Polytech' Grenoble – Info 4
Examen Accès et Recherche d'Information – 5 avril 2022 – 1 heure 30

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

Partie I

Recherche d'information multilingue

On va décrire un système de recherche d'information manipulant des documents écrits en français ($D._{fr}$) et des documents écrits en anglais ($D._{en}$).
Ce système repose sur les **4 étapes** suivantes :

1. Tokenisation basique sur la chaîne en minuscules, comme vu en TD, qui sépare sur les espaces, les apostrophes, les ponctuations, etc.
2. Détection de la langue : français(fr)/anglais(en)
3. Si langue == en, alors anti-dictionnaire anglais + troncature anglais + modèle vectoriel
4. Si langue == fr, alors anti-dictionnaire français + troncature français + modèle vectoriel

Nous voyons que ce système est en fait composé de deux systèmes de recherche d'information vus en cours, avec l'ajout de la détection de la langue du texte (document ou requête).

Dans la suite, nous allons nous baser les deux anti-dictionnaires suivants :
- pour l'anglais : {am, and, at, be, he, i, is, it, on, she, the}
- pour le français : {à, aux, de, es, est, le, les, la, suis, travers, un, une}

Question 1 : ***Décrivez*** une solution de détection de la langue qui se base uniquement sur les occurrences de mots de l'anti-dictionnaire anglais d'un côté, et de l'anti-dictionnaire français de l'autre. ***Expliquez*** les étapes de votre proposition, et la décision de détection, pour les 2 chaînes de caractères suivantes :
- "The cat is on the table and its looks at the mouse"
- "Le chien Labrador aboyait à travers la fenêtre de la maison"

Question 2 : Notre système de recherche d'information décrit plus haut manipule donc deux vocabulaires V_{en} et V_{fr} (un par langue), ainsi que deux index inversés (un par langue).
Quand on traite une requête, on commence par trouver la langue dans laquelle elle est écrite, puis on exécute les traitements sur "la bonne langue", avant de rechercher sur "le bon index inversé".

Considérons les vocabulaires (avec les « inverse document frequency », idf, fournis) anglais V_{en} et français V_{fr} suivants :

Vocabulaire anglais V_{en} :

Terme	idf
…	…
cat	0.3
labrador	10
look	0.4
mou	1.4
muscl	3.0
tabl	2.1
…	…

Vocabulaire français V_{fr} :

Terme	idf
…	…
aboi	0.4
chien	1.3
fenêtr	2.4
labrado	2.1
maison	0.6
tabl	0.5
…	…

Considérons de plus :
1. des poids de requête comme vus en cours et en TP : t^q_w = tf * idf
2. une troncature sur l'anglais qui utilise les règles suivantes :
   1_{en}: e → /
   2_{en}: s → /
   3_{en}: ies → y
3. une troncature sur le français qui utilise les règles suivantes :
   1_{fr}: e → /
   2_{fr}: s → /
   3_{fr}: ait → /
   4_{fr}: y → i
   5_{fr}: r → /

Soit la requête Q1 = "Le chien est un Labrador"
Détectez la langue de la requête Q1 en vous basant sur votre proposition pour la question 1.
Donnez 1) les termes de la requête (en analysant Q1 par les traitements classiques vus en cours et en TP), avec 2) leurs poids respectifs (rappel : t^q_w = tf * idf).

Question 3 :
Supposons maintenant une nouvelle requête, Q2, telle que :
Q2 = "Labrador"
Est-ce que votre étape de décision sur la langue de la question 1 fonctionne sur la requête Q2 ? ***Expliquez*** votre réponse.

Question 4 :
Dans le cas où la décision sur la langue n'est pas possible, on *modifie* les étapes décrites au début de l'exercice 2 en **ajoutant une nouvelle étape 5.**:

5. pour une requête, si le choix de langue est impossible, alors on traite la requête sur les deux langues et on interclasse toutes les documents suivant les valeurs de RSV des deux langues.

Suivant ce principe, ***calculez*** le résultat du système pour la requête Q2, "Labrador". Pour répondre à cette question, ***donnez*** la norme de la requête, les scores RSV, et la liste triée des résultats obtenus.
(toutes les informations dont vous avez besoin pour réaliser ces calculs sont fournies ci-dessous en (*) et (**)).

(*) Les lignes "utiles" des index inversés (les poids w_{id} fournis ci-dessous sont les "ptf*pdf*n", ils correspondent aussi aux t^d_w dans la partie 2 du cours. Ils sont utilisés dans le calcul de RSV) sont les suivantes (ces lignes sont complètes) :

- Anglais

Terme		Liste (d, w_{id}) associée à chaque terme
…		…
cat	→	($d1_{en}$, 5.2) ($d303_{en}$, 5.2) ($d310_{en}$, 4) ($d321_{en}$, 4)
labrador	→	($d1_{en}$, 5) ($d3_{en}$, 3) ($d4_{en}$, 5)
look	→	($d6_{en}$, 2.1) ($d3068_{en}$, 2.2)
mou	→	($d303_{en}$, 1.2) ($d3120_{en}$, 0.4)
muscl	→	($d1_{en}$, 3.2)
tabl	→	($d102_{en}$, 5.2) ($d301_{en}$, 5.2) ($d350_{en}$, 4) ($d500_{en}$, 4)
…		…

