Analyse de sentiments d'un corpus de « tweets » comportant toxicité.

Dominique Loyer et Serigne Saliou Mbacke Diakhate
supervisé par Richard Khoury

Université Laval, 15 février 2017

1- Les données

But :

Nous allons :

- Discuter sur les données, les types d'attributs, et les propriétés statistiques des données. Quelles difficultés présentent-elles ? (ex.: bruit, fléau de dimensionnalité, informations manquantes, déséquilibre des classes, etc.).

- Implanter des algorithmes de prétraitement des données afin de corriger ces difficultés.

- Décrire les algorithmes et leurs résultats.

- Discuter comment les différents jeux de données vont interagir dans notre système. Comment fonctionne la liaison des données en fonction des valeurs qu'ils ont en commun ? Comment allons-nous utiliser les données ? Comment allons-nous structurer les données, les stocker et les charger en mémoire?

-
- Et discutez également de la procédure de tests que nous envisageons de faire.

Le jeu de données (attacks_data) contient 6 attributs (Date, Country, City, Killed, Injured et Description) et 29 364 enregistrements décrit comme suit pour chaque enregistrement :

- Date : étant un attribut de type temps (date (AAAA-MM-JJ)) qui correspond à la date exacte de l'attentat.

- Country et City : correspondant aux attributs de type lieu (Varchar) qui spécifient dans quel pays et ville s'est passé l'attentat.

- Killed : étant un attribut de type nombre positif (SMALLINT [M] [UNSIGNED]) qui donne le nombre de mort dans un attentat.

- Injured : étant un attribut de type nombre positif (SMALLINT [M] [UNSIGNED]) qui donne le nombre de blessées dans un attentat.

- Description : étant un attribut de type texte (TEXT) qui décrit comment l'attentat c'est passé.

Voici quelques statistiques sur le jeu de données.

10 attentats au Canada dont le plus meurtrie est l'attentat 30 juin 2009 à Kingston faisant 4 morts. L'attentat le plus meurtrie est celui du Nigeria à Baga faisant 2000 morts survenu le 7 janvier 2015. L'Irak est représenté par près du tiers des crimes avec 8701 crimes suivi par le Pakistan avec 4494 crimes. On voit d'ailleurs dans ce graphique la distribution des crimes par importance en commençant par les villes du Pakistan (avec Lahore, 729 crimes en guise d'exemple)

Bruit :
Nous n'avons pas constaté de bruit dans notre enregistrement mais juste quelques remarques :

Nous constatons dans l'enregistrement 16400 la description manque, et aussi que dans le champ pays la Palestine est représentée comme suit Pal.Auth.

Aussi nous constatons dans l'enregistrement 16994 La ville n'est pas spécifiée. Nous avons juste un attentat survenu le 7 février 2012 à Palestine et que comme nous n'avons pas pu trouver d'informations sur le net à propos de cet attentat à Palestine nous avons décidé de l'enlever.

D'après notre observation les remarques que nous avons citées ci-dessus ne vont pas influencer dans le traitement de nos données.

Nous avons ci-dessous deux figures (1) et (2)

Figure(1) : la première figure représente le nombre d'attentat qu'on enregistre pour une ville donnée et d'après cette figure nous constatons que c'est dans la ville de Palu que nous enregistrons le plus d'attentat.

Figure(2) : la deuxième figure représente le nombre de personnes tuées dans chaque grande ville canadienne suite à un attentat. et nous constatons que c'est la ville de Kingston qu'on enregistre le plus de tués.

