Groupe: Jonas SEBAONI, Gary PEREZ

Projet R

« Mini-système de recommandation »

I) Introduction

Le projet consiste à construire un mini-système de recommandation avec le logiciel R. Afin d'arriver au résultat escompté les données passent par plusieurs étapes. Dans ce rapport nous proposons la démarche que nous avons utilisée afin de réaliser notre mini-système de recommandation.

II) Réalisation projet

a) Choix des données

Afin de pouvoir tester notre futur mini-système de recommandation nous avons dû choisir un dataset. Nous avons décidé de nous orienter sur un petit dataset d'Amazon d'approximativement 10 000 reviews. Celui-ci est téléchargeable à l'adresse suivante: http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Digital_Music_5.json.gz

Ce dataset brut contient un certain nombre de champs qui sont:

- reviewerID,
- asin
- reviewerName
- helpful
- reviewText
- overall
- summary
- unixReviewTime
- reviewTime.

2) Echantillonnage

Pour éviter tout problème de calcul, nous avons limité le nombre d'échantillons pour l'entrainement de notre système à 8 000 reviews. Le choix de la sélection des reviews se fait de façon aléatoire.

3) Mise en forme + « tidytext »

Les données brutes du dataset ne peuvent pas en elle-même être directement utilisées. C'est pourquoi vient une étape qui consiste à mettre en forme le dataset pour qu'il soit plus exploitable et ne garde que les champs utiles à l'analyse sentimentale qui sera réalisée par la suite.

Cette mise en forme consiste à se débarrasser d'une partie des données que l'on considère ne pas être utile à l'exploitation et à l'entrainement de notre système de recommandation.

Voici les champs que nous avons retenus pour réaliser l'analyse sentimentale:

- reviewerID
- asin
- reviewText
- overall

4) Analyse de sentiment + scoring

L'analyse sentimentale permet de déterminer dans une phrase si le message peut être perçu comme étant positif ou négatif. Nous avons choisi de nous orienter sur deux types d'analyse qui sont:

- BING
- NRC

Ces deux types d'analyse ont l'avantage d'avoir été entrainé sur un large dataset et donc d'offrir des résultats qui nous ont semblé plus réalistes comparés à d'autre types.

Pour déterminer le côté sentimental dans le lexique nous utilisons un système de notation. Pour réaliser la notation, ce système ajoute 1 point par mot considéré positif et enlève 1 point si le mot est considéré négatif. Il réalise enfin

la somme de l'ensemble des points et en déduit si la phrase est plutôt positive ou négative.

