régularisation présentées au chapitre 6 permettent de pallier partiellement ce problème[1].

- Comme les bases d'apprentissage ne sont pas infinies, (on verra qu'elles sont souvent de petite taille), des corrélations fortuites peuvent apparaître entre un descripteur non

[1] Il faut noter néanmoins que, pour les modèles linéaires par rapport à leurs paramètre tels que les polynômes, la nécessité de sélectionner les descripteurs est encore plus grande puisque le nombre de paramètres varie exponentiellement avec le nombre de descripteurs.

informatif et les individus d'une classe ; elles peuvent avoir une influence négative sur la qualité du modèle.

Les méthodes de sélection de descripteurs ont donc pour but de choisir parmi un ensemble de descripteurs possibles, les "bons" descripteurs, c'est-à-dire ceux qui vont permettre d'obtenir de bonnes performances sur une base différente de la base d'apprentissage.

La problématique de la sélection de descripteurs est très générale, mais le traitement du langage naturel présente des spécificités que nous allons présenter ci-dessous.

5.5.2 Spécificité de la sélection de descripteurs pour le filtrage de documents

Pour la problématique du filtrage de documents avec le modèle vectoriel, l'ensemble des descripteurs potentiels est constitué de l'ensemble des mots du corpus, ce qui peut représenter plusieurs centaines de milliers d'individus sur un corpus de taille raisonnable. Même si l'on a montré que les mots les plus fréquents et les plus rares pouvaient être éliminés facilement, soit parce qu'ils n'étaient pas discriminants, soit parce qu'ils n'étaient pas exploitables statistiquement, le nombre de candidats reste de plusieurs milliers. Parmi l'ensemble des descripteurs restants, on peut penser que tous ne sont pas nécessairement discriminants pour un thème donné, et que certains peuvent être très corrélés.

On cherche donc à supprimer ces mots de la représentation des textes, tout en sachant que chaque suppression de mot entraîne une perte d'information ; il faut trouver le bon compromis entre, d'une part, la nécessité de réduire l'espace des descripteurs et, d'autre part, le besoin de garder suffisamment d'information.

L'une des difficultés du filtrage provient du fait qu'il existe beaucoup de tournures différentes pour exprimer le même concept ou la même idée. La Figure 5.11 montre des exemples de passages pertinents pour le thème *participation*, qui traite des échanges de participations entre entreprises. Cet exemple montre la variété du vocabulaire employé, et la nécessité de conserver suffisamment de descripteurs, sachant que la présence d'une seule de ces phrases suffit à rendre le texte pertinent pour ce thème.

```
Telia est sur le point d'être privatisé, un tiers de son capital devant
être introduit en Bourse

BSCH a porté sa participation de 4,89% à 5,1% dans la banque allemande
Commerzbank
```

KPN a annoncé mercredi qu'il allait **céder les 21% qu'il détient** dans son homologue irlandais Eircom

Une compagnie nationale dont la **majorité de contrôle reste détenue** par l'état.

Tennessee **détenait 12,64% des actions** et **11,24% des droits de vote**.

L'entrée du Crédit Agricole **au capital** du Crédit Lyonnais comme **premier actionnaire** (avec 10%).

Figure 5.11 : *Exemples de phrases pertinentes pour le thème participation.*

La sélection de descripteurs est une étape primordiale de la construction d'un filtre, car, quel que soit le modèle statistique utilisé ultérieurement, si la représentation des textes n'inclut pas certains descripteurs ou si les descripteurs retenus sont trop nombreux ou mal choisis, le filtre aura des performances médiocres.

5.5.3 Les méthodes de sélection de descripteurs

5.5.3.1 Principe des méthodes

La méthode idéale consisterait à tester tous les sous-ensembles possibles de descripteurs afin de conserver l'ensemble donnant les meilleurs résultats sur une base de test. Une telle solution n'est évidemment pas possible, car si l'on considère un ensemble de p candidats, le nombre d'ensembles à tester s'élève à 2^p, donc si l'ensemble initial comporte cent descripteurs, le nombre de combinaisons s'élève à environ 10^{30} ce qui est évidemment beaucoup trop grand pour être réalisable.

Parmi les méthodes testées ici, deux approches différentes sont utilisées ; toutes deux tiennent compte de la tâche que l'on cherche à accomplir : différencier les textes pertinents des textes non pertinents.

La première approche consiste à calculer un score pour chaque descripteur, indépendamment des autres, en s'appuyant sur les statistiques d'apparition et d'absence du descripteur en fonction de la classe à laquelle appartiennent les textes. Les descripteurs sont ensuite classés selon ce score, les descripteurs en tête de liste étant les plus discriminants pour distinguer les textes pertinents des textes non pertinents. Les méthodes de l'information mutuelle et du chi-2 exposées ci-après reposent toutes deux sur ce principe.

La deuxième approche est constructive : elle construit itérativement un modèle, en partant d'un ensemble vide et en ajoutant successivement de nouveaux descripteurs en tenant compte des

descripteurs déjà sélectionnés. Cette construction est faite en utilisant l'algorithme d'orthogonalisation de Gram-Schmidt.

Pour toutes ces méthodes le résultat se présente sous la même forme : il s'agit d'une liste de mots ordonnés du plus discriminant au moins discriminant.

