Projet de Programmation Numérique
Analyse de Champs Lexicaux dans un Corpus de Texte

YANN KERGUTUIL ROLIH MEYNARD JOHN GLIKSBERG
M1 IHPS UVSQ

1 Conception du projet : Mise en valeur d'un champ lexical dans un corpus de texte

Introduction

Le thème du projet est de parvenir à extraire des groupes de mots contenu dans de nombreux documents. Les documents en question, dans le cadre de nos travaux, sont des articles Wikipédia. Les démarches proposées consistent à effectuer une analyse matricielle, des calculs de distances et de comparaison vis-à-vis des dictionnaires de mots.
Les deux principales étapes à adopter sont d'une part l'analyse statistique, et de l'autre les méthodes de clustering. Il est à noter aussi qu'un des axes de notre travail est aussi la recherche de mots similaires dans un texte.

Définition d'une méthode statistique

On appelle une méthode statistique une méthode qui se base sur les probabilités d'un événement, notamment sur la probabilité d'occurrence lié à un grand nombre d'événements pour en déduire des corrélations entre ces événements.

Rapport entre analyse statistique et extraction de mots dans des documents

Les méthodes statistiques permettent de repérer de façon automatique la présence de mots dans un corpus de texte. Cette méthode calcule, entre autre, la fréquence d'apparition du mot dans un texte.

Intérêt de l'analyse d'un champ lexical

La numérisation des documents, entre autres des articles, et le développement des technologies du web génèrent d'énormes quantités d'informations écrites dans des documents. Il en ressort, compte-tenu de cette masse de données qui croit de façon exponentielle, il convient de fournir aux internautes un outil capable de sélectionner des documents pour avoir accès à ceux qui sont pertinents. Ainsi, afin de filtrer les documents souhaités il serait intéressant de parvenir à faire ressortir les mots clés présents dans un document. On entend par mots clés les termes précis qui soulignent le sujet du document. Un des intérêts de l'analyse d'un champ lexical est d'améliorer, voire d'optimiser la recherche documentaire sur internet. La problématique est de pouvoir accéder à l'information, le plus rapidement possible.

État de l'art

L'extraction de mots dans un document suppose que l'on soit capable de représenter ce que c'est un corpus de texte. La représentation la plus simple d'un corpus de texte est dite "représentation sac de mots". Concrètement, il s'agit de transformer des textes en vecteurs où chaque élément représente un mot. Cette représentation ne tient pas compte de l'aspect grammatical du mot et de toute notion de distance entre les mots. Par ailleurs, l'autre représentation de documents textuels que l'on peut citer est la représentation dite "sac de phrases". Celle-ci consiste de favoriser les mots susceptibles d'avoir une signification. Par conséquent, on peut supposer que la méthode appelée "sac de phrases" est beaucoup plus pertinente que la méthode nommée "représentation sac de mots". D'autre part, il existe une autre méthode à propos de la représentation des documents textuels. Il est question d'une méthode qui repose sur les techniques de lemmatisation et qui vise à chercher la racine lexicale d'un terme. Par exemple, on peut citer la forme de l'infinitif singulier pour les verbes et pour les noms. Enfin, on peut également évoquer la méthode qui traite les documents de textes indépendamment de la langue utilisée. Cette méthode est basée sur les "n-grammes". Par définition, un "n-gramme" est une séquence de n caractères consécutif.
Mathématiquement, un document de texte (d_i) peut être représenté par un vecteur à composante numérique de la manière suivante :

$$D_i = (V(1_i), V(2_i), \cdots V(|T|_i))$$

On note que T est l'ensemble des mots qui apparaissent au moins une fois dans le corpus. De plus, $|T|$ représente la taille du vocabulaire et les $V(k_i)$ sont la fréquence du mot dans le document.
Certaines méthodes statistiques cherchent à calculer la fréquence $V(ki)$ pour chaque mot. De ce fait, on peut en déduire qu'il y a moyen aussi de calculer le nombre de documents qui contient le même terme.

