Accès et Recherche d'Information

RICM 4 - 23 avril 2018

Durée 2 heures

Documents autorisés

Partie I. Recherche d'information textuelle

Question I.1 : On considère une requête q contenant les termes *OS, Jaguar* et trois documents de même taille d_1, d_2 et d_3 qui contiennent respectivement *Jaguar, Jaguar, Jungle, Jungle, Jungle*, et *Système d'exploitation, Jaguar, Mac, Système d'exploitation, Système d'exploitation* et *Jaguar, Bentley, Mercedes, Jaguar, Jaguar*. On prend l'abréviation S.E. pour *Système d'exploitation* et on suppose que le vocabulaire associé est :

$$\mathcal{V} = \{bentley, jaguar, jungle, mac, mercedes, os, S.E.\}$$

I.1.1) Donner les vecteurs associés aux documents et à la requête. Dans le cas où on privilégie une représentation à base de tf, ordonner les documents par rapport à leur score *produit scalaire* avec la requête.

I.1.2) *Jaguar* est un terme polysémique et on voit bien sur l'exemple précédent que si un terme polysémique d'une requête est répété plusieurs fois dans des documents traitants d'autres sujets que ce que l'on recherche, ces documents obtiendront un meilleur score que ceux traitant du sujet mais contenant moins d'occurrences de ce terme polysémique. Une solution est d'augmenter la couverture des termes du vocabulaire en prenant en compte, dans la représentation vectorielle des documents et de la requête, les termes synonymes des termes apparaissant dans les documents et la requête. Un moyen simple pour cela consiste à définir une matrice de similarité W entre les termes et de projeter les documents et la requête sur cette matrice avant de calculer leurs scores. Pour notre exemple, considérons la matrice de similarité entre termes suivante:

$$\begin{array}{c} \\ B \\ Ja \\ Ju \\ Mac \\ Mer \\ OS \\ S.E. \end{array} \begin{array}{c} \begin{array}{ccccccc} B & Ja & Ju & Mac & Mer & OS & S.E. \end{array} \\ \begin{pmatrix} 0,5 & 0,1 & 0 & 0 & 0,4 & 0 & 0 \\ 0,1 & 0,5 & 0,05 & 0,05 & 0,1 & 0,1 & 0,1 \\ 0 & 0,05 & 0,95 & 0 & 0 & 0 & 0 \\ 0 & 0,05 & 0 & 0,8 & 0 & 0,05 & 0,1 \\ 0,4 & 0,1 & 0 & 0 & 0,5 & 0 & 0 \\ 0 & 0,1 & 0 & 0,05 & 0 & 0,55 & 0,3 \\ 0 & 0,1 & 0 & 0,1 & 0 & 0,3 & 0,5 \end{pmatrix} \end{array}$$

Quelles sont les nouvelles représentations des documents d_1, d_2 et d_3 que de la requête q? Calculer les nouveaux scores produits scalaires entre ces documents et q et ordonner ces derniers par rapport à ces scores. Conclure.

I.1.3) Si on suppose que les termes qui apparaissent dans les mêmes documents avec les mêmes fréquences sont sémantiquement similaires, donner un moyen simple de calculer la matrice de similarité entre termes, W.

Question I.2: Le codage par *indice-valeur* consiste à obtenir une représentation compacte des vecteurs de documents en ne codant que leurs caractéristiques non nulles ainsi que les indices associés à ces caractéristiques.

I.2.1) Quel est le codage par indice-valeur du vecteur suivant

$$\vec{v} = (0,0 \quad 0,1 \quad 0,0 \quad 3,1 \quad 5,2 \quad 0,0 \quad 0,0 \quad 1,3 \quad 0,0)$$

I.2.2) Pour une collection contenant 1 349 539 documents et un vocabulaire de taille 604 244, quelle est la taille sur le disque du fichier contenant la représentation pleine des documents de la collection de Wikipédia dans le cas où on utiliserait la pondération *tf-idf* ?

I.2.3) Même question que précédemment mais dans le cas où on utiliserait le codage par indice-valeur. On suppose que chaque document de la collection contient 225 termes. Par quel facteur ce codage permet-t-il de comprimer l'information contenue dans la représentation pleine des documents ?