5.5.3.2 La méthode du chi-2

La statistique du χ^2 (chi-2) mesure l'indépendance entre un descripteur *t* et un thème *T*. Cette mesure a été utilisée pour la sélection des descripteurs dans [Schütze *et al.*, 1995] et [Wiener *et al.*, 1995]. Le calcul nécessite de construire pour chaque descripteur *t* du corpus le tableau de contingences de la Figure 5.12 :

	Descripteur *t* présent	Descripteur *t* absent	
Thème *T* présent	*a*	*c*	$T_1 = a+c$
Thème *T* absent	*b*	*d*	$T_0 = b+d$
	$D_1 = a+b$	$D_0 = c+d$	$N = a+b+c+d$

Figure 5.12 : *Tableau de contingences pour l'absence ou la présence d'un descripteur.*

On définit :

$$\chi^2(t,T) = \frac{N(ad - cb)^2}{(a + c)(b + d)(a + b)(c + d)}$$

Si un descripteur *t* et le thème *T* sont totalement indépendants, alors *t* apparaît avec la même fréquence dans le sous-ensemble des textes pertinents et dans le sous-ensemble des textes non pertinents, ce qui se traduit par (ad = bc) et la valeur de $\chi^2(t,T)$ est nulle.

A l'inverse, si le descripteur *t* apparaît systématiquement dans l'ensemble des textes pertinents et jamais dans l'ensemble des textes non pertinents, on a c = b = 0 et $\chi^2(t,T)$ vaut N, ce qui est sa valeur maximale. Cette valeur est également atteinte si un descripteur apparaît systématiquement dans l'ensemble des textes non pertinents et jamais dans l'ensemble des textes pertinents.

Entre ces deux valeurs extrêmes, plus la valeur de $\chi^2(t,T)$ est grande, plus t et T sont liés. Les descripteurs du corpus sont donc classés par ordre décroissant de $\chi^2(t,T)$, les plus discriminants figurant en tête de liste.

5.5.3.3 La méthode de l'information mutuelle

L'information mutuelle est employée pour mesurer la quantité d'information apportée par la présence ou l'absence d'un descripteur dans un document. Cette mesure a été fréquemment utilisée pour la catégorisation de textes pour effectuer la sélection de descripteurs [Lewis, 1992] [Mouliner, 1997] [Dumais *et al.*, 1998].

Dans le cas d'une classification à deux classes notées C_0 et C_1, si l'on note $P(t=1)$ la probabilité de présence d'un descripteur t dans un document et $P(t=0)$ son événement complémentaire, l'information apportée par la présence ou l'absence de t se mesure par :

1

La première somme fait intervenir les probabilités *a priori* de chaque classe ; elle est indépendante des descripteurs et n'est donc pas prise en considération. En reprenant les notations du tableau de la Figure 5.12, on calcule pour chaque descripteur t la quantité (sans tenir compte du terme constant) :

$$G'(t) = \frac{D_0}{N}\left(\frac{d}{D_0}\log\left(\frac{d}{D_0}\right) + \frac{c}{D_0}\log\left(\frac{c}{D_0}\right)\right) + \frac{D_1}{N}\left(\frac{a}{D_1}\log\left(\frac{a}{D_1}\right) + \frac{b}{D_1}\log\left(\frac{b}{D_1}\right)\right)$$

Les descripteurs sont ensuite classés par ordre décroissant de valeurs de G' ; comme les termes constants ont été supprimés, la valeur maximale de G' est 0. Cette valeur est obtenue pour un descripteur qui apparaît dans tous les textes pertinents et dans aucun texte non pertinent ou vice-versa, la formule étant symétrique.

5.5.3.4 Une méthode constructive : l'orthogonalisation de Gram-Schmidt

Cette méthode est issue des méthodes utilisées pour trouver la solution des moindres carrés d'un problème linéaire par rapport à ses paramètres. Une description détaillée de cet algorithme et son application à la sélection de modèle NARMAX peut être trouvée dans

[Chen *et al.*, 1989]. Cette procédure itérative classe les descripteurs par ordre décroissant d'importance tout en tenant compte de ceux déjà classés.

Elle a été utilisée avec succès en classification pour sélectionner les entrées de réseaux de neurones pour un problème de classification des collectivités locales en fonction d'un ensemble de ratio financiers [Stoppiglia, 1997] ou pour déterminer les caractéristiques les plus discriminantes d'un fichier client, pour une application dont le but est de mieux cibler les clients potentiels pour un nouveau produit [Stricker et Haré, 1998].

Il existe deux techniques de mise en œuvre de l'algorithme de Gram-Schmidt. La méthode dite "classique" est économe en termes d'occupation de la mémoire, mais elle est très sensible aux erreurs d'arrondi contrairement à la méthode dite "modifiée" qui est numériquement plus stable [Björck, 1967]. Précisons que ces méthodes seraient strictement équivalentes en l'absence d'erreurs d'arrondi. Puisque la taille de la mémoire des machines à notre disposition le permet, nous utilisons l'algorithme de Gram-Schmidt modifié, qui assure la stabilité numérique.

Le principe de cette méthode itérative est de choisir à chaque itération, le "meilleur" descripteur puis de supprimer l'influence de ce descripteur sur les descripteurs restants.

La mise en œuvre est décrite ci-dessous :

Soit Q le nombre de descripteurs candidats et N le nombre d'exemples d'apprentissage avec leurs Q descripteurs possibles et leur sortie associée (qui vaut +1 si l'exemple est pertinent et –1 s'il ne l'est pas). Notons $X_i = {}^T [x_1^i, x_2^i, ..., x_N^i]$ le vecteur des réalisations pour le descripteur i et $Y = {}^T [y_1, ..., y_N]$ le vecteur de dimension N contenant les sorties à modéliser. Le codage des descripteurs est une fonction de la fréquence du descripteur *i*. Soit la matrice X (N, Q)

$$X = \begin{matrix} x_1^1 & \dots & x_1^j & \dots & x_1^Q \\ \dots & & \dots & & \\ \dots & & x_j^j & \dots & = [X_1, \dots X_Q] \\ \dots & & \dots & & \\ x_N^1 & & x_N^j & \dots & x_N^Q \end{matrix}$$

Le modèle s'écrit $Y=X\theta$, la matrice X étant la matrice des entrées et le vecteur Y le vecteur de sortie.

