PJE : Analyse de comportement avec Twitter

Annotation de tweets par mots clés

Objectif : afin de pouvoir classifier de nouveaux tweets, il faut constituer une base d'apprentissage contenant des tweets déjà étiquetés comme positif, négatif ou neutre.

Prérequis :un fichier CSV contenant une liste de tweets français sur les sujets suivant Vlille, Equipe de France, Election présidentielle

Format du fichier CSV : id; utilisateur; tweet; Date; Requete; Polarité (classe -1 non annoté, 0 négatif, 2 neutre, 4 positif);

Nettoyage de la base d'apprentissage

Dans cette partie, nous allons préparer les données. A partir du fichier CSV contenant les tweets récupérés par votre API vous allez créer une application qui crée un nouveau fichier avec la même structure CSV en réalisant les actions suivantes :

- Nettoyer les données (enlever les @, #, RT, URL et l'URL associée, "") voir slide de cours
- Enlever les tweets qui comportent des émoticônes positifs et négatifs en MEME temps
- Enlever les tweets redondants (i.e. qui sont dupliqués)
- Vérifier que les tweets récupérés sont plutôt français

Vous utiliserez pour cela les expressions régulières et les trois Java :

- Pattern : permet d'obtenir une version compilée d'une expression régulière.
- Matcher : permet d'analyser une chaîne en entrée à partir d'un Pattern.
- PatternSyntaxException : exception levée lorsque la syntaxe d'une expression régulière n'est pas correcte.

Vous vous servirez de l'aide suivante pour créer votre nettoyeur de tweets :

http://imss-www.upmf-grenoble.fr/prevert/Prog/Java/CoursJava/expressionsRegulieres.html

Annotation automatique de tweets pour créer la base d'apprentissage

Le champ polarité de votre CSV est pour le moment à -1 pour tous les tweets car aucun tweet n'est annoté. Afin de créer la base d'apprentissage pour classifier de nouveaux tweets, nous pouvons les annoter à la main. Un volontaire ???

Nous pouvons aussi essayer une première méthode assez naïve qui consiste à regarder les corpus de mots utilisés dans les tweets et voir si les mots du tweet sont issus de corpus plutôt positif ou plutôt

négatif. Pour cela nous avons établi des corpus de mots positifs et négatifs à partir du site http://twitrratr.com/ qui a été complété avec des termes français. Vous trouverez ce fichier dans http://www.fil.univ-lille1.fr/~jourdan/PJE/keywords/

Un tweet sera considéré positif s'il contient plus de mots du corpus positifs que négatifs. De la même façon il sera considéré négatif s'il contient plus de mots du corpus négatifs que positifs. S'il ne contient pas de mots positifs ou négatifs ou autant de termes positifs que négatifs, il sera neutre. Vous utiliserez la grammaire sur la polarité 0 négatif, 2 neutre, 4 positif pour annoter les tweets.

Réaliser une application qui à partir de votre fichier CSV nettoyé génère un nouveau fichier CSV qui contient les tweets annotés.