Le tableau ci-dessous donne la taille supposée des variables intervenant dans ces calculs.

variable	taille (en octet)
caractère	1
entier	4
réel	8

Partie II. Classification par réseaux de neurones profonds

La figure 1 représente de manière simplifiée une variante du réseau "AlexNet" qui a remporté le "ImageNet Large Scale Visual Recognition Challenge" (ILSVRC) en 2012. Ce réseau prend en entrée des images normalisées en taille (224×224). Il donne directement en sortie des scores de détection pour chacun des 1000 concepts cibles. Les éléments suivants ne sont pas détaillés sur la figure:

- les couches neuronales contiennent toutes une partie linéaire (affine en fait car des biais sont utilisés) et une partie non linéaire (la fonction d'activation) ;
- les sixième et septième couches neuronales contiennent en plus en entrée une étape dite de "dropout" ;
- une étape finale de "softmax" est ajoutée à la sortie du réseau.

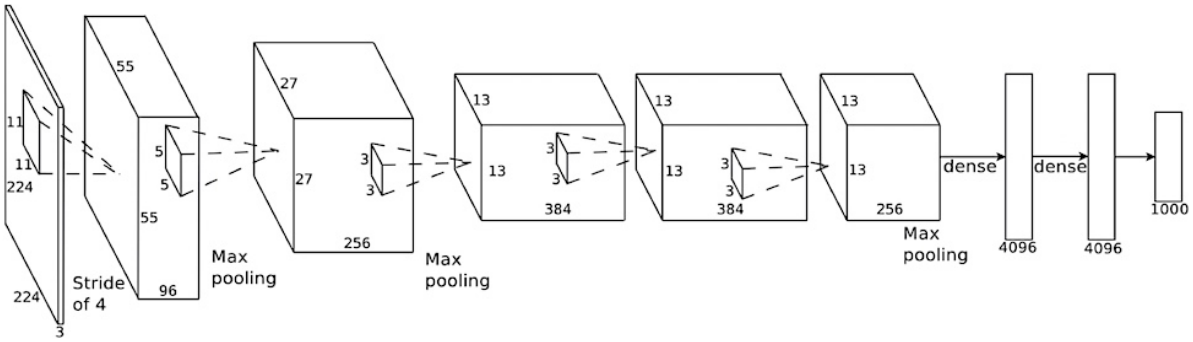

Figure 1: Représentation simplifiée du réseau AlexNet pour la catégorisation d'images.

Question II.1 : Combien de couches "neuronales" ce réseau contient-il ? Quels types de couches neuronales contient-il et combien en contient-il de chaque type ?

Question II.2 : Combien de paramètres contient la première couche neuronale ?

Question II.3 : Combien de connexions contient la troisième couche neuronale ?

Question II.4 : Combien de paramètres contient la huitième couche neuronale ?

Question II.5 : En comptant 1 pour chaque étape linéaire, 1 pour chaque fonction d'activation, 1 pour chaque étape de pooling, 1 pour chaque étape de "dropout", et 1 pour l'étape finale de "softmax", de combien d'étapes au total le réseau est-il composé ?

Question II.6 : Comment la transition entre les couches de convolution et les couches complètement connectées est-elle effectuée ?

Question II.7 : À quoi sert l'étape finale de "softmax" (non représentée sur la figure) et comment est-elle implémentée ?

Question II.8 : Dans quel cas l'étape finale de "softmax" est-elle adaptée et dans quel cas ne l'est-elle pas ?

Partie III. Recherche d'images par similarité visuelle

On veut faire un système de recherche d'image par l'exemple prenant en compte la couleur ou la texture. Pour la couleur, on utilise des histogrammes de couleur par blocs. Les couleurs sont considérées dans l'epace RVB avec 4 intervalles par composante et on prend 4×3 blocs par image. Pour la texture, on considère une repréentation basée sur des filtres de Gabor avec 8 orientations et 6 échelles.

Question III.1 : Quel est le nombre de composantes de chacun des deux descripteurs ?

Question III.2 : Le descripteur de couleur est-il invariant par symétrie gauche-droite ?

Question III.3 : Le descripteur de texture est-il invariant par symétrie gauche-droite ?

Question III.4 : Quelle(s) distance(s) peut-on utiliser pour comparer les images avec chacun de ces descripteurs ?