A la première itération de l'algorithme, il faut trouver le vecteur d'entrée le plus corrélé avec la sortie. Pour cela, on calcule le carré des cosinus des angles entre le vecteur de sortie et les vecteurs d'entrée selon la formule :

$$\cos^2(X_P, Y) = \frac{(X_P^T.Y)^2}{(X_P^T.X_P).(Y^T.Y)}$$

Le vecteur sélectionné est celui pour lequel cette quantité est maximale. On élimine ensuite la contribution de l'entrée sélectionnée en projetant le vecteur de sortie ainsi que tous les vecteurs d'entrée restants sur le sous-espace orthogonal au vecteur sélectionné.

La procédure se poursuit en choisissant, une nouvelle fois, le vecteur d'entrée projeté qui explique le mieux la sortie projetée. La procédure se termine lorsque tous les vecteurs d'entrées ont été ordonnés.

La mise en œuvre de cette méthode nécessite la mise en mémoire de la matrice X ; pour que les calculs soient réalisables en un temps raisonnable, il est nécessaire de limiter la taille de cette matrice. Cette méthode doit donc être précédée d'une première sélection de descripteurs pour limiter l'espace de recherche. La sélection de descripteurs se fait donc uniquement dans l'espace des descripteurs sélectionnés lors de la détermination du vocabulaire spécifique, expliquée au paragraphe 5.3.

On cherche donc dans le vocabulaire spécifique d'un thème les descripteurs les plus discriminants.

5.5.3.5 Comparaison des approches

Avant de voir la mise en œuvre de ces différentes méthodes sur des exemples pratiques de catégorisation de textes, nous faisons quelques remarques sur leurs différences.

Utilisation des fréquences

Les méthodes du chi-2 et de l'information mutuelle reposent toutes les deux sur un décompte des apparitions des mots. Ces deux mesures se fondent uniquement sur la présence ou l'absence d'un mot dans un document sans prendre en considération l'occurrence. À l'opposé, la méthode d'orthogonalisation de Gram-Schmidt utilise explicitement l'occurrence des mots dans un texte ou une fonction de cette occurrence.

Prise en considération des corrélations

La méthode de Gram-Schmidt peut être qualifiée de constructive : le modèle est construit progressivement en partant d'un ensemble vide de descripteurs puis en les ajoutant un à un en tenant compte des descripteurs préalablement sélectionnés.

Donc, contrairement aux deux méthodes statistiques, elle tient compte des corrélations éventuelles entre les descripteurs. Dans le cas limite où deux descripteurs t_1 et t_2 sont systématiquement associés, ils sont sélectionnés tous les deux par les méthodes de l'information mutuelle et du chi-2 avec le même score, alors qu'un seul de ces deux descripteurs est sélectionné par la méthode de Gram-Schmidt. Or la présence de descripteurs redondants dans le classifieur ajoute un paramètre inutile, et risque de dégrader les performances.

Mots négatifs et positifs

Les formules de calcul de l'information mutuelle et du chi-2 sont totalement symétriques pour la classe des documents pertinents et celle des documents non pertinents. Plus précisément, cela signifie que certains descripteurs peuvent être choisis parce qu'ils sont caractéristiques des documents pertinents, et d'autres parce qu'ils sont caractéristiques des documents non pertinents.

Dans toute la suite, les descripteurs caractéristiques des textes pertinents sont appelés *mots positifs* et les descripteurs caractéristiques des textes non pertinents sont appelés *mots négatifs*. Cette dénomination vient du fait que si l'on considère un modèle linéaire du type $Y=X\theta$ (la matrice X étant la matrice des descripteurs et le vecteur Y étant le vecteur de sortie avec le codage +1 pour les textes pertinents et –1 pour les textes non pertinents), alors les composantes de θ associées aux mots dits positifs seront positives et les composantes associées aux mots dits négatifs seront négatives.

La présence de mots positifs dans un texte tend donc à indiquer que ce texte est pertinent, et la présence de mots négatifs tend à indiquer qu'il n'est pas pertinent.

En revanche, dans notre implémentation de l'algorithme de Gram-Schmidt nous ne prenons comme descripteur potentiel dans la matrice X que les mots issus du vocabulaire spécifique du thème donc, par construction, des mots positifs. Il faut noter cependant que rien n'interdit d'ajouter dans la matrice X des descripteurs négatifs afin qu'ils puissent éventuellement être

sélectionnés par la méthode de Gram-Schmidt ; l'impact de l'utilisation de ces mots est étudié au paragraphe 5.5.6.

Défauts communs

Chacune de ces méthodes sélectionne les descripteurs pour leur pouvoir discriminant, mais deux descripteurs peuvent avoir un pouvoir discriminant très faible pris séparément, alors que la présence simultanée de ces deux descripteurs peut avoir un rôle important. Pour le thème des *participations*, par exemple, les deux descripteurs *droits* et *vote* ne sont pas des descripteurs très intéressants pris séparément, mais l'interaction de ces deux mots forme le concept de *droits de vote* qui a un sens très précis. Aucune des méthodes présentées ci-dessus ne répond à ce problème.

