Fouille de données, Fouille de textes
2018/2019
02.07.19

Nom : _____

Master 2
Durée : 40mn

- **Notes et supports de cours autorisés.**
- **Calculatrice autorisée, téléphone interdit.**
- Justifier vos calculs en détaillant les étapes. Un résultat sans détails ni explication ne donnera aucun point.

Exo	Points	Score
1	6	
2	7	
3	7	
Total:	20	

1. **Pré-traiter et représenter les documents**
   (a) (1 point) Qu'appelle-t-on *mots vides* ? Quel est l'intérêt de disposer d'une liste de mots vides ?
   (b) (2 points) Qu'est ce que la lemmatisation ? Quel en est l'objectif ?
   (c) (3 points) Donner une méthode pour indexer (i.e. obtenir les mots clés) les documents des corpus. Détailler son fonctionnement.

2. **Plongements lexicaux : *word embeddings***
   (a) (1 point) Qu'est ce qu'un vecteur *creux* ?
   (b) (2 points) Qu'utilise-t-on comme distance pour comparer deux vecteurs dans le cadre du traitement de la langue ? Expliquer son utilisation.
   (c) (2 points) À quoi sert l'apprentissage de word embeddings ?
   (d) (2 points) Expliquer cette formule qui est celle du modèle de Word2vec :
   $$p\left(w_O|v_I\right) = \frac{\exp\left({v'_{w_O}}^{\top} v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left({v'_w}^{\top} w_I\right)}$$

3. **Clustering**
   (a) (3 points) Expliquer ce qui se passe au fur et à mesure du déroulement de l'algorithme des k-moyennes avec une initialisation des centroïdes telle que celle de la Figure 1. Cette initialisation permet-elle de découvrir correctement les trois groupes ?
   (b) (2 points) L'initialisation aléatoire des centroïdes fait que l'algorithmes des k-moyennes n'est pas déterministe, i.e. les résultats peuvent être différents d'une exécution à l'autre. Quel protocole peut être mis en place pour faire le clustering d'un jeu de données en tenant compte de cet aspect ?
   (c) (2 points) Après avoir réalisé le clustering d'un ensemble de documents textuels, qu'est il utile de faire pour faciliter l'interprétation des résultats, i.e. décrire la contenu des clusters ?

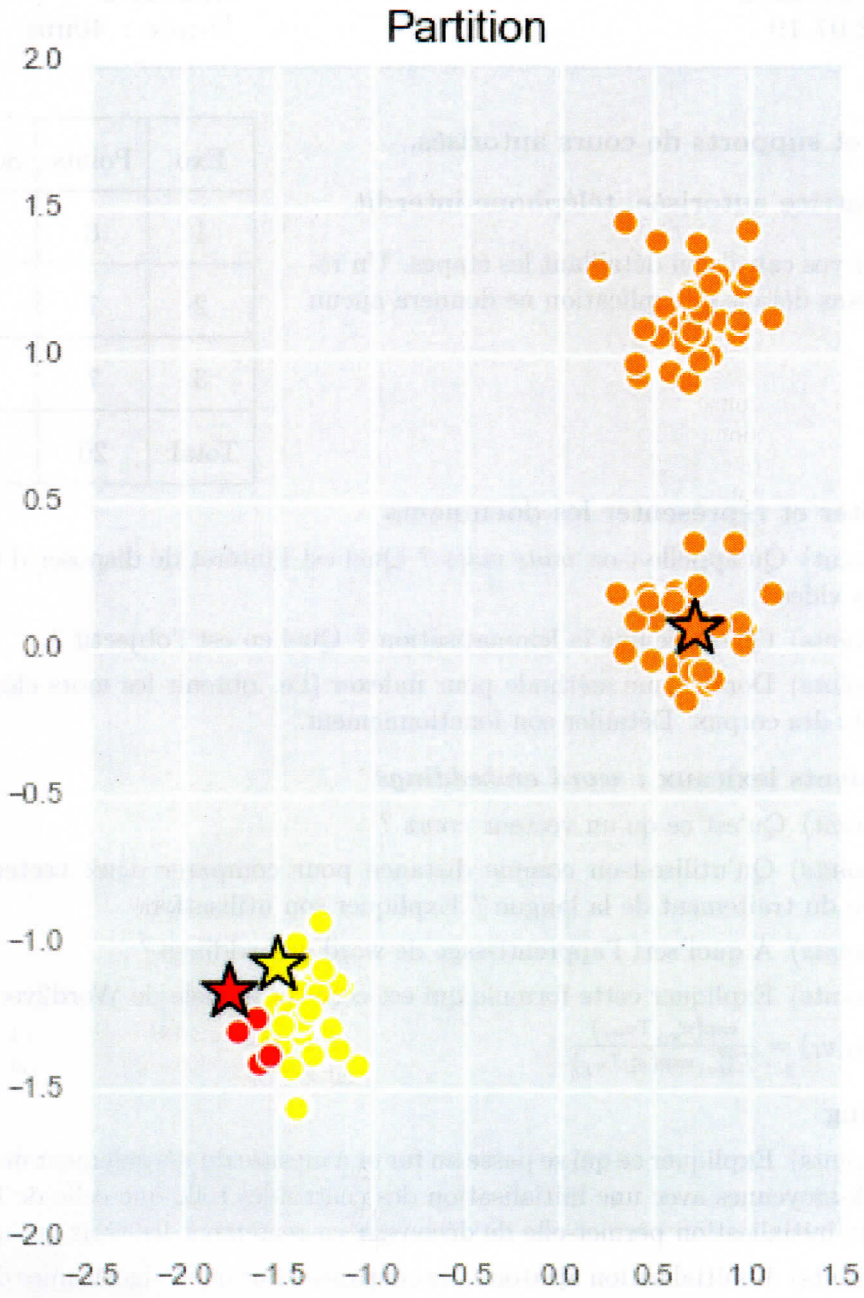

Figure 1: Les cercles pleins sont des données, et les étoiles des centroïdes.