- Français

Terme		Liste (d, w_{id}) associée à chaque terme
…		…
aboi	→	($d1_{fr}$, 4.2) ($d405_{fr}$, 4.2) ($d512_{fr}$, 5) ($d722_{fr}$, 3)
chien	→	($d6_{fr}$, 5.2) ($d8_{fr}$, 2.2)
fenêtr	→	($d33_{fr}$, 2.1) ($d123_{fr}$, 1.4)
labrado	→	($d2_{fr}$, 3.2) ($d3_{fr}$, 2) ($d4_{fr}$, 9) ($d5_{fr}$, 6)
maison	→	($d1_{fr}$, 3.2)
tabl	→	($d1_{fr}$, 5.2) ($d124_{fr}$, 4)
…		…

(Note : les documents en français et en anglais sont totalement indépendants : **les documents $d1_{en}$ et $d1_{fr}$ n'ont rien en commun, à part leur indice**)

(**) Les normes $||d||=\sqrt{(t_w^d)^2}$ des documents sont :

- pour l'anglais :
$|| d1_{en} || = 65$
$|| d2_{en} || = 31$
$|| d3_{en} || = 22$
$|| d4_{en} || = 25$
$|| d5_{en} || = 51$
$|| d6_{en} || = 47$
…

- pour le français :
$|| d1_{fr} || = 27$
$|| d2_{fr} || = 32$
$|| d3_{fr} || = 41$
$|| d4_{fr} || = 53$
$|| d5_{fr} || = 19$
$|| d6_{fr} || = 18$
….

Question 5 :

Proposez et décrivez une solution pour la détection de la langue quand l'approche à base d'anti-dictionnaire seul n'est pas suffisante (cf. la question 3).

Partie 2

Les réponses doivent être justifiées de manière concise.

Exercice 1. Descripteurs histogrammes

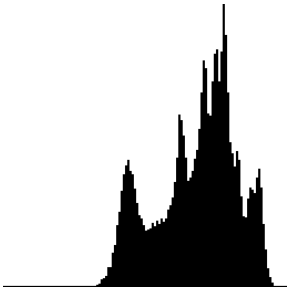

image 1

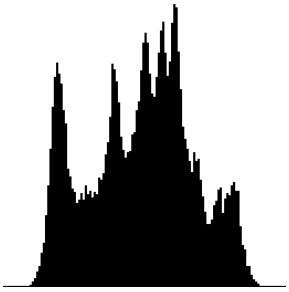

image 2

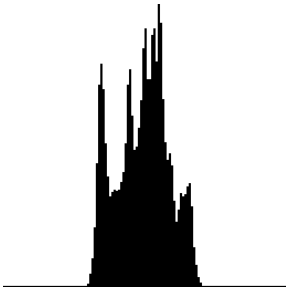

image 3


image 4

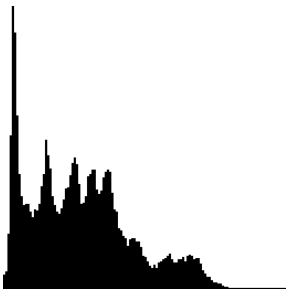

image 5


histogramme a


histogramme b


histogramme c


histogramme d


histogramme e

Question 1.1 : Associez les images ci-dessus avec leurs histogrammes, justifiez vos réponses. Les histogrammes sont monodimensionnels et portent sur l'intensité (luminance) de l'image.

Question 1.2 : Quelle est la dimension d'un descripteur d'image basé sur un histogramme de couleurs tridimensionnel dans l'espace LAB avec respectivement 8, 5 et 5 « bins » pour les composantes L, A et B ?

Question 1.3 : Quelles distances peuvent être utilisées pour évaluer la similarité de ces histogrammes dans le cadre d'une recherche par l'exemple ?

Exercice 2. Apprentissage profond pour la classification

On considère un réseau à propagation avant qui prend en entrée des images RGB de taille 64×64 et qui applique successivement les opérateurs suivants :

- un module de 64 filtres de convolution de taille 5×5 sans « padding » ;
- une rectification linéaire « ReLU »
- une réduction de taille par « max pooling » par blocs de taille 3×3
- un module de 16 filtres de convolution de taille 5×5 sans « padding » ;
- une rectification linéaire « ReLU »
- une réduction de taille par « max pooling » par blocs de taille 2×2
- un module linéaire avec 100 sorties
- une couche de « softmax »

Question 2.1 : Combien de couches neuronales ce réseau contient-il ?

Question 2.2 : Quel est le nombre total de plans à la sortie du premier et du second module de convolution ?

Question 2.3 : Quelle est la taille des images de sortie après chacune des deux modules de convolution ?

Question 2.4 : Combien de paramètres « apprenables » contiennent chacune des fonctions (ne pas oublier les biais) ?

Question 2.5 : Combien d'opérations sont effectuées dans le second module de convolution pour une image ?

Question 2.6 : Combien d'opérations sont effectuées dans le module linéaire pour une image ?

Question 2.7 : Quel est le but de la couche de « softmax » ?