L'interaction éventuelle entre plusieurs variables peut être prise en considération dans la méthode d'orthogonalisation de Gram-Schmidt en ajoutant des polynômes. Par exemple, au lieu de se contenter de construire la matrice X avec uniquement x_1 et x_2 il est possible d'ajouter le produit $x_1.x_2$ dans la matrice. Dans le cas de la sélection de descripteurs pour le filtrage de textes, le nombre de candidats potentiels est très grand, donc le nombre de monômes l'est également. De plus il est nécessaire de tenir compte de la distance entre les mots : *droits* et *vote* peuvent être dans le même texte sans que l'association *droits de vote* ne soit présente.

Finalement, même avec l'orthogonalisation de Gram-Schmidt, il est difficile de prendre en considération les interactions éventuelles ; néanmoins nous verrons au chapitre 8 comment ce problème est résolu au moins partiellement.

Notons enfin que ces méthodes ne tiennent pas compte des synonymes de la langue ; si un descripteur représentant un mot a été sélectionné dans la liste des descripteurs, ses synonymes n'y figurent pas.

5.5.3.6 Exemple de mise en œuvre des différentes méthodes

Les trois méthodes décrites ci-dessus ont été mises en œuvre sur trois thèmes du corpus Reuters, afin de comparer leur comportement. Les thèmes étudiés sont *interest*, *oilseed* et *nat-gas*. La Figure 5.13 montre les quinze meilleurs mots sélectionnés par chaque méthode ; les mots figurant en gras dans ces listes sont des mots négatifs.

Il n'est pas possible de juger de la qualité d'une méthode à partir de la liste des mots sélectionnés, mais il est intéressant de constater qu'il existe peu de différences entre toutes ces listes. Globalement, toutes ces méthodes semblent conduire aux même choix de descripteurs.
On peut remarquer la présence du mot *lt* dans la liste des mots sélectionnés par l'information mutuelle : il s'agit en fait d'une marque utilisée dans le corpus pour indiquer un nom d'entreprise. Pour les catégories étudiées ici, il s'agit clairement d'un mot négatif.

La Figure 5.14 montre l'évolution des performances sur la base de test pour chaque modèle en fonction du nombre de descripteurs retenus pour la représentation des textes. Le modèle utilisé est un simple neurone logistique avec un paramètre de régularisation fixé à la valeur 1.0 comme expliqué au chapitre 7. Le nombre de descripteurs varie par incrément de 10, entre 10 et 500.

Vocabulaire spécifique	Information mutuelle	Chi-2	Gram-Schmidt
money	rate	rate	rate
rates	bank	money	customer
england	money	rates	money
rate	rates	bank	prime
prime	**pct**	england	band
point	**lt**	lending	rates
discount	market	prime	england
fed	england	repurchase	dollar
stg	prime	market	discount
k	lending	discount	mcentee
repurchase	repurchase	customer	cuts
funds	**cts**	**pct**	advances
lending	point	point	repurchase
bills	discount	fed	leaves
effective	**company**	band	floating

Vocabulaire spécifique	Information mutuelle	Chi-2	Gram-Schmidt
soybean	soybean	soybean	soybean
soybeans	soybeans	soybeans	soybeans
tonnes	**lt**	rapeseed	rapeseed
usda	agriculture	oilseeds	oilseed
corn	corn	oilseed	oilseeds
agriculture	usda	crushers	seaforth
grain	tonnes	corn	inspections
crop	rapeseed	agriculture	romero
rapeseed	grain	usda	sorghum
shipment	oilseeds	sunflower	peanuts
oilseeds	wheat	bushels	rouen
grains	oilseed	meal	asa
farmers	crushers	crushing	disappearance
oilseed	**u**	sorghum	agriculture
bought	crop	inspections	hrw

Vocabulaire spécifique	Information mutuelle	Chi-2	Gram-Schmidt
gas	gas	cubic	gas
natural	natural	gas	cubic
cubic	cubic	natural	butane
feet	feet	feet	exploration
barrels	oil	barrels	favored
exploration	barrels	exploration	cameron
energy	exploration	trillion	natural
reserves	production	liquids	elecetric
petroleum	energy	oil	pel>
oil	reserves	condensate	nmot
pipeline	petroleum	thousand	alaskan
drilling	trillion	discoveries	waterflooding
offshore	drilling	proven	gop>
crude	offshore	proved	stalon
barrel	liquids	offshore	dependency

Figure 5.13 : *Listes des quinze premiers mots sélectionnés par chaque méthode pour les thèmes interest, oilseed et nat-gas. Les mots négatifs sont en gras.*

