IT-633
Data Mining and Warehousing

Group 32

Mentor : Prof. P.M.Jat

GROUP MEMBERS

Nihit Alamuru (201301036)

Ankit Taral (201301087)

Rahul Kumar (201301204)

Samarth Desai (201301456)

Problem Definition

- The goal of our project is to apply machine learning for sentiment analysis or opinion mining on user generated comments of our subset(subreddit).
- We are looking to classify comments as being positive or negative.
- An opinion is defined as a positive or negative sentiment, view, attitude, emotion or appraisal about an entity or an aspect of the entity from an opinion holder.

Tools used

Python 2.7

Numpy

Pandas

BeautifulSoup4

Sklearn

NLTk(Natural Language Toolkit)

Dataset

	A	B
1	id	review
2	9999_0	Watching Time Chasers, it obvious that it was made by a bunch of friends. Maybe