|  |
| --- |
|  |
| Sentiment analysis |
| Base de données avancée |
|  |
| **Mathieu HENTGES**  **Lucien LOPEZ**  **Christian WANG** |
| **03/06/2014** |

|  |
| --- |
|  |

# Introduction

Pour le projet de base de données avancée, nous avons décidé d’étudier l’analyse des sentiments, dans le cadre des sites internet Deezer et Tweeter. Pendant les dernières années, les médias sociaux se sont développés très rapidement, favorisant avec grande simplicité les interactions sociales à travers le monde. Ils permettent aux personnes de communiquer, partager des documents, ou encore des musiques.

Ces médias sont devenus populaires auprès de la population, mais aussi auprès des entreprises. En effet, ils regorgent d’informations de tout type, pouvant être exploité à des fins commerciales. La discipline « sentiment analysis » permet de détecter les opinions des personnes, d’analyser leurs comportements et habitudes pour, au final, proposer des solutions telle que la recommandation de produit. De ce fait, de nombreux systèmes automatisés ont été conçus afin d’analyser toutes ces données.

Pour notre projet, nous avons choisi de créer une interface permettant de voir le pourcentage des genres écoutés sur Deezer dans les pays, mais aussi de compter les tweets ayant dans leur hashtag l’un des trois genres les plus écoutés.

Dans un premier temps, ce rapport présentera rapidement l’état de l’art de l’analyse des sentiments. Puis il présentera l’interface réalisé ainsi que ses objectifs.

# L’analyse des sentiments

## Principe

Tout d’abord, une opinion est un « jugement, avis, sentiment qu’un individu ou un groupe émet sur un sujet, des faits, ce qu’il en pense » (Larousse). C’est un jugement personnel que l’on porte sur une question, et qui n’est pas forcément juste.L’analyse des sentiments permet de suivre l’opinion, l’humeur du public à propos d’un sujet, d’un produit en particulier en analysant des textes, commentaires, messages. Cette discipline a été appliqué sous diverses formes à travers l’histoire ; en effet, connaître l’opinion des autres a toujours été important avant d’émettre une décision, et ce quel que soit la période. Cependant, le principal défaut résidait dans le nombre d’avis que l’on pouvait obtenir.

C’est dans les années 2000 que ce sujet a marqué un grand pas. Avec l’avènement d’Internet, du web participatif et plus particulièrement des médias sociaux, il est maintenant possible de connaître l’opinion des personnes réparties dans plusieurs pays. La quantité de données à traiter étant colossale, il faut, pour cela, construire un système permettant de collecter ces données, de les examiner afin d’en sortir des informations pertinentes, et de les mettre à jour si nécessaire.

L’analyse des sentiments présente plusieurs défis à relever. Premièrement, la signification d’un mot peut dépendre du contexte ; il peut très bien exprimer quelque chose de positif dans une phrase, et quelque chose de négatif dans une autre. De plus, il faut prendre en compte le fait qu’une personne exprime plusieurs opinions dans une même phrase. Cela devient compliqué à analyser pour un programme informatique.

## Méthodes d’analyse

En général, les différentes méthodes proposent une même première approche. Elles cherchent d’abord à classe un texte dans une catégorie. Ceci peut être fait en analysant la nature des mots ou des phrases qui le composent ; en effet, un vocabulaire précis est employé lorsqu’on exprime une opinion sur un sujet. Il existe trois méthodes principalement utilisées, la méthode symbolique, la méthode statistique, et la méthode hybride.

### La méthode symbolique

Pour cette méthode, l’analyse des sentiments se fait au niveau des phrases. Elle traite d’abord un texte phrase par phrase, et extrait de celles-ci des informations sous forme de relations. Cette extraction se base sur un lexique de mots ayant un rapport avec le texte ; celui-ci classe les mots selon leur domaine d’utilisation (ex : mots liés au domaine du cinéma) mais aussi selon leur polarité (positive, négative, neutre).

Il y a plusieurs types de relations. La première représente les relations de base. Ce sont, par exemple les modifieurs d’un nom (ex : une jolie fleur) ou d’un verbe (ex : écouter attentivement). La deuxième représente les relations complexes telles que le sujet d’un verbe.

La méthode regarde si ces relations sont des relations de sentiments. Ce sont les relations comprenant des mots exprimant des sentiments positifs, négatifs. Pour les sentiments neutres, la méthode calcul d’après la construction de la phrase (coordination d’un sentiment positif et d’un sentiment négatif, présence de termes telles que « malgré », « pourtant »).

Au final, pour connaître l’opinion du texte, le nombre de sentiments positifs, négatifs et neutre est retenu pour chaque phrase. A la fin du texte, durant la synthèse, la somme des valeurs est calculée, permettant de donne une valeur globale du sentiment du texte.

Le principal défaut de cette méthode est qu’elle est mal adaptée pour des analyses globales de texte. Elle semble plus adaptée à des phrases, ainsi que des messages courts. Cette méthode peut être particulièrement intéressante si on veut analyser, par exemple, des tweets. Elle nécessite aussi différents lexiques suivant le sujet qu’elle traite.

### La méthode statistique

La méthode statistique est basée sur les techniques de l’apprentissage automatique. Elle consiste à utiliser un corpus de données déjà traité comme modèle dans le but de le faire apprendre au système.

La première étape de cette méthode est la catégorisation d’un texte. Cette phase peut se dérouler de différentes manières, par exemple, en classant les mots du texte selon des caractéristiques prédéfinies (fréquence des mots, sens, nature, etc.).

On crée ensuite un corpus d’entrainement avec ces données, qui sera traité par un algorithme d’apprentissage automatique et qui va créer des modèles se basant dessus.

Ensuite, on fait traiter un nouveau lot de données de la même catégorie que celle de départ. Ceci permettra d’améliorer le modèle crée et de le compléter avec de nouvelles informations.

Il existe de nombreux classificateurs utilisant des algorithmes différents, tels que :

* SVM (Support Vector Machine)
* Classificateur naïve bayésienne.

La méthode statistique peut s’appliquer à tous les domaines. Elle permet, contrairement à la méthode symbolique, d’analyser l’opinion d’un texte en général. Elle possède néanmoins des défauts. Elle nécessite un grand corpus d’entrainement afin de s’améliorer. De plus, chaque modèle crée est réservé à un domaine spécifique. Un modèle basé sur un jeu de données de commentaires en cuisine ne donnera pas de bon résultat pour un jeu de données de commentaires en automobile.

### La méthode hybride

Les méthodes symbolique et statistique présentes des limitations. L’intérêt des méthodes hybrides est de combiner les points forts des méthodes vues précédemment. Elle utilise la méthode symbolique pour la catégorisation d’un texte, afin de relever les phrases contenant des sentiments, avant de lancer le processus d’apprentissage comme dans la méthode statistique.