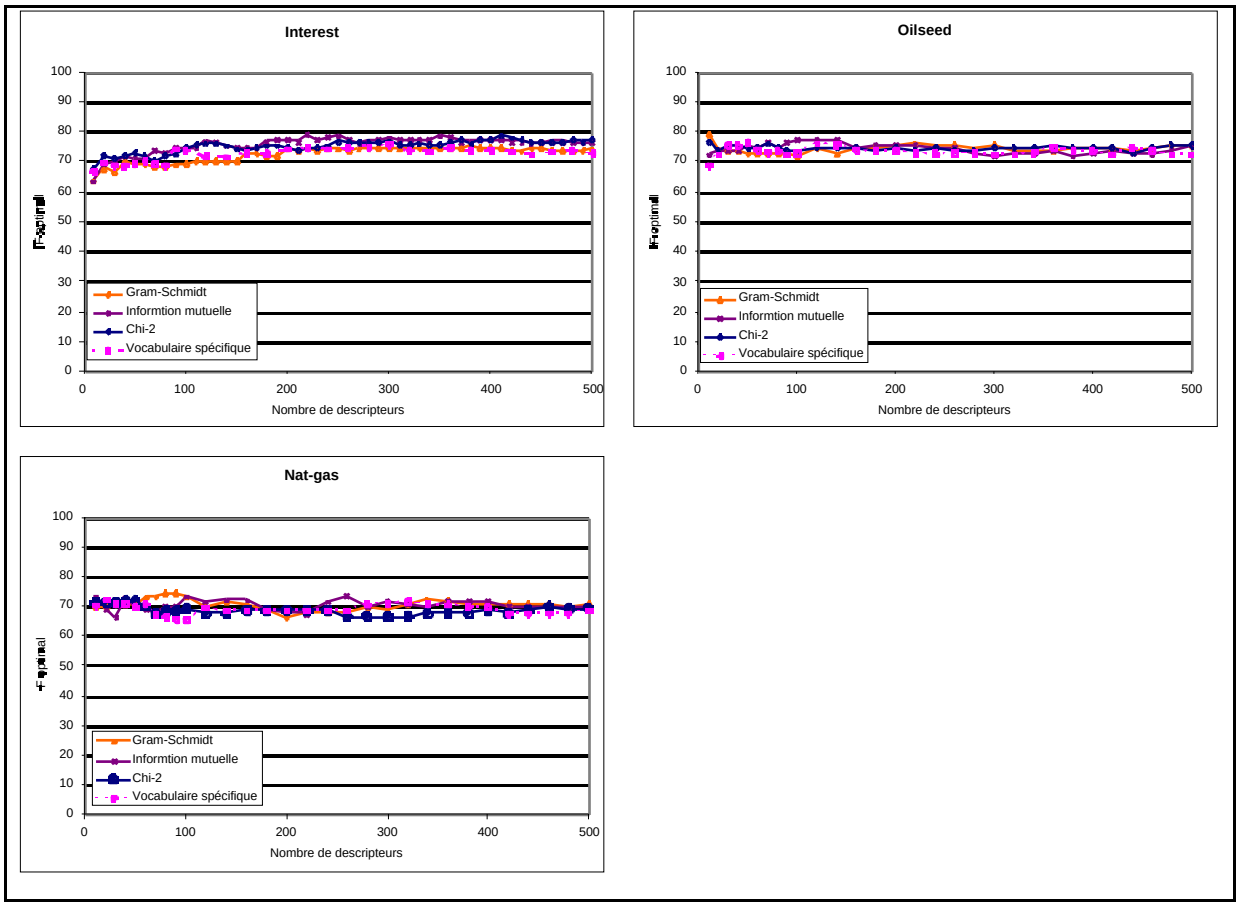

Figure 5.14 : *Évolution des performances sur la base de test en fonction du nombre de descripteurs utilisés pour la représentation des textes. La mesure de performance utilisée est la mesure de F optimal sur la base de test définie au chapitre 4.*

Ces courbes montrent qu'au-delà de dix descripteurs, les performances n'évoluent presque plus, ce qui suggère que peu, voire très peu, de descripteurs suffisent, et qu'il existe un nombre optimal de descripteurs à retenir. Avec les dix premiers descripteurs trouvés par chaque méthode, les performances sur ces trois catégories sont déjà très proches de l'optimum.

Des expériences ont également été effectuées en augmentant encore le nombre de descripteurs utilisés : soit les performances stagnent, soit elles décroissent.

Sur ces catégories, aucune méthode ne semble être largement supérieure ou inférieure aux autres, et, malgré leurs différences, ces méthodes donnent des résultats assez proches, y compris la méthode du vocabulaire spécifique qui est pourtant la plus élémentaire.

Des expériences supplémentaires sont nécessaires avant d'affirmer la supériorité d'une méthode par rapport une autre. Nous les présenterons au chapitre 7.

5.5.4 Utilisation d'un critère d'arrêt pour choisir le nombre de descripteurs

5.5.4.1 Faut-il utiliser un critère d'arrêt ?

Les courbes obtenues à la Figure 5.14 suggèrent qu'au-delà d'un certain nombre de descripteurs, il n'est plus utile d'ajouter de nouveaux descripteurs, car les performances sur la base de test n'augmentent plus. Il semble donc utile d'utiliser un critère d'arrêt afin de limiter le nombre de descripteurs retenus pour la représentation des textes : le nombre de paramètres du modèle sera moins élevé et tous les calculs d'optimisation et de traitement seront plus rapides. Cependant, lorsque l'optimum est atteint, les performances ne décroissent pratiquement pas ; par conséquent, le nombre de descripteurs à retenir ne semble pas être un paramètre très critique et n'a pas besoin d'être déterminé très précisément.

Si l'on appelle k le nombre de descripteurs retenus, beaucoup d'auteurs se contentent de choisir simplement k *a priori* [Lewis, 1992] [Mouliner, 1997] [Dumais *et al.*, 1998] en fixant par exemple $k = 50$ ou $k = 100$.

Une solution extrême consiste à n'utiliser aucun critère d'arrêt : l'ensemble des descripteurs est utilisé. Ainsi, dans ses expériences sur le corpus Reuters, [Joachims, 1998] conserve 9947 descripteurs correspondant à 9947 mots différents pour ses modèles utilisant les machines à vecteurs supports. Ce nombre correspond à l'ensemble des descripteurs disponibles sur le corpus en utilisant des racines lexicales et après suppression des mots vides, et des mots apparaissant dans moins de trois documents. Ce cas correspond à un cas critique où aucune sélection de descripteurs n'est faite. Pour prouver son assertion, il classe les 9947 mots selon leur information mutuelle moyenne pour la catégorie *acquisition (acq)* et il choisit comme entrées, soit les 200 premiers descripteurs de la liste, soit les descripteurs de rang 201 à 500, puis de 501 à 1000, puis de 1001 à 2000, puis de 2001 à 4000, et enfin de 4001 à 9947. Dans son expérience, même avec la dernière partie de la liste donc avec les descripteurs les moins discriminants, il obtient des performances qui sont nettement meilleures que le hasard. Il en déduit donc que, même vers la fin de la liste, il reste des descripteurs discriminants et que, comme il existe peu de chance que tous ces descripteurs soient redondants avec les descripteurs du début de la liste, il est nécessaire de les conserver tous.