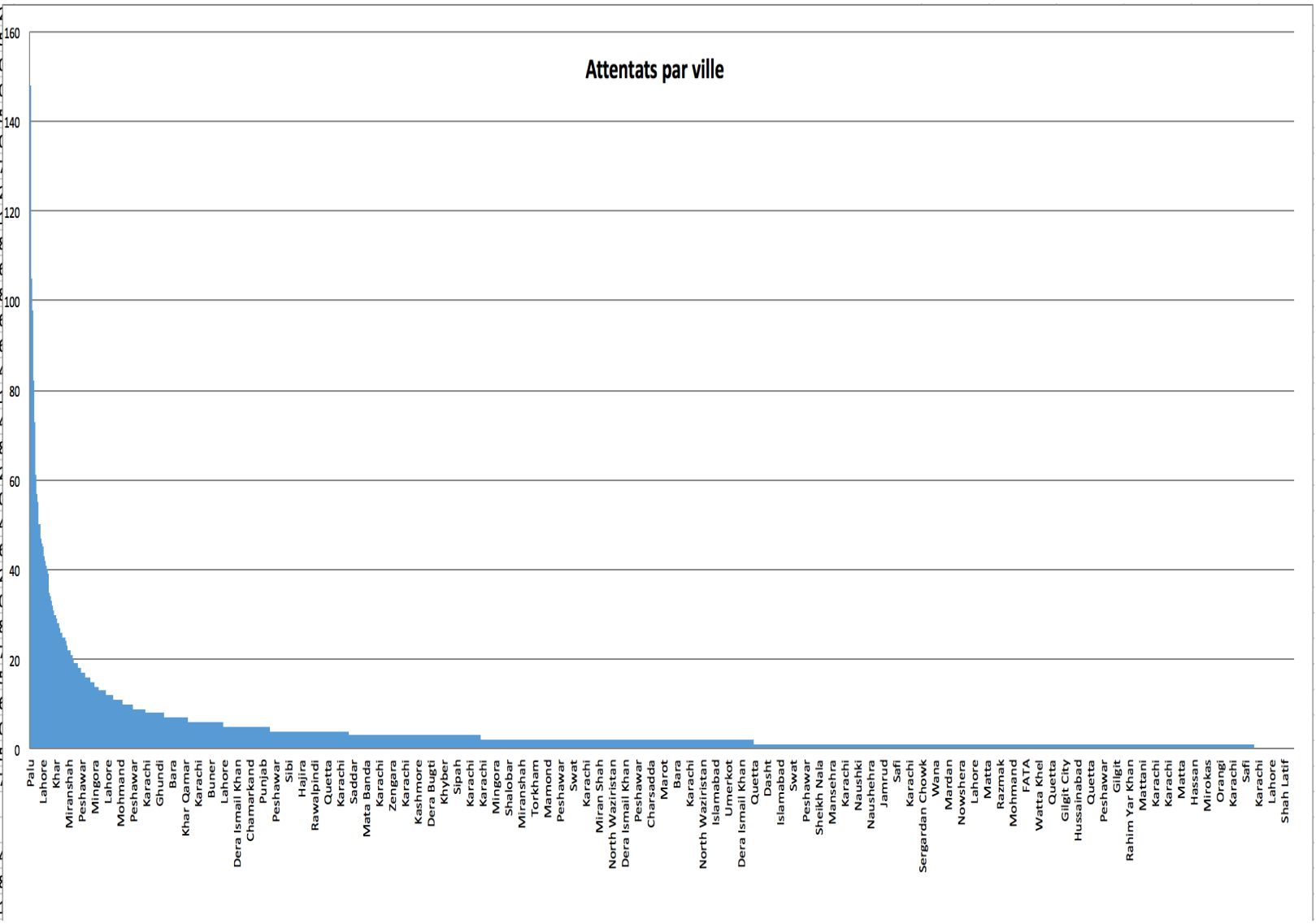
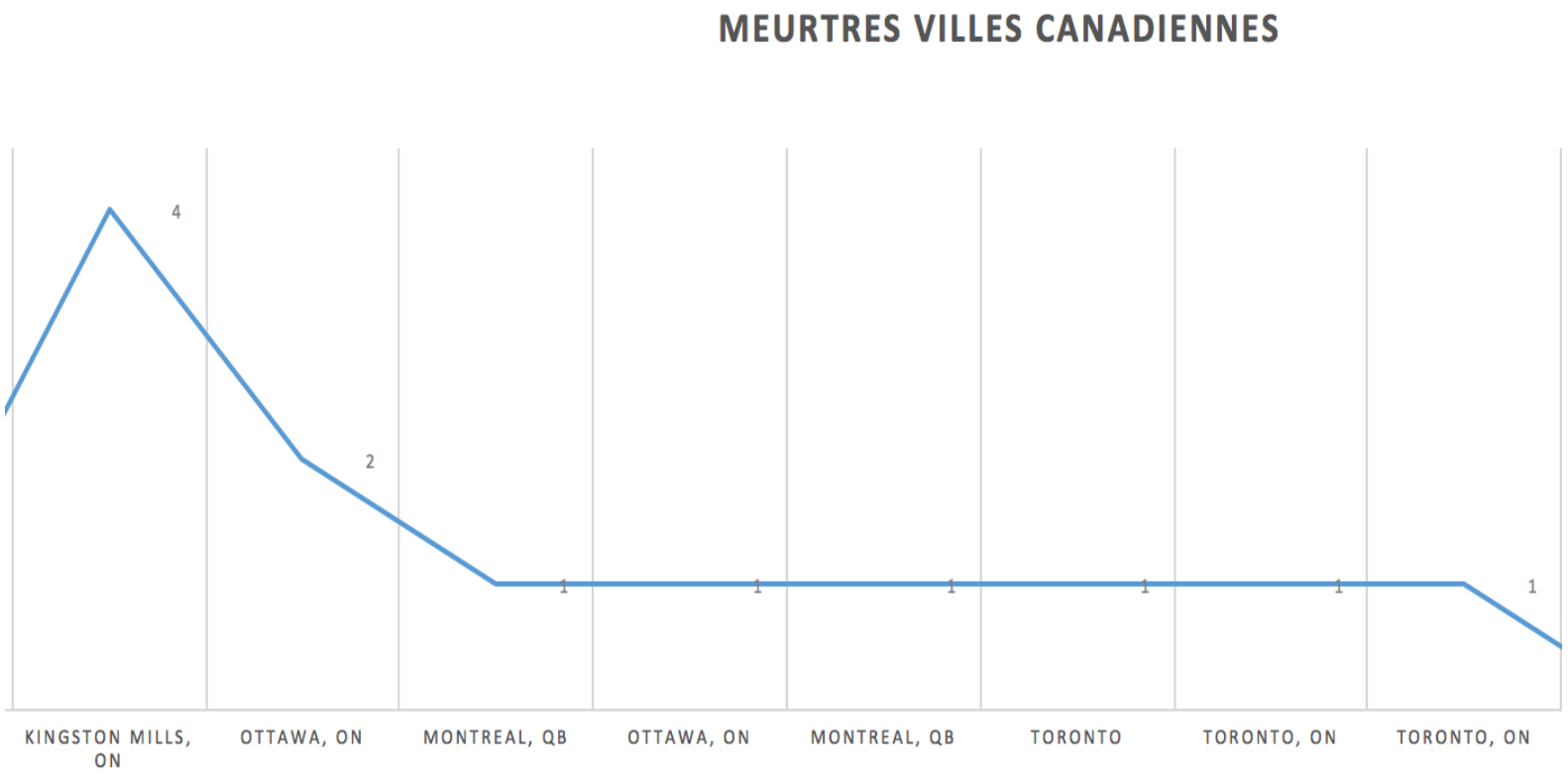


Figure (1) : Nombre attentat par ville



Figure(2) : Nombre de meurtres dans chaque grande ville canadienne suite à un attentat.

La difficulté principale qui se présentent est que le concours est maintenant fermé et qu'on travaille en vase clos, c'est-à-dire qu'on ne peut tester les données et avoir de la rétroaction. Toutefois, compte-tenu qu'il y a très peu de bruit, nous pouvons dégager des connaissances de ces données et faire des inférences statistiques.