Traitement Automatique de la Langue
Examen 2017, première session

1 heure 45 - aucun document autorisé - barème donné à titre indicatif

Exercice I - Classification de textes

Retour sur le problème de classification de locuteur (Chirac/Mitterrand).

Question 1

On hésite entre une représentation en sac de mots et une représentation en tri-grammes de lettres. Quels sont les avantages et inconvénients de chacune des représentations ? (par exemple, en terme de taille, de bruit généré, d'interprétabilité...)

Question 2

Etant donnée la nature particulière de ce problème, quel choix de représentation du texte feriez-vous et pourquoi ? Indiquez quelques pré-traitements qui vous semblent utiles et quelqu'uns que vous éviteriez ici.

Utiliseriez-vous la même représentation pour un problème de classification de sentiments (indépendamment de la langue visée) ?

Question 3

Donner la formulation d'un classifieur linéaire type SVM (à deux classes) utilisant un tel codage de l'information :

— quelles sont les dimensions des données en entrée ?
— combien de paramètres y a-t-il dans le modèle ? Y a-t-il des hyper-paramètres à donner lors de l'apprentissage ?
— comment interpréter les poids du modèle linéaire ?

Question 4

Le problème de l'équilibrage des classes était critique (environ 80% des échantillons de texte appartenant à l'une des classes). Comment faire face à cette problématique du point de vue des techniques d'apprentissage et du point de vue de l'évaluation des performances ?

Question 5

Un linguiste vous explique (tardivement) que l'un des locuteurs préfère les tournures verbales avec beaucoup d'adverbes tandis que l'autre privilégie des tournures nominales avec plus d'adjectifs...

Comment prendre en compte ces informations ? Dans le détail, quels outils utiliser pour extraire ce type d'information et comment les intégrer dans votre processus de classification ?

Exercice I (suite)

On dispose d'un corpus de 1000 documents représentés en sacs de mots sur un dictionnaire de 9000 termes et étiqueté en 10 classes :

Question 1

Est-il toujours possible de trouver un classifieur ne faisant aucune erreur sur ce corpus (en apprentissage donc) ?

Question 2

Le fait d'avoir plus de variables explicatives que d'échantillons entraine un risque : lequel ? Comment y faire face ? (a) lors de la formulation du problème d'apprentissage (b) lors de l'évaluation des résultats

Exercice II - Sémantique

Lisez l'ensemble des questions : vous pouvez proposer des réponses jointes si vous le souhaitez.

Question 1

Qu'est ce qu'un espace sémantique en TAL, à l'inverse, qu'est ce que le *semantic gap* ?

D'après cette définition, la matrice brute représentant un corpus de N documents en sacs de mots (dictionnaire de taille D) permet-elle de définir une sémantique ? Dans l'affirmative, comment ?

Question 2

Nous appliquons maintenant l'algorithme Word2Vec (W2V) sur le précédent corpus.

1. Quels sont les hyper-paramètres à fixer ?
2. Qu'est ce qui distingue l'algorithme d'apprentissage W2V de PLSA (ou LDA) ?
3. Quelle est l'intérêt de la sémantique obtenue par rapport à un algorithme comme PLSA (ou LDA) ?

Latent Dirichlet Analysis. Vecteurs.

Exercice III - Construction d'un système industriel

Un industriel vous demande de créer un système de classification automatique de documents pour structurer un corpus de centaines de milliers de documents. D'après l'expert, il y a 10 classes de documents. Cet expert est capable de vous donner une dizaine de mots clés associés à chaque classe, mais il ne dispose pas de documents étiquetés. Par contre, il est prêt à mettre à votre disposition une personne chargée de corriger/valider les sorties du systèmes quelques heures par semaine durant la phase de réglage du système.

Question 1

Comment procéder ?

Bootstrapping

Question 2

Comment optimiser le travail de la personne mise à votre disposition ?

Question 3

Si on réfléchit bien, il manque des informations concernant l'usage que feront les clients de notre système... Quelles sont les bonnes questions à poser pour développer un système utile ? Quelles pourraient être les solutions techniques pour rendre votre système efficace ?

Exercice IV - Extraction d'information

On considère le texte suivant :

> Dr House (House, M.D., puis House) est une série télévisée américaine en 177 épisodes de 43 minutes et répartis sur huit saisons. Elle a été créée par David Shore et sa diffusion s'est déroulée du 16 novembre 2004 au 21 mai 2012 sur le réseau Fox.
> (Wikipédia)

ORG. PERS

1. Citez trois entités nommées contenues dans ce texte et leurs catégories.
2. En vous appuyant sur un exemple du texte, présentez un exemple de difficulté d'annotation des entités nommées pour un système d'extraction d'information.
3. Si l'on décide d'utiliser les métriques de rappel et de précision classiques de la recherche d'information, quels problèmes se posent lors de l'évaluation de la sortie d'un système de reconnaissance d'entités nommées ? Comment adapter ces métriques à cette tâche ?
4. Citez deux critères qui peuvent être utilisés pour désambiguïser une entité nommée (par exemple pour déterminer que « Fox » se rapporte à une organisation et non pas par exemple à l'actrice Megan Fox).
5. Donnez deux types d'attributs utilisés en extraction de relations supervisée, permettant par exemple de reconnaître une relation de type oeuvre-créateur telle que celle entre « Dr House » et « David Shore ».
6. Quel est le principe de la supervision distante ?