Dans une autre étude, [Yang et Perdersen, 1997] étudient l'impact de plusieurs sélections de descripteurs sur deux modèles différents : les k plus proches voisins et une méthode de régression fondée sur les moindres carrés. Selon leurs résultats, lorsque le nombre de descripteurs utilisés devient trop élevé, les performances de leurs classifieurs diminuent ou n'augmentent plus. Leurs modèles utilisent environ 300 descripteurs.

5.5.4.2 Utilisation d'un critère d'arrêt

Dans notre approche, un critère d'arrêt est couplé à la méthode de d'orthogonalisation de Gram-Schmidt afin de déterminer automatiquement le nombre de descripteurs optimal pour chaque thème.

Cette opération est effectuée en utilisant un vecteur aléatoire qui est classé par la méthode de Gram-Schmidt exactement comme les autres descripteurs. Les descripteurs classés après ce vecteur aléatoire sont considérés comme non pertinents pour le problème posé.

Dans la pratique, le rang de ce vecteur aléatoire est en fait un nombre aléatoire. Il faut donc calculer la fonction de distribution de probabilité de l'angle entre un vecteur aléatoire et le vecteur de sortie. Le calcul de la probabilité qu'un vecteur aléatoire soit plus pertinent que l'un des n descripteurs sélectionnés après n itérations a été développé dans [Stoppiglia, 1997]. Chaque fois qu'un descripteur est sélectionné, on peut calculer la probabilité qu'un descripteur aléatoire soit plus pertinent que ce descripteur, et au-delà d'un seuil prédéfini (typiquement 1% ou 5%), tous les descripteurs sont éliminés.

Lors de l'implémentation, cette probabilité est calculée juste après la sélection d'un descripteur et, si le seuil est dépassé, la procédure s'arrête automatiquement sans chercher à classer les descripteurs restants. Donc, même si le nombre de descripteurs candidats est élevé, la procédure s'achève et aucun calcul inutile n'est effectué.

5.5.5 Comparaison des performances des méthodes de sélection de descripteurs

5.5.5.1 Description de l'expérience

Les méthodes de l'information mutuelle, de l'orthogonalisation de Gram-Schmidt et du vocabulaire spécifique sont testées sur un ensemble de thèmes pour comparer leurs performances relatives. Dans une première série d'expériences le nombre de descripteurs

retenus en entrée du classifieur est choisi arbitrairement, et, dans une deuxième série, le critère d'arrêt décrit au paragraphe 5.5.4.2 est couplé à la méthode d'orthogonalisation de Gram-Schmidt.

Pour ces expériences, le modèle neuronal est identique à celui utilisé pour les expériences du paragraphe 5.5.3.6.

Les expériences sont effectuées sur les thèmes 1 à 40 du corpus Reuters[1]. Le nombre de documents pertinents sur la base d'apprentissage varie de 2877 pour le thème 1 (thème *earn)* à 24 pour le thème 40 (thème *sorghum*). Ces quarante thèmes permettent de comparer les différentes méthodes de sélection de descripteurs sur des thèmes comportant beaucoup d'exemples pertinents comme sur des thèmes en comportant peu.

La méthode du chi-2 a été éliminée des comparaisons, car

- il a été montré que les résultats obtenus par cette méthode étaient très proches de ceux obtenus par la méthode de l'information mutuelle [Yang et Perdersen, 1997] et cette tendance a été vérifiée sur les trois thèmes de la Figure 5.14,
- selon cette même étude, la méthode de l'information mutuelle semble être légèrement plus performante.

De plus, on souhaite comparer des méthodes avec des approches différentes et les deux méthodes de l'information mutuelle et du chi-2 reposent sur les mêmes principes. Les trois différences majeures entre les trois méthodes retenues sont :

- la prise en considération de mots négatifs pour l'information mutuelle,
- la prise en considération des corrélations éventuelles avec l'algorithme de Gram-Schmidt,
- l'utilisation explicite de la fréquence d'apparition des descripteurs pour la méthode de Gram-Schmidt.

[1] Les thèmes sont ordonnés selon le nombre de documents pertinents sur la base d'apprentissage selon la présentation du corpus faite au chapitre 3.

5.5.5.2 Comparaison sans optimisation du nombre de descripteurs retenus

Pour cette première série d'expériences, le nombre de descripteurs retenus pour chaque méthode (le paramètre noté *k*) est fixé à 50, 100, ou à 200.

Pour chacun des thèmes, les performances sont mesurées avec la mesure de *F* optimale décrite au chapitre 4 sur l'ensemble de test et les macro-moyennes sont calculées sur l'ensemble des thèmes. La Figure 5.15 présentent les performances obtenues pour chacune des méthodes sur les thèmes 1 à 40 et sur les deux sous-ensembles de thèmes 1 à 20 et 21 à 40 pour observer d'éventuelles corrélations entre le comportement d'une méthode et le nombre de documents pertinents sur la base d'apprentissage.

