Fouille de texte **Nom :** _____
2018/2019 **Master 2 ATAL**
07.11.18 **Durée : 1 heure**

- **Notes et supports de cours autorisés.**
- **Calculatrice autorisée, téléphone interdit.**
- Justifier vos calculs en détaillant les étapes. Un résultat sans détails ne donnera aucun point.
- La première partie de l'examen (Q1-3) concerne tout ce qui a été vu et exploité en cours.
- La seconde partie de l'examen (Q4) concerne la séance d'exposés et est composée de questions bonus. Privilégiez à cette section des réponses brèves détaillant les intuitions ou des exemples.

Exo	Points	Score
1	5	
2	9	
3	8	
4	3	
Total:	25	

1. **Représentation vectorielle des documents**

   (a) (2 points) Soit le document suivant :

   ```
   J'aime la fouille de texte.
   La fouille de texte est devenue ma passion.
   Donnez moi des textes et je me ferai un plaisir de les fouiller.
   ```

   Détailler votre approche pour le transformer en un vecteur utilisable par les algorithmes d'apprentissage automatique. Ne négligez pas d'appliquer des pré-traitements : précisez lesquels.

   (b) (2 points) Calculer le *tf-idf* des mots *texte* et *plaisir* pour le document précédent. On considère que ce document fait partie d'un corpus dans lequel *texte* apparaît dans 52 documents, et *plaisir* dans 9 documents. Quel mot semble le plus important pour décrire le document dans ce corpus ?

   (c) (1 point) Quel est le nom de la méthode qui applique une réduction de dimension par Décomposition en valeurs singulières (SVD) à une matrice documents-termes ?

2. **Plongements lexicaux**

   Soit le document suivant :

   ```
   Avant de savoir comment transporter la congolexicomatisation des lois du
   marché, il faut plutôt savoir ce qu'on va transporter. La congolexicomatisatio
   des lois du marché seule, dans son plus simple appareil avec seulement
   son contenu théorique, peut tenir sans problème dans un fichier PDF.
   ```

   (a) (2 points) Nous souhaitons connaître le sens de *congolexicomatisation*. Pour cela, nous nous basons sur l'hypothèse de Firth : "You shall know a word by the company it keeps."

Calculer la matrice de co-occurrence du document suivant en utilisant une fenêtre de taille 2. Utilisez les termes suivants dans la matrice : *congolexicomatisation*, *transporter*, *loi*, *marché*, *simple*, *appareil*.

(b) (3 points) On souhaite appliquer la Positive Mutual Information (PMI) sur cette matrice. À quoi sert ce calcul ? Calculer les vecteurs PMI de *congolexicomatisation* et *marché*.

(c) (2 points) Avec quelle métrique comparer les vecteurs lignes de cette nouvelle matrice ? Définir cette métrique et donner le calcul de similarité de *congolexicomatisation* et *marché*.

(d) (2 points) Quel est le principal défaut des vecteurs que nous manipulons jusque là ? Quelle méthode appliquer sur nos vecteurs *PMI* pour obtenir des vecteurs qui seront mieux utilisés par les algorithmes d'apprentissage ?

3. **Clustering**

(a) (2 points) Soit la matrice documents-termes suivante :

$$\begin{pmatrix} 0 & 5 \\ 1 & 4 \\ 4 & 4 \\ 2 & 1 \\ 3 & 0 \end{pmatrix}$$

Choisissez une initialisation *aléatoire* de deux centroïdes puis itérer l'algorithme k-moyennes jusqu'à convergence, ou au maximum quatre étapes (une étape = affectation des données + réaffectation des centroïdes).

(b) (2 points) Que peut-on utiliser pour interpréter les clusters obtenus ? Donner une méthode utilisant directement un *output* de l'algorithme des k-moyennes, et proposer une approche appliquée *a posteriori* sur les clusters.

(c) (2 points) Sur ces données, comment calculer la qualité de la partition résultant du clustering ? Expliquer les concepts de critère interne, critère externe.

(d) (2 points) Soit les données de la Figure 1. Quel algorithme vaut-il mieux appliquer : DBScan ou K-moyennes ? Détailler les différences entre ces algorithmes.

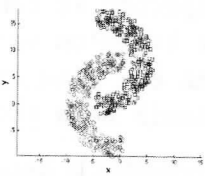

Figure 1: Des données...

4. (3 points) **Questions bonus : Exposés (0,5 pt par question)**
   (a) Que signifie supervision distante ? Dans quel cas cela peut-il être utile ? Citer un exemple.
   (b) Que signifie la phrase : le graphe de co-occurrence du langage est un *petit-monde* ?
   (c) Quel est l'intérêt des Convolutional Neural Networks pour la classification de texte ?
   (d) Quelle métrique est utilisée dans le cadre de l'évaluation d'un résumé automatique. Expliquer brièvement son fonctionnement.
   (e) Expliquer brièvement le fonctionnement des *plongements lexicaux* (word embeddings) dans le cadre diachronique.
   (f) Donner un exemple d'application des *plongements lexicaux* (word embeddings) dans le cadre diachronique.