Credit Estimation for Chinese Construction Companies

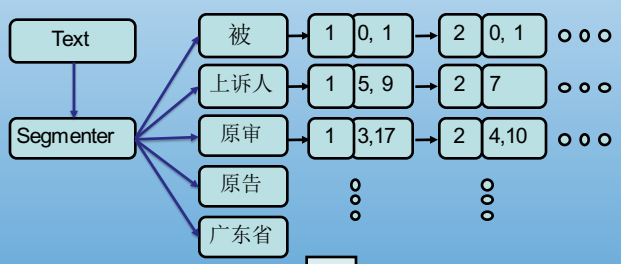
Christopher Wang

Objective

- Provide another view of companies: court cases involving the studied companies.
- Facilitate investment decisions other than financial status of the company.
- Extendable to news, magazine articles, etc.

Data Processing

- Crawled court files from Chinese government website.
- Parsed document text from the html.
- Segmented the document text.
- Built inverted index with positional posting.
- Found court cases where construction companies are the defendants.
- Built ready-to-use data on the qualified cases.



	被	上诉人	原审	原告	广东省
Doc1	5	2	4	4	2
Doc2	4	1	2	2	0
Doc3	3	5	3	2	1
Doc4	1	3	0	1	0

Classification Models and Results

- Feature selection: Filtering out most frequent (0 – 10%) and less frequent words (0-10%).
- Methods used: SVM and Naïve Bayes.

Error rate:

Naïve Bayes Test Results (70/30 Train/Test)

< y% of the doc / > x% of the doc	90	91	92	93	94	95	96	97	98	99	100
0	0.20	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23
1	0.20	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23
2	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17
3	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
4	0.13	0.13	0.13	0.13	0.13	0.17	0.17	0.20	0.20	0.20	0.20
5	0.13	0.13	0.13	0.13	0.13	0.17	0.17	0.17	0.20	0.20	0.17
6	0.17	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
7	0.13	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17
8	0.13	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17
9	0.13	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17
10	0.20	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17

SVM Test Results (70/30 Train/Test)

< y% of the doc / > x% of the doc	90	91	92	93	94	95	96	97	98	99	100
0	0.00	0.23	0.23	0.23	0.20	0.27	0.27	0.27	0.27	0.27	0.27
1	0.27	0.23	0.23	0.23	0.20	0.27	0.27	0.27	0.27	0.27	0.27
2	0.23	0.20	0.23	0.23	0.17	0.23	0.23	0.23	0.23	0.23	0.23
3	0.23	0.30	0.30	0.30	0.20	0.23	0.23	0.23	0.23	0.23	0.23
4	0.27	0.27	0.27	0.27	0.20	0.23	0.23	0.23	0.23	0.23	0.23
5	0.27	0.23	0.23	0.27	0.20	0.23	0.23	0.23	0.23	0.23	0.23
6	0.27	0.27	0.23	0.27	0.20	0.23	0.23	0.23	0.23	0.23	0.23
7	0.23	0.23	0.23	0.23	0.20	0.23	0.23	0.23	0.23	0.23	0.23
8	0.23	0.23	0.23	0.23	0.20	0.23	0.23	0.23	0.23	0.23	0.23
9	0.23	0.27	0.27	0.27	0.20	0.23	0.23	0.23	0.23	0.23	0.23
10	0.23	0.23	0.23	0.23	0.20	0.23	0.23	0.23	0.23	0.23	0.23

Better

Naïve Bayes Test Results (50/50 Train/Test)

< y% of the doc / > x% of the doc	90	91	92	93	94	95	96	97	98	99	100
0	0.16	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26
1	0.16	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26
2	0.22	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26
3	0.18	0.22	0.22	0.22	0.22	0.24	0.24	0.24	0.24	0.24	0.24
4	0.18	0.22	0.22	0.22	0.26	0.26	0.26	0.26	0.26	0.26	0.26
5	0.18	0.18	0.18	0.18	0.20	0.22	0.22	0.20	0.20	0.20	0.20
6	0.18	0.18	0.18	0.20	0.18	0.20	0.20	0.20	0.20	0.20	0.20
7	0.18	0.18	0.18	0.20	0.18	0.18	0.18	0.18	0.18	0.18	0.18
8	0.18	0.18	0.18	0.20	0.20	0.18	0.18	0.20	0.20	0.20	0.20
9	0.18	0.18	0.18	0.20	0.22	0.22	0.22	0.22	0.22	0.22	0.22
10	0.16	0.18	0.18	0.20	0.20	0.20	0.20	0.20	0.22	0.22	0.22

SVM Test Results (50/50 Train/Test)

Overfitting?

< y% of the doc / > x% of the doc	90	91	92	93	94	95	96	97	98	99	100
0	0.24	0.22	0.22	0.20	0.20	0.24	0.24	0.24	0.22	0.24	0.24
1	0.24	0.22	0.22	0.20	0.20	0.24	0.24	0.24	0.22	0.24	0.24
2	0.20	0.26	0.26	0.26	0.18	0.26	0.26	0.22	0.22	0.22	0.22
3	0.22	0.18	0.18	0.18	0.18	0.22	0.22	0.20	0.20	0.20	0.20
4	0.22	0.20	0.20	0.20	0.14	0.20	0.20	0.18	0.18	0.18	0.18
5	0.18	0.16	0.14	0.16	0.14	0.20	0.20	0.20	0.18	0.20	0.20
6	0.20	0.16	0.16	0.14	0.12	0.20	0.20	0.20	0.20	0.20	0.20
7	0.18	0.14	0.14	0.14	0.12	0.20	0.20	0.20	0.20	0.20	0.20
8	0.18	0.14	0.14	0.14	0.12	0.20	0.20	0.20	0.20	0.20	0.20
9	0.16	0.10	0.10	0.10	0.10	0.20	0.20	0.16	0.18	0.18	0.18
10	0.16	0.12	0.12	0.12	0.10	0.20	0.20	0.16	0.18	0.18	0.18

Discussion

Most Indicative Tokens in Naïve Bayes Model

0% ~ 100%

10% ~ 90%

Number of Tokens

< y% of the doc / > x% of the doc	90	91	92	93	94	95	96	97	98	99	100
0	10750	10729	10726	10724	10722	10716	10716	10713	10711	10710	10709
1	10750	10729	10726	10724	10722	10716	10716	10713	10711	10710	10709
2	4343	4322	4319	4317	4315	4309	4309	4306	4304	4303	4302
3	2893	2872	2869	2867	2865	2859	2859	2856	2854	2853	2852
4	2291	2270	2267	2265	2263	2257	2257	2254	2252	2251	2250
5	1862	1841	1838	1836	1834	1828	1828	1825	1823	1822	1821
6	1622	1601	1598	1596	1594	1588	1588	1585	1583	1582	1581
7	1438	1417	1414	1412	1410	1404	1404	1401	1399	1398	1397
8	1295	1274	1271	1269	1267	1261	1261	1258	1256	1255	1254
9	1163	1142	1139	1137	1135	1129	1129	1126	1124	1123	1122
10	1082	1061	1058	1056	1054	1048	1048	1045	1043	1042	1041

10 times deduction

- Feature Selection is a very important task in NLP.
  - Further work needs to be done to make the prediction even better: numbers, Chinese representation of numbers, names, locations, punctuations, etc.
  - Further deduction also enables other methods to become viable, such as logistic regression, decision tree, etc.
- There could be multiple defendants in a case and the classification is done assuming all defendants are either positive or negative. Further detection can be done to distinguish one another.