5) Résultats de l'analyse

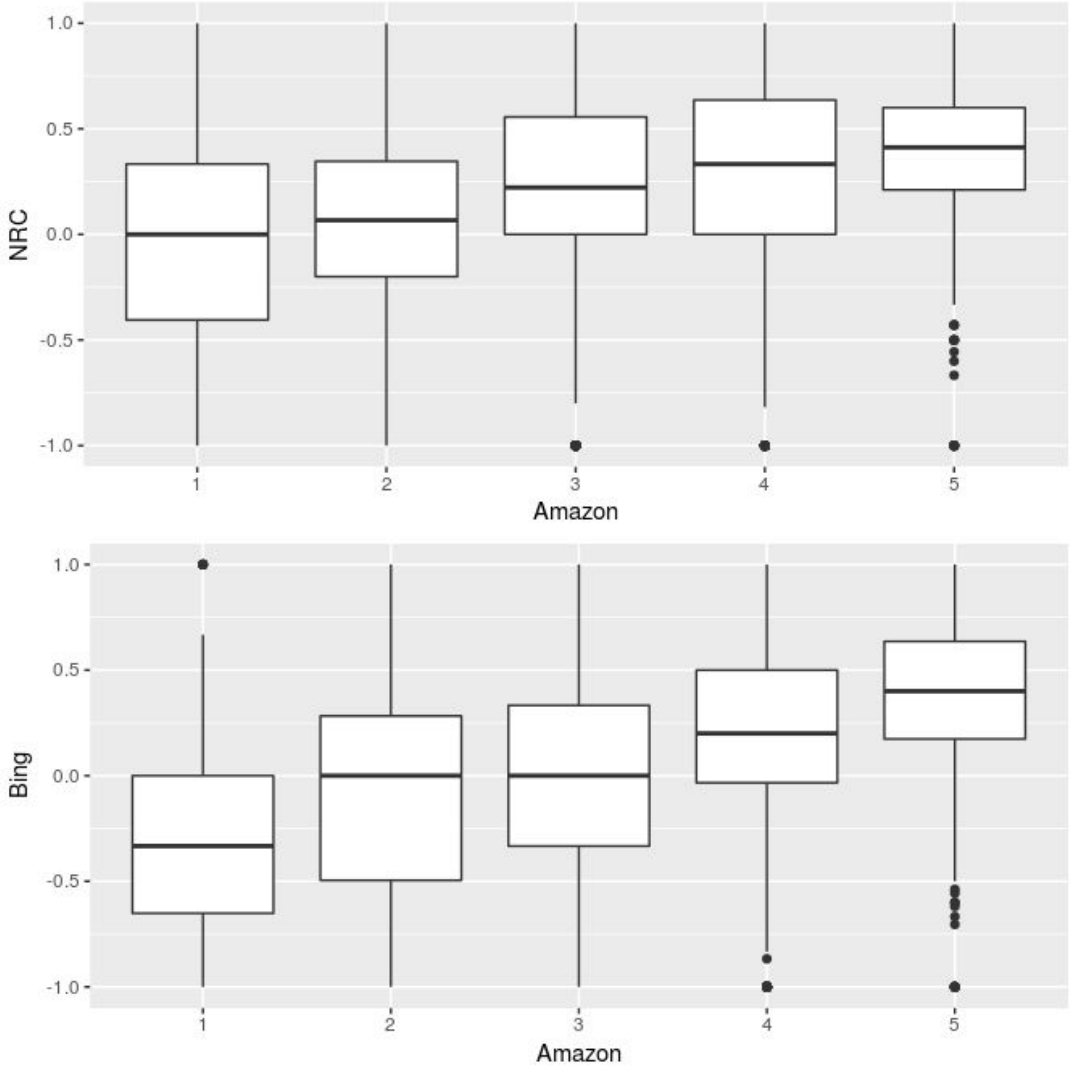

Figure 1: Analyse sentimentale appliquée au overall

En observant les résultats obtenus nous pouvons voir que les deux types d'analyse dégage plus ou moins les mêmes résultats ce qui est bonne chose et montre que les modèles sont bien entrainé.

5) Système de recommandation

Pour le système de recommandation nous avons décidé d'utiliser le système dit UCBF. Ce système est intéressant dans le fait qu'il est capable de recommander à une personne quelque chose en fonction de ce qu'il a déjà pu recommander. En effet les recommandations de ce dernier seront comparé aux recommandations d'autres utilisateurs. Et le système essayera de matcher le plus possible les résultats avec les utilisateurs qui ont les profils les plus similaires.

Concrètement, si deux utilisateurs ont mis des notes similaires sur un même produit, on peut alors raisonnablement prédire que si l'un achète un autre produit, l'autre utilisateur aura potentiellement de l'intérêt pour ce produit !

5) Conclusion

Le système de recommandation que nous avons pu mettre en place n'est pas très précis (8 à 10% selon les échantillons pris aléatoirement). Ceci peut s'expliquer de deux façons.

Premièrement le nombre de review que nous avons sélectionné est trop faible. Pour les tests cela permet d'avoir des résultats plus rapidement mais le résultat est tout de fois moins intéressant.

Et pour finir la méthode d'analyse que nous avons utilisée, la recommandation dite explicite n'est peut-être pas la plus adaptée à cette problématique.

Pour aller plus loin il serait bon de tester d'autres approches et comparer les résultats obtenus afin de déterminer la meilleure façon d'appréhender une problématique comme la nôtre.

III) Résultat

L'ensemble du projet réalisé peut être récupéré sur la plateforme Github à l'adresse suivante:

https://github.com/gary93/systeme_recommandation