	Information mutuelle		
	k = 50	*k* = 100	*k* = 200
Thèmes 1 à 40	81,9	81,2	80,3
Thèmes 1 à 20	83,0	83,4	82,8
Thèmes 21 à 40	80,8	79,0	77,7

	Gram-Schmidt		
	k = 50	*k* = 100	*k* = 200
Thèmes 1 à 40	81.6	81.4	81,1
Thèmes 1 à 20	82,3	82,2	82,5
Thèmes 21 à 40	80,8	80,7	79,8

	Vocabulaire spécifique		
	k = 50	*k* = 100	*k* = 200
Thèmes 1 à 40	81,3	81,2	81,1
Thèmes 1 à 20	82,2	82,8	82,5
Thèmes 21 à 40	80,4	79,6	79,8

Figure 5.15 : *Macro-moyennes sur les thèmes 1 à 40, 1 à 20, 21 à 40 du corpus Reuters pour chaque méthode de sélection de descripteurs. k est le nombre de descripteurs utilisés en entrée du classifieur.*

Ces résultats montrent que les trois méthodes de sélection de descripteurs testées conduisent à des classifieurs dont les performances sont très proches en moyenne ; ces méthodes semblent

être globalement équivalentes. La méthode du vocabulaire spécifique donne des résultats similaires aux deux autres méthodes bien qu'elle soit beaucoup plus simple.

Comme l'avaient suggéré les courbes de la Figure 5.14, le nombre de descripteurs retenus n'est pas une valeur très critique puisque, quelle que soit la méthode de sélection, les performances sont peu affectées par la valeur de *k* ; cependant, à performance égale, il est toujours préférable de choisir des valeurs de *k* faibles afin d'obtenir des modèles comportant moins de paramètres ajustables par apprentissage.

5.5.5.3 Utilisation du critère d'arrêt

Le critère d'arrêt est couplé à la méthode de Gram-Schmidt comme expliqué au paragraphe 5.5.3.4 pour déterminer le nombre optimal de descripteurs pour chaque thème, avec un seuil de 1%. Les résultats sur les mêmes thèmes que précédemment sont présentés Figure 5.16.

	Gram-Schmidt + critère d'arrêt
Thèmes 1 à 40	81,5
Thèmes 1 à 20	82,3
Thèmes 21 à 40	80,7

Figure 5.16 : *Résultats avec la méthode de Gram-Schmidt couplée avec le critère d'arrêt.*

Les résultats sont là encore très proches de ceux trouvés à la Figure 5.15 ; l'utilisation du critère d'arrêt n'améliore pas les performances, mais, sans les dégrader, elle permet de déterminer automatiquement le nombre de descripteurs retenus.

5.5.6 Impact des mots négatifs pour la méthode de Gram-Schmidt

Comme expliqué précédemment, l'utilisation de mots négatifs est l'une des différences entre l'information mutuelle et l'utilisation de la méthode de Gram-Schmidt choisie ici.

Il est également possible d'utiliser des mots négatifs par la méthode de Gram-Schmidt : il suffit pour cela de définir le vocabulaire spécifique de l'ensemble des documents non pertinents pour chaque thème. On dispose alors, pour chaque thème, de deux listes de mots : une liste du vocabulaire spécifique des documents pertinents, et une liste du vocabulaire spécifique des documents non pertinents.

On effectue deux séries d'expériences sur les thèmes précédents, qui se distinguent uniquement par la construction de la matrice *X* en entrée de l'algorithme de Gram-Schmidt :

1. La matrice *X* est construite à partir des 200 premiers mots de la liste du vocabulaire spécifique de l'ensemble des documents pertinents.
2. La matrice *X* est construite à partir des 200 premiers mots de la liste du vocabulaire spécifique de l'ensemble des documents pertinents plus les 50 premiers mots de la liste du vocabulaire spécifique de l'ensemble des documents non pertinents.

Donc, pour la première expérience, seuls des mots positifs peuvent être choisis, et, pour la seconde, l'algorithme peut choisir soit des mots positifs soit des mots négatifs.

Les résultats sur les thèmes 1 à 40 du corpus Reuters sont présentés à la Figure 5.17 (les résultats de l'expérience prenant en considération uniquement les mots positifs sont différents de ceux présentés à la Figure 5.16, car le choix des documents non pertinents utilisés pour fabriquer les bases d'apprentissage est légèrement différent).

	Gram-Schmidt mots positifs	Gram-Schmidt mots positifs et négatifs
Thèmes 1 à 40	82,1	82,2
Thèmes 1 à 20	82,7	82,8
Thèmes 21 à 40	81,6	81,7

Figure 5.17 : *Comparaison des résultats entre la méthode de sélection ne prenant en compte que les mots positifs et celle prenant en compte les mots positifs et négatifs.*

Les résultats obtenus sont identiques avec la méthode qui prend en compte les mots négatifs et les mots positifs, ce qui laisse à penser que la méthode de Gram-Schmidt ne considère pas les mots négatifs comme discriminants. La Figure 5.18 qui montre la comparaison catégories par catégories entre les deux méthodes confirme qu'il n'existe pratiquement aucune différence entre les deux expériences.