< PERS ??

(Bootstrapping)

Exercice V - Similarité thématique de textes

Soient les deux textes suivants appartenant tous deux au même domaine, les élections politiques :

Texte1 :	Après l'appel du Front national à " faire battre " la candidate de la majorité au second tour de la législative partielle de la troisième circonscription de l'Orne, le secrétaire général du RPR, Jean-François Mancel, est allé soutenir, mercredi 27 mars, Sylvia Bassot (UDF-PR). Il espère que " les mots d'ordre donnés par les uns et par les autres n'auront que peu d'impact sur le choix des électeurs ".
Texte2 :	Créant une surprise, car nul sondage n'avait su le prévoir, M. Le Pen avait obtenu, à l'élection présidentielle de 1988, 14,39% des suffrages exprimés. Sa "percée" était encore récente : commencée aux élections municipales de 1983, elle avait été confirmée par les européennes de 1984 et par les législatives de 1986, qui lui avaient permis de faire élire 35 députés. Les résultats des élections depuis 1988 montrent que le FN dispose d'un noyau d'électeurs oscillant entre 12% et 14%. Elections législatives de 1988 : 9,65%, 1 élu ; cantonales de 1988 : 5,24% ; municipales de 1989 : 2,17%, 804 élus ; européennes de 1989 : 11,73%, 10 élus ; régionales de 1992 : 13,90%, 239 élus ; cantonales de 1992 : 12,18%, 1 élu ; législatives de 1993 : 12,52% ; cantonales de 1994 : 9,78%, 3 élus ; européennes 1994 : 10,52%, 11 élus.

Lorsqu'on ne garde que les lemmes et qu'on pondère ceux-ci par la formule tf.idf calculée par rapport à une collection de 110 textes, on obtient les deux listes suivantes :

```
                        2.88 général
0.09 ,                  1.01 il
0.02 .                  4.67 impact
2.96 "                  3.57 Jean-François
0.96 (                  0.00 le
0.96 )                  3.29 législatif
0.18 à                  3.06 majorité
2.88 aller              4.67 Mancel
3.06 appel              2.03 mars
1.96 après              2.37 mercredi
0.40 au                 4.67 mot
2.19 autre              2.48 national
0.20 avoir              0.91 ne
4.67 Bassot             3.29 ordre
2.88 battre             4.67 orne
3.29 candidat           1.22 par
0.13 @card@             3.29 partiel
4.67 choix              3.57 peu
4.67 circonscription    1.48 que
0.00 de                 3.57 RPR
3.06 donner             3.98 second
0.09 du                 3.29 secrétaire
3.57 électeur           3.98 soutenir
3.06 espérer            0.72 sur
0.29 et                 4.67 Sylvia
0.24 être               2.59 tour
1.63 faire              2.73 troisième
3.98 front              4.67 UDF-PR
                        0.16 un
```

```
0.25 ,                  1.63 faire
24.51 ;                 3.98 FN
13.41 :                 0.00 le
0.04 .                  9.86 législatif
1.48 "                  2.48 lui
0.18 à                  2.48 monsieur
0.40 au                 3.06 montrer
0.78 avoir              7.96 municipal
14.02 cantonal          0.91 ne
3.06 car                4.67 noyau
4.40 @card@             4.67 nul
3.06 commencer          2.88 obtenir
3.29 confirmer          4.67 osciller
4.67 créer              1.22 par
0.00 de                 3.98 Pen
1.90 depuis             4.67 percée
3.57 député             3.06 permettre
4.67 dispos             3.57 présidentiel
0.06 du                 2.03 prévoir
3.57 électeur           0.74 que
9.19 élection           0.91 qui
4.67 Elections          3.98 récent
37.38 élire             4.67 régional
2.27 elle               3.57 résultat
2.48 encore             3.57 savoir
2.11 entre              0.82 son
0.58 et                 3.98 sondage
0.48 être               3.98 suffrage
7.78 européen           4.67 surprise
3.98 exprimer           0.32 un
```

Question 1

Lorsqu'on calcule leur similarité par la formule du cosinus, on obtient une valeur de similarité très basse.

1. Indiquer pourquoi. Sur quels termes repose la similarité obtenue ?
2. Indiquer différents traitements qui permettraient de construire des représentations de ces deux textes rendant mieux compte de leur sens et conduisant à une plus grande similarité. Citez en au moins trois.

Question 2

Questions de cours

1. Qu'est-ce que l'implication textuelle ? Comment s'applique-t-elle dans le contexte de systèmes de question-réponse ?
2. Quelles caractéristiques permettent de mesurer la cohésion lexicale d'un passage comportant plusieurs phrases ? Citez en au moins trois.