Mesure de similarité

Il existe plusieurs mesures de similarité pour évaluer la similitude des mots dans les documents. En effet, on peut les lister ainsi :
La distance euclidienne
La distance euclidienne entre deux vecteurs T_i et T_j dans un espace multidimensionnel est :

$$D(T_i, T_j) = \sqrt{\sum_k k(x(k) * T_i - x(k) * T_j)^2}$$

La distance Manhattan
La distance Manhattan entre deux vecteurs T_i et T_j dans un espace multidimensionnel est défini par :

$$D(T_i, T_j) = \sum_k |x(k) * T_i - x(k) * T_j|$$

La distance cosinus
Enfin, la distance cosinus entre deux vecteurs Ti et Tj dans un espace multidimensionnel est :

$$cos(T_i, T_j) = \frac{T_i \times T_j}{|T_i| \times |T_j|}$$

La matrice de similarité de mots est une matrice symétrique de taille $N_d \times N_d$, où N_d est le nombre de documents à classifier, de diagonale nulle (pour les distances euclidiennes et Manhattan) et de diagonale égale à 1 pour la distance cosinus. Les indices i et j représentent les index des documents du corpus à classifier.
Algorithme de clustering
Pour pouvoir extraire des mots dans des documents, on peut appliquer l'algorithme de clustering. Il existe différentes méthodes liées à cet algorithme :

- Classification hiérarchique : arbres de classes
  - Classification hiérarchique ascendante : Agglomérations hiérarchiques
  - Classifications hiérarchique descendante : division successives
- Classification à plat : algorithmes des k-moyennes (partition)

Les différentes étapes de l'analyse d'un champ lexical dans un corpus de textes
Pour effectuer l'analyse d'un champ lexical dans un document, on propose de suivre les étapes suivantes :

- Le découpage en mots : on va remplacer chaque article par la liste de ses mots. Ainsi, on posera N_d le nombre de documents
- L'indexation : dans chaque document, un mot sera associé à sa fréquence d'apparition dans ce document. De ce fait, on notera la fréquence du mot $freq_w$ et N_w le nombre de mots uniques total dans le document.
- Création du dictionnaire : on affectera à chaque mot son identifiant w_{id}. Aussi,les documents auront leur propre identifiant noté d_{id}
- La compression : on devra construire une matrice où l'on pourra récupérer le couple $(w_{id}, score_w)$. On définit le score d'un mot de la façon suivante :

$$\text{score} = \frac{\text{fréquence\_du\_mot\_dans\_le\_document}}{\text{fréquence\_du\_mot\_dans\_le\_dictionnaire}}$$

- L'accès du score devra se faire d'une part par une matrice de la forme
  Documents[d_{id}][w_{id}]
- L'indexation bis inversée : le but est d'obtenir le score d'un mot par une matrice de la forme
  Index_inv[w_{id}][d_{id}]
- Construction des matrices Matrice_document et Matrice_Words de taille respectivement N_d pour nombre de documents et N_w pour nombre de mots
- Calculs de distances entre les matrices Matrice_document et Matrice_Words
- Clusteriser les matrices Matrice_document et Matrice_words

2 Étude de la recherche de mots similaires

Introduction

La recherche de mots similaires est importante, car on peut avoir dans plusieurs documents différents l'apparition d'un même terme. De plus, cette étude est une partie intégrante dans la recherche fondamentale en informatique. Les applications de la recherche de mots sont, pour l'essentiel, dans la gestion et le traitement de données textuelles. Les exemples d'applications que l'on peut énumérer sont les suivants :

- La recherche documentaire
- La détection de plagiat
- Les moteurs de recherche
- La compression de texte
- La correction d'erreurs
- L'encodage de l'information
- Le traitement de texte

Dans notre recherche de mots identiques, on étudie les mots qui ont le même nombre de lettres et qui s'écrivent de la même manière dans le corpus de texte. Pour ce faire, on va devoir construire une matrice qui contient un ensemble de mots. On va considérer que la présence d'un mot correspond à son identifiant et à chacune d'elles on associe la fréquence du mot dans les documents textuels.

