

Info 4

Examen Accès et Recherche d'Information – 8 avril 2019 – 2 heures

Partie I

Note importante : les parties I et II sont à rédiger sur des copies séparées.

Exercice 1 – Généralités sur les modèles de recherche d'information

Question 1 : Avec le modèle booléen, une solution pour augmenter le rappel est-elle de :

- a) utiliser l'opérateur ET
- b) utiliser l'opérateur OU
- c) utiliser l'opérateur NON

entre les termes de la requête ?

Expliquer brièvement votre réponse sur un exemple.

Question 2 : Lister et expliquer brièvement deux critères qui peuvent être utilisés pour évaluer et comparer les systèmes de recherche d'information.

Exercice 2 : Obtention des termes d'indexation

On considère les 3 documents suivants:

D₁ = "la recherche d'information"

D₂ = "rechercher des informations de recherche"

D₃ = "la recherche en informatique"

Question 1 : En ne considérant aucun anti-dictionnaire ni aucun traitement de troncature, indiquer les termes qui indexent ces 3 documents.

Question 2 : En considérant uniquement l'anti-dictionnaire suivant {à, au, d, de, du, des, elle, elles, en, est, je, il, ils, le, la, les, lui, son, sa, ses, tu, un, une}, indiquer les termes qui indexent les documents.

Question 3 : En considérant, en plus de la question 2, l'utilisation d'une version simplifiée de l'algorithme de Porter ayant les règles suivantes :

1. S → /
2. R → /
3. QU → /
4. ION → /
5. E → /
6. MENT → /

indiquer les termes qui indexent les documents (en décrivant les règles de réécritures utilisées dans chaque cas). Commenter les regroupements de mots obtenus.

Question 4 : Expliquer la difficulté de définir des règles pour l'algorithme de Porter dans le cas général.

Exercice 3 : Le modèle vectoriel

Chacun des 4 documents suivants a été indexé et est représenté dans le modèle vectoriel de recherche d'information par un vecteur à 4 dimensions :

$$D_1 = (1 \ 0 \ 0 \ 0)$$

$$D_2 = (1 \ 1 \ 1 \ 1)$$

$$D_3 = (1 \ 1 \ 0 \ 0)$$

$$D_4 = (0 \ 1 \ 0 \ 1)$$

Question 1 : Si une requête avec deux termes, $Q=(1 \ 0 \ 1 \ 0)$ (basée sur une pondération par tf) est posée sur cette base de 4 documents, indiquer les valeurs de correspondance de ces documents en utilisant la fonction de correspondance cosinus vue en cours. Expliquer brièvement les valeurs obtenues.

Question 2 : Dédurre de votre réponse à la question 1 le résultat trié par ordre de pertinence décroissant pour la requête Q.

Question 3 : En supposant que le corpus ne contienne que ces quatre documents, donner le vecteur requête Q'' si l'on considère uniquement l'idf (formule vue en cours avec log en base 10, noté $\log_{10}$) comme poids et non plus le tf pour la requête Q. Expliquer la modification des poids entre le vecteur requête de la question 1 et ce que vous obtenez. Indiquer dans ce cas en quoi ces poids ont des chances d'être plus intéressants pour la recherche d'information.

Note : $\log_{10}(1)=0$, $\log_{10}(4/3)=0.125$, $\log_{10}(4/2)=0.301$, $\log_{10}(4)=0.602$

Question 5 : Reprendre la question 1 avec la requête Q'' , et expliquer de façon détaillée les différences obtenues et pourquoi ces différences apparaissent.

Exercice 4 : Recherche d'information sur domaine de spécialité

Mettons-nous dans le cas d'une application médicale, dans laquelle on gère des documents de diagnostics médicaux d'examens. Dans certains diagnostics, on peut lire : "tumeur à droite du poumon gauche". Dans d'autres : "tumeur à gauche sur le poumon droit". Supposons que l'on utilise un antidictionnaire avec les mots utilitaires habituels.

Question 1 : Expliquer en quoi le modèle vectoriel vu en cours est inadapté à ce cadre. Donnez brièvement des exemples pour étayer vos arguments.

Question 2 : Proposer une approche (pas trop détaillée !) qui permettrait de résoudre ce problème au niveau de l'indexation des documents. Votre proposition devra être en mesure de traiter les exemples cités ci-dessus.

Question 3 : Discutez votre proposition en indiquant comment vous pourrez, lors de la recherche, être capable de rechercher des documents parlant de "tumeur à droite du poumon gauche" qui ne retienne pas ceux qui parlent de "tumeur à gauche sur le poumon droit".

