Rapport de mineure – 6 (Francis Dupé & Aubry Hervé)

Introduction

Le but de ce projet est de réaliser un programme trouvant les comptes de réseaux sociaux appartenant à une même personne.

Pour ce faire notre programme va d'abord tourner sur un Training Set : un ensemble de couples de réseaux sociaux dont on sait qu'ils appartiennent à une seule et même personne.

On lui donne ensuite l'url d'un compte et il trouve parmi le reste des candidats les comptes susceptibles d'appartenir à la même personne.

Stratégie de corrélation

Nous avons choisi d'utiliser les critères suivants pour déterminer la corrélation de deux comptes :

- Le « nickname » : on utilise la distance de Levenshtein pour déterminer la similitude de deux surnoms. On considère qu'en dessous de 4 lettres cette distance n'est pas pertinente et la remplace par 100% si les deux surnoms sont identique, 0 %sinon.
- Le « realname » : on utilise le rapport entre le nombre de mots en commun et le nombre de mots total pour établir le pourcentage de ressemblance.
- Les adresses « emails » : on part du principe que si les deux comptes ont une adresse en commun alors ils ont une ressemblance de 100%, 0% sinon.
- De même pour les « websites » et les « locations »
- Pour la liste des « profilesLinks » : si l'url de l'un est dans les profilesLinks de l'autre alors ils ont une ressemblance de 100%, 0% sinon.

Pour s'entrainer sur le Training Set nous avons essayé plusieurs méthodes d'apprentissage : SMO, DecisionTable, KStar

Résultats des expériences :

Pour chaque url, on affiche les 10 premiers sites obtenus avec leurs taux de similarité pour : SMO et DecisionTable, un Training Set de 500 couples ou un training set de 50 couples. Les résultats présentés sont quelques exemples assez significatifs de ce qui a été observés : Seul e classifier SMO a donné des résultats satisfaisants...

Utilisateur sur un seul réseau social :

D'abords nous avons essayé de chercher les autres profiles de http://twitter.com/legendaryishuge alors qu'il n'en a pas.

Voici les résultats obtenus (en rouge les similitude supérieur à 95%) :

	Training Set/Methode d'apprentissage			
Urls	Decision Table/50	SMO/50	Decision Table/500	SMO/500
http://www.flickr.com/photos/naked_ffish/	**0,951219512**	??	0,480000000	??
http://www.flickr.com/photos/makkiephoto/	**0,951219512**	??	0,480000000	??
http://www.flickr.com/photos/thedarlinglife/	**0,951219512**	??	??	??
http://www.flickr.com/photos/feaverish/	**0,951219512**	??	??	??
http://www.flickr.com/photos/eugeniagrasso/	**0,951219512**	??	??	??
http://www.flickr.com/photos/andaria/	**0,951219512**	??	??	??
http://www.flickr.com/photos/alexandrasophie/	**0,951219512**	??	??	??
http://www.flickr.com/photos/reprise/	**0,951219512**	??	??	??
http://www.flickr.com/photos/electro_cute/	**0,951219512**	??	??	??
http://www.flickr.com/photos/genergrohl/	**0,951219512**	??	??	??
http://www.flickr.com/photos/fredarmitage/	??	0,677880019	**0,995073892**	0,736277889
http://www.flickr.com/photos/leefabrique/	??	0,677880019	**0,995073892**	0,736277889
http://www.livejournal.com/users/elementary-game/profile	??	0,677880019	**0,995073892**	0,736277889
http://www.livejournal.com/users/legend-of-mike/profile	??	0,677880019	**0,995073892**	0,736277889
http://www.livejournal.com/users/mignardise/profile	??	0,677880019	**0,995073892**	0,736277889
http://www.livejournal.com/users/legenda/profile	??	0,677880019	**0,995073892**	0,736277889
http://www.flickr.com/photos/my_ill/	??	0,677880019	**0,995073892**	0,736277889
http://www.flickr.com/photos/lasegundaconigriega/	??	0,592002137	??	0,654104120
http://www.flickr.com/photos/eleventhirtythree/	??	0,573597171	??	0,635943789
http://www.flickr.com/photos/gunnargestur/	??	0,549994697	??	0,612378505
http://www.flickr.com/photos/kristamasklousch/	??	??	0,480000000	??