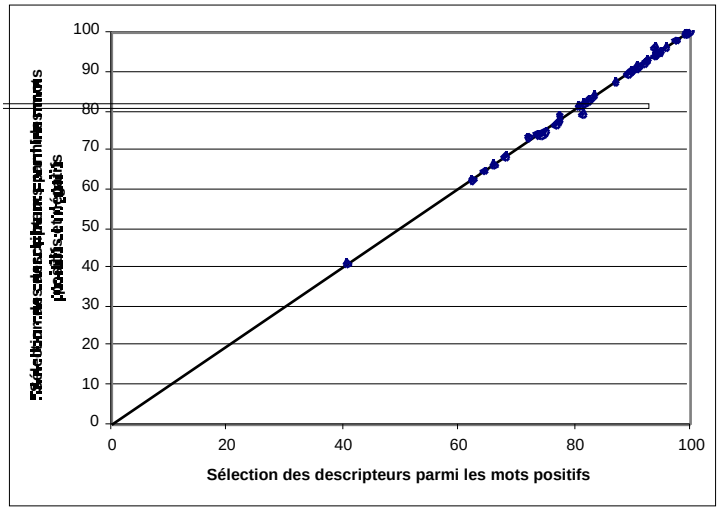

Figure 5.18 : *Comparaison catégorie par catégorie.*

Il est important de noter que sur le corpus Reuters, la catégorie *earn* représente 2877 documents sur la base d'apprentissage, et 1087 documents pertinents sur la base de test, soit à peu près 33% de la base de test. De plus, le vocabulaire utilisé par les textes de cette catégorie est assez particulier, comme le montre la liste des vingt premiers mots de la liste du vocabulaire spécifique de ce thème présentée à la Figure 5.19.

```
revs, th, shr, note, div, prior, loss, profit, dividend, shrs, avg, qtly,
qtr, includes, mths, jan, sets, vs, gain, nine
```

Figure 5.19 : *Liste des vingt premiers mots du vocabulaire spécifique de la catégorie earn.*

La Figure 5.20 montre la liste du vocabulaire négatif pour les trois thèmes *interest, oilseed* et *nat-gas* obtenue en cherchant le vocabulaire spécifique des documents non pertinents.

Interest		*Oilseed*		*Nat-gas*	
418	revs	390	revs	389	revs
373	th	374	th	370	th
299	note	285	note	287	note
292	shr	279	shr	277	shr
266	prior	251	prior	259	prior
262	div	241	div	255	div
246	dividend	235	loss	239	loss
234	profit	229	dividend	230	profit
228	shrs	212	profit	225	shrs
224	includes	211	manager	218	avg

Figure 5.20 : *Listes du vocabulaire négatif pour les trois thèmes interest, oilseed et nat-gas.*

Ces deux remarques signifient que le corpus Reuters présente la particularité de se prêter *a priori* à l'utilisation de mots négatifs puisqu'une catégorie prédomine largement, et que, de plus, elle utilise un vocabulaire spécifique comme les mots *revs*, *vs*, *shr*, *qtr* qui sont des abréviations rarement utilisées dans un contexte différent de celui de la catégorie *earn ;* le texte Figure 5.21 est un exemple de l'utilisation de ces abréviations.

```
UNIVERSAL HOLDING CORP 4TH QTR NET
Shr NA Net profit 2,000 vs profit 195,000 Revs 2,623,000 vs 2,577,000 Year
Shr NA Net loss 425,000 vs profit 278,000 Revs 15.4 mln vs 8,637,000
```

Figure 5.21 : *Exemple de texte appartenant à la catégorie earn.*

Pour le corpus utilisé dans TREC-8, le thème le plus dense contient 186 documents pertinents sur la base de test, ce qui représente une densité de l'ordre de 0,1% ; il est donc peu probable que le fait de savoir qu'un document n'appartient pas à cette catégorie constitue une information pour trouver les autres thèmes.

5.6 Conclusion

Ce chapitre a permis d'introduire la représentation la plus utilisée pour la catégorisation de textes et le niveau d'information principalement utilisé : la présence ou l'absence de certains mots pour trouver automatiquement le sens d'un texte.

Le très grand nombre de descripteurs potentiels rend nécessaire l'utilisation d'une méthode de sélection. Nous avons introduit une méthode originale par rapport à la littérature de la catégorisation de textes : la méthode d'orthogonalisation de Gram-Schmidt précédée de la détermination du vocabulaire spécifique d'un ensemble de textes.

Dans les expériences effectuées, toutes ces méthodes se sont révélées à peu près équivalentes ; néanmoins, dans la suite, nous utiliserons la méthode d'orthogonalisation de Gram-Schmidt couplée à la détermination du vocabulaire spécifique pour plusieurs raisons :

- On montrera au chapitre 7, qu'en modifiant la construction de la matrice X, il est possible d'améliorer les performances.
- Elle tient compte des corrélations éventuelles entre les mots, et, dans le cas limite où deux mots apparaissent systématiquement ensemble, la méthode de Gram-

Schmidt, contrairement aux autres méthodes, ne sélectionne qu'un seul de ces deux mots.

- Elle tient compte des mots précédemment sélectionnés.
- Elle utilise la fréquence des mots dans les textes et non pas uniquement la présence ou l'absence de mots.
- L'utilisation d'un critère d'arrêt optimise le nombre de descripteurs pour chaque thème. Il est également possible d'utiliser des critères d'arrêt avec les méthodes de l'information mutuelle et du chi-2, mais la justification théorique de ces critères semble moins bien établie.
- L'utilisation de mots dits négatifs ne semble pas apporter d'amélioration, même sur le corpus Reuters qui a pourtant une configuration particulière du fait de la répartition des catégories.

Cette méthode présente néanmoins des défauts, comme les deux autres méthodes ; les deux principaux semblent être :

- Le système ne tient pas compte des synonymes.
- Deux mots peuvent avoir un pouvoir discriminant très faible pris séparément et n'être pas sélectionnés alors que l'union de ces deux mots a un pouvoir discriminant fort. Sur le thème *participation*, par exemple, les deux mots *entrée* et *hauteur* pris séparément ne sont pas de bons descripteurs relativement à ce thème, mais l'expression *entrée à hauteur* est souvent utilisée dans des phrases dans un contexte de participation comme le montre la phrase ci-dessous :

  Toyata a annoncé son entrée à hauteur de 5,4% dans le capital de Yamaha.

Néanmoins l'ensemble des expériences a montré que la détermination du vocabulaire spécifique qui est beaucoup plus simple permet d'obtenir de bons résultats, et des comparaisons supplémentaires seront effectuées au chapitre 7.