Description de mots identiques

Par ailleurs, on peut décrire ce que l'on entend par mots identiques. Ainsi, différents critères peuvent intervenir dans la définition de ce qu'est mot identique :

- Les fréquences des lettres dans un mot
- Le nombre de lettre dans un mot
- Le positionnement des lettres les unes par autres dans un mot

Éxemple de la recherche de mots identiques dans la bio-informatique

La recherche en bio-informatique met en adéquation l'informatique et les mathématiques pour résoudre les problématiques liées à la biologie. Il s'agit du décryptage de la bio-information. La recherche de mots similaires a une grande importance dans la bio-informatique, notamment dans le problème de l'analyse des séquences biologiques. Les problématiques de séquençage de génomes sont assez récurrentes en bio-informatique. Ainsi, le but est de pouvoir prévoir et d'analyser les séquences des génomes. Les principales problématiques de séquençage sont les suivantes :

- La recherche de motif dans une séquence
- La recherche de similarité entre les séquences
- La recherche de similitude d'une séquence avec l'ensemble des séquences d'une base de données
- L'alignement de séquences
- La découverte de motifs

On va considérer qu'un mot est défini et composé l'ensemble des lettres suivantes : $\{A, C, G, T\}$. Ainsi, on va chercher à déterminer combien de fois on retrouve le motif $ACGT$. Il s'agira d'une part d'évaluer la structure de ce mot, puis de déterminer sa fréquence dans la séquence du génome. De ce fait, il est question d'effectuer une analyse du champ lexical car on cherche à extraire ce mot (motif) et à calculer sa fréquence dans le document, en sachant qu'ici le document se réfère à l'ADN.

Construction de la matrice des occurrences

Pour pouvoir effectuer une analyse complète du champ lexical dans un corpus de texte, la première démarche consiste à construire une matrice des fréquences des mots dans les documents de taille $N_{wd} \times N_{docs}$, où N_{wd} est le nombre de mots uniques au total dans un document et N_{docs} est le nombre de documents. Chaque ligne de la matrice représente un terme donné par son identifiant w_{id}, et chaque colonne représente son document donné par son identifiant d_{id}. On notera que les mots considérés non pertinents sont pour la plupart les déterminants ou les pronoms par exemple. Ainsi, on peut proposer l'exemple ci-dessous qui construit une matrice d'occurrences.
On va considérer les documents suivants :

- Docs1 : "Lorem ipsum dolor sit amet"
- Docs2 : "sed consectetur dolor dolor dolor sed ispum ipsum"
- Docs3 : "ipsum dolor sit amet"
- Docs4 : "consectetur sit sed amet"

Nous pouvons, par conséquent, représenter la fréquence simple des mots dans les documents par la matrice X suivante :

	Docs1	Docs2	Docs3	Docs4
lorem	1			
ipsum	1	2	1	
dolor	1	3	1	
sit	1		1	1
amet	1		1	1
sed		2		1
consectetur		1		1

À partir de calculs numériques, on devra réaliser cette matrice.

3 Étude de la méthode de clustering

Introduction

La méthode de clustering est appelée aussi méthode de regroupement. Il s'agit d'effectuer une classification de mots.

Proposition de l'algorithme pour notre analyse de champs lexicaux

Dans le cadre de notre étude, on peut proposer d'utiliser la méthode de regroupement agglomératif hiérarchique. De ce fait, il conviendrait de parvenir à implémenter l'algorithme suivant :

- On calcule la matrice M d'occurrences de mots dans les documents avec $M_{i,j}$ la similarité entre les groupes i et j
- On fusionne les deux groupes les plus similaires, c'est-à-dire les plus proches
- On met à jour la matrice M avec les nouveaux mots constitués
- On répète les étapes 2 et 3 fois jusqu'à l'obtention d'un unique groupe