On constate alors que DecisionTable trouve beaucoup trop de profiles similaires, est n'est pas du tout discriminant.

A l'inverse SMO rejette bien tous les profiles avec une similarité inférieur à 68% pour un training set de 50 et inférieur à 73% pour celui de 500.

Utilisateur sur deux réseaux sociaux au nickname proche.

D'abords nous avons essayé de chercher les autres profiles de http://www.flickr.com/photos/tokioshi/. On s'attend à en trouver un avec un nickname similaire.

Voici les résultats obtenus (en rouge les similitude supérieur à 95%) :

	Training Set/Methode d'apprentissage			
Urls	Decision Table/50	SMO/50	Decision Table/500	SMO/500
http://twitter.com/igorhosse	**0,951219512**	??	??	??
http://twitter.com/antonioteoli	**0,951219512**	??	??	??
http://twitter.com/joshkoscheck	**0,951219512**	??	??	??
http://youtube.com/user/talkinboutme2	**0,951219512**	??	??	??
http://youtube.com/user/taoofpooh26	**0,951219512**	??	??	??
http://youtube.com/user/yorktonfilm	**0,951219512**	??	??	??
http://youtube.com/user/todmaffin	**0,951219512**	??	??	??
http://youtube.com/user/hotelslive	**0,951219512**	??	??	??
http://twitter.com/newtothekitchen	**0,951219512**	??	??	??
http://www.livejournal.com/users/trains/profile	**0,951219512**	??	??	??
http://www.livejournal.com/users/tokioshi/profile	??	**0,980932330**	**0,960784314**	**0,992818684**
http://www.livejournal.com/users/takca/profile	??	0,861400084	??	**0,963320302**
http://www.livejournal.com/users/neoromantic/profile	??	0,840502493	??	**0,958591661**
http://www.livejournal.com/users/robynbright/profile	??	0,816011052	??	**0,950052627**
http://www.livejournal.com/users/nickinuse/profile	??	0,781704512	??	0,933435745
http://www.livejournal.com/users/violet-milk/profile	??	0,780951642	??	0,936185394
http://www.livejournal.com/users/emsi/profile	??	0,755599218	??	0,919406384
http://www.livejournal.com/users/mekish/profile	??	0,734142889	**0,995073892**	??
http://www.livejournal.com/users/ff-mortsanscafe/profile	??	0,725375371	??	0,915679942
http://www.livejournal.com/users/posta0_/profile	??	0,703447801	??	0,906088164
http://www.livejournal.com/users/toropchin/profile	??	??	**0,995073892**	??
http://www.livejournal.com/users/jokishop/profile	??	??	**0,995073892**	??
http://www.livejournal.com/users/takie-sny/profile	??	??	**0,995073892**	??
http://www.livejournal.com/users/tat-oshka/profile	??	??	**0,995073892**	??
http://twitter.com/thejoshhd	??	??	**0,995073892**	??
http://www.livejournal.com/users/zojirushi/profile	??	??	**0,960784314**	??
http://www.livejournal.com/users/forioscribe/profile	??	??	**0,960784314**	??
http://www.livejournal.com/users/tokarchuk/profile	??	??	**0,960784314**	??
http://www.livejournal.com/users/otter78/profile	??	??	??	0,899715103

On constate que la méthode DecisionTable ne trouve pas le bon lien avec un training set de 50, et la trouve avec un mauvais classement (7ème position) avec un training set de 500.

A l'inverse, SMO trouve à chaque fois le bon compte. Avec un écart plus notable entre premier et second compte pour un training set de 50 (98%/86% contre 99%/96%).

Utilisateur sur deux réseaux sociaux au nickname complétement différent.