Question 4 : Discutez, pour votre proposition, la difficulté potentielle pour extraire les termes, pour la pondération, pour les index inversés, etc.

Partie 2

Apprentissage profond (deep learning)

La figure 1 donne une vue générale de l'architecture du réseau "VGG-16" :

Question 1 : combien y a-t-il de couches (ou sous-couches) :

- a) convolution ?
- b) complètement connectée ?
- c) "pooling" ?
- d) "softmax" ?

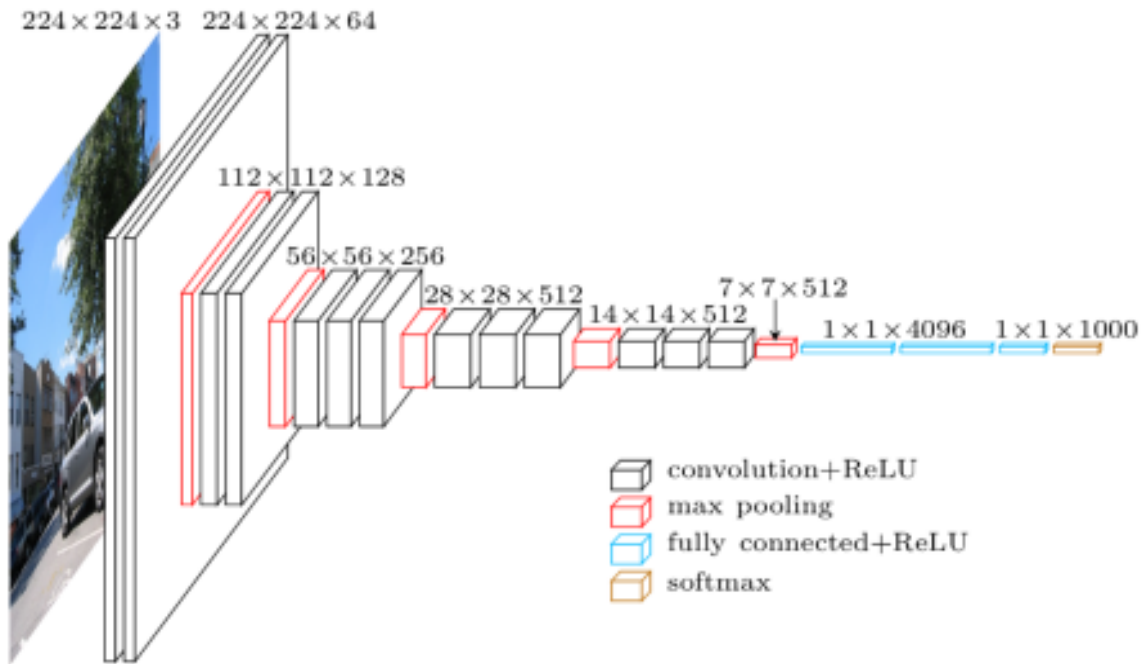


Figure 1 : réseau "VGG-16"

Question 2 : couche "pooling" :

- a) quel est le but de cette couche ?
- b) comment est-elle implémentée ?
- c) quelle autre opérateur associatif peut être utilisé ?
- d) quelle alternative pourrait être utilisée à la place de cette couche ?

Question 3 : couche "softmax" :

- a) quel est le but de cette couche ?
- b) comment est-elle implémentée ?
- c) quelle est fonction de coût adaptée à cette couche de sortie ?
- d) comment se calcule-t-elle ?

Question 4 : (sous-)couche "ReLU" :

- a) quel est le but de cette couche ?
- b) comment est-elle implémentée ?

Question 5 : combien y a-t-il de paramètres à apprendre dans :

- a) la cinquième couche de convolution ($56 \times 56 \times 256 \rightarrow 56 \times 56 \times 256$) ?
- b) la troisième couche de "pooling" ($56 \times 56 \times 128 \rightarrow 28 \times 28 \times 512$) ?
- c) la première couche complètement connectée ($7 \times 7 \times 512 \rightarrow 4096$) ?
- d) la couche de "softmax" ($1000 \rightarrow 1000$) ?

Rappel : toutes les convolutions sont de taille 3×3 . Compter aussi les biais.

- e) Lesquels de ces nombres dépendent de la taille de l'image d'entrée ?

Question 6 : combien d'opérations flottantes sont-elles effectuées dans :

- a) la dernière couche de convolution ($14 \times 14 \times 512 \rightarrow 14 \times 14 \times 512$) ?
- b) la dernière couche complètement connectée ($4096 \rightarrow 1000$) ?

Question 7 : comment les effets de bord des convolutions sont-ils gérés dans ce réseau ?