Detecting spammers in Chinese microblog Sina weibo

DNDS 6005 - Data Mining and Big Data Analytics

Name: Hao CUI

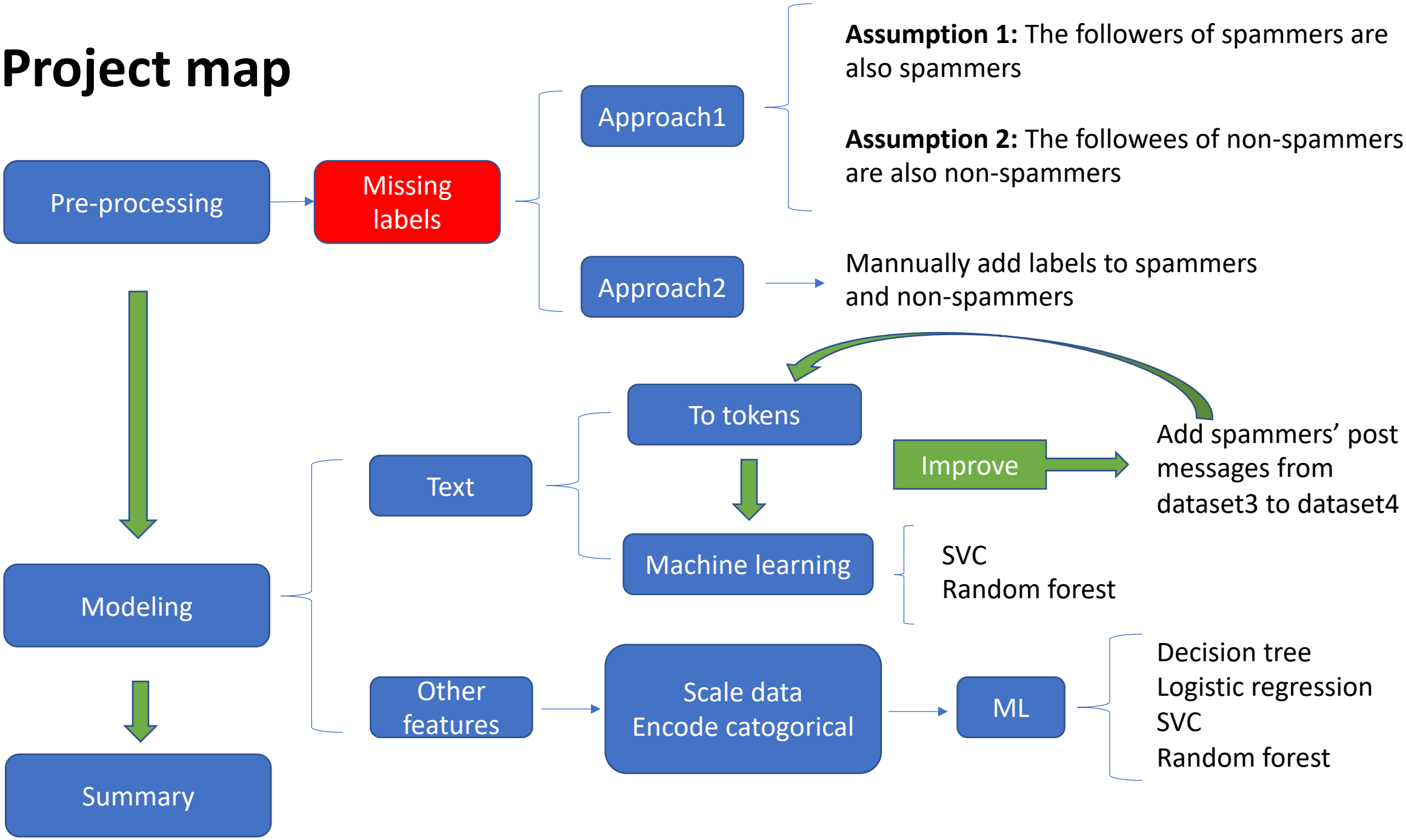
March 4th, 2019

Dataset overview

Dataset\Properties	Name	Dimension	Main features
Dataset 1	follower_followee.csv	(142368, 10)	follower_id, followee_id, post_num, gender
Dataset 2 (Discarded)	post.csv	(35, 11)	post_time, content, poster_id, poster_url, repost_num, comment_num
Dataset 3	user_post.csv	(48813,8)	post_time, content, poster_id, poster_url, repost_num, comment_num
Dataset 4	weibo_user.csv	(781, 10)	user_id, user_name, user_image, gender, class, message, post_num, follower_num, followee_num, is_spammer (incomplete labels)

Chinese microblog dataset crawled from sina weibo
UCI machine learning repository

Project map

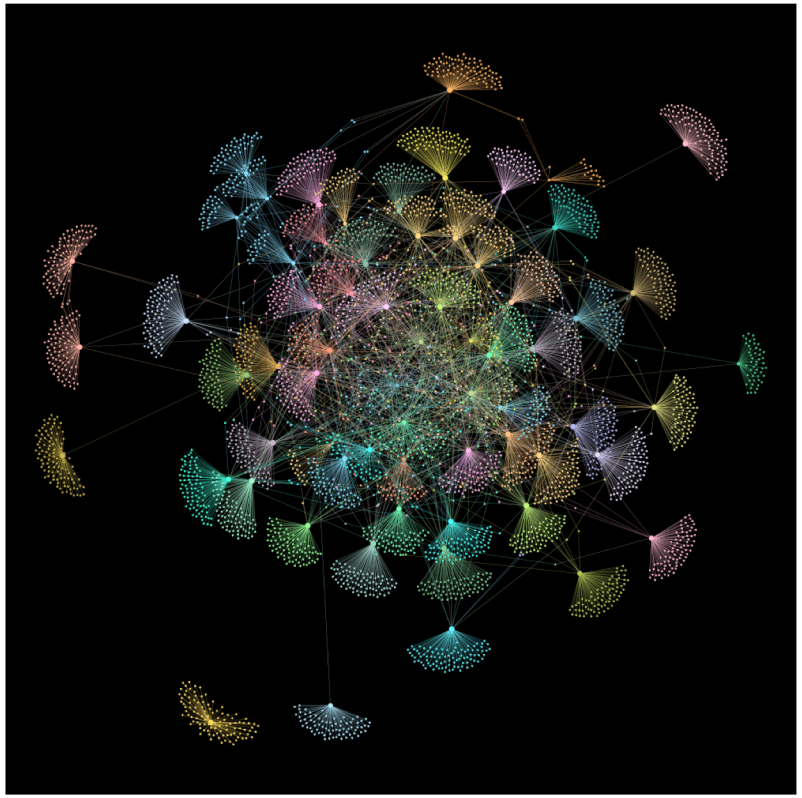

Pre-processing
Missing labels
Approach1
Assumption 1: The followers of spammers are also spammers
Assumption 2: The followees of non-spammers are also non-spammers
Approach2
Mannually add labels to spammers and non-spammers
Modeling
Text
To tokens
Improve
Add spammers' post messages from dataset3 to dataset4
Machine learning
SVC
Random forest
Other features
Scale data
Encode catogorical
ML
Decision tree
Logistic regression
SVC
Random forest
Summary

Missing labels: directed social network

Spammers and spammers' followers

Nodes: 7079; Edges: 7189; Labeled: 112
133 potential spammers (21 +1 and 21 -1) found
Possible explanations: 1) some users follow many kinds of people 2) some initial labels are wrong

Non-spammers and non-spammers' followees



Nodes: 4850; Edges: 5772; Labeled: 66
No potential non-spammers found in dataset4

Text mining

Different from English, there are no spaces in Chinese written texts.

Wrote a function to separate words and phrases by spaces, an example is shown below:

Test the correctness of the function.

```
df4_new.iloc[500]['message']
```

```
'    Lv.3      香港其他     简介：在当前现实的狭隘基础上，有高尚理想，全面的计划；在一步一步行动上，想到远大前途，脚踏实地地稳步前进，才能有
所成就   '
```

```
a=df4_new.iloc[500]['message']
```

```
to_token(a)
```

```
'          Lv.3                    香港   其他          简介   ：   在   当前   现实   的   狭隘   基础   上   ，   有   高尚   理想   ，   全面
的   计划   ；   在   一步   一步   行动   上   ，   想到   远大   前途   ，   脚踏实地   地   稳步前进   ，   才能   有所   成就        '
```

Text mining - Machine learning

Technique	Parameter	Accuracy
SVC, CountVectorizer	kernel=rbf, linear, gamma=scale	0.702
SVC, CountVectorizer	Kernel=sigmoid	0.708
SVC, CountVectorizer	Kernel=poly	0.68
SVC, TfidfVectorizer	kernel=rbf,gamma=scale	0.691
SVC, TfidfVectorizer	Kernel=linear	0.704
Random forest,CountVectorizer	n_estimators=200, max_depth=40	0.785
Random forest, TfidfVectorizer	n_estimators=500, max_depth=20	0.752

Extract users' id from poster_url in df3.
Make a dictionary such that the keys are the spammers' User_ID, and the values are the aggregated contents of the spammers.
Add the spammers' messages from df3 to df4 .
Transform the updated messages in df4 to tokens for machine learning.

Text mining - Machine learning

Technique	Parameter	Accuracy
SVC, CountVectorizer	kernel=rbf, gamma=scale	0.857
SVC, CountVectorizer	Kernel=linear	0.809
SVC, CountVectorizer	Kernel=sigmoid	0.866
SVC, CountVectorizer	Kernel=poly	0.771
SVC, TfidfVectorizer	kernel=rbf, sigmoid, gamma=scale	0.788
SVC, TfidfVectorizer	Kernel=linear	0.802
Random forest,CountVectorizer	n_estimators=200, max_depth=80	0.882
Random forest, TfidfVectorizer	n_estimators=500, max_depth=20	0.88
GaussianNB, CountVectorizer		0.721
GaussianNB, TfidfVectorizer		0.691

Using other features - Machine learning

Gender, class, post_num, follower_num, followee_num

Technique	Parameter	Pipeline accuracy
Decision tree	max_depth= 2	0.852
Random forest	'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'n_estimators': 100	0.857
LogisticRegression, MinMax	C=1	0.821
LogisticRegression, Standard	C=1000	0.821
LogisticRegression, Robust	C=100	0.837
KNeighbors, MinMaxScaler	n_neighbors=7	0.816
KNeighbors, StandardScaler	n_neighbors=27	0.806
Kneighbors, RobustScaler	n_neighbors=10	0.842
SVC, MinMaxScaler	kernel='linear'	0.821
SVC, StandardScaler	kernel=rbf	0.821
SVC, RobustScaler	kernel='linear'	0.842

Summary

Difficulties

- Determining labels: use social network relations, add mannually, some original labels are not correct ...
- Language: difficulty in reading in the dataset, encoding, transform to tokens ...

Future research

- Classification using image
- Semi-supervised learning

Your suggestions


Thanks for watching!