D'abords nous avons essayé de chercher les autres profiles de http://twitter.com/lolrenee. On s'attend à en trouver un avec un nickname complétement différent (http://youtube.com/user/missdoesntmiss).

Voici les résultats obtenus (en rouge les similitude supérieur à 95%) :

	Training Set/Methode d'apprentissage			
Urls	Decision Table/50	SMO/50	Decision Table/500	SMO/500
http://www,flickr,com/photos/alenachendler/	**0,951219512**	??	??	??
http://www,flickr,com/photos/bonnabelle/	**0,951219512**	??	??	??
http://www,flickr,com/photos/ballena53/	**0,951219512**	??	??	??
http://www,flickr,com/photos/colourfullife/	**0,951219512**	??	??	??
http://www,flickr,com/photos/elgatococo/	**0,951219512**	??	??	??
http://www,flickr,com/photos/lona/	**0,951219512**	??	??	??
http://www,flickr,com/photos/indie_ingenue/	**0,951219512**	??	??	??
http://www,flickr,com/photos/waterandsleep/	**0,951219512**	??	??	??
http://www,flickr,com/photos/soul2squeez/	**0,951219512**	??	??	??
http://www,flickr,com/photos/louobedlam/	**0,951219512**	??	??	??
http://www,livejournal,com/users/oh-velveteen/profile	??	0,910309339	??	**0,976492409**
http://www,flickr,com/photos/aaarenee/	??	0,884383691	??	0,915212952
http://www,livejournal,com/users/dolzhenkov/profile	??	0,859133970	??	**0,954737549**
http://www,livejournal,com/users/lalaranel/profile	??	0,829218798	??	0,931206816
http://www,livejournal,com/users/otter78/profile	??	0,819288009	??	0,946082844
http://www,flickr,com/photos/lolitanie/	??	0,812840771	??	??
http://www,flickr,com/photos/oliviarenee/	??	0,799991721	**0,995073892**	??
http://www,flickr,com/photos/aliciarenee/	??	0,799991721	**0,995073892**	??
http://www,livejournal,com/users/gloomleage/profile	??	0,767123563	??	0,910458481
http://www,livejournal,com/users/neoromantic/profile	??	0,761879972	??	0,926084591
http://www,flickr,com/photos/lockergnome/	??	??	**0,995073892**	??
http://www,flickr,com/photos/moline/	??	??	**0,995073892**	??
http://www,flickr,com/photos/laurencephilomene/	??	??	**0,995073892**	??
http://www,flickr,com/photos/juliedee/	??	??	**0,995073892**	??
http://www,flickr,com/photos/julianzee/	??	??	**0,995073892**	??
http://www,flickr,com/photos/l4urenella/	??	??	**0,995073892**	??
http://www,flickr,com/photos/moretrees/	??	??	**0,995073892**	??
http://www,flickr,com/photos/llorensot/	??	??	**0,995073892**	??
http://www,livejournal,com/users/golovorez/profile	??	??	??	0,885580684
http://www,livejournal,com/users/loree/profile	??	??	??	0,882575834
http://www,livejournal,com/users/loreleei/profile	??	??	??	0,882575834

Ici les deux méthodes ne trouvent pas le profil attendu, et aucun des profils trouvés ne semblent correspondre à la même personne.

Cependant alors que DecisionTable se fourvoie complétement, SMO ne trouve que deux profile supérieur à 95% de similarité lorsqu'elle a un training de 500 et aucun avec un training de 50. Elle donc l'avantage de moins sortir d'aberration à défaut de trouver le bon compte.

Conclusion

On peut déduire des trois expériences précédente que :

1. SMO nous a semblé relativement performant pour ce problème.
2. Un Training set de 50 semble préférable à un de 500. En effet SMO semble surapprendre avec un training set de 500 et même s'il est plus sûr de lui pour les réponses justes, il augmente aussi significativement la similarité des couples non-liés.
3. Les attributs autres que le « nickname » étant mal renseignés, il est difficile de trouver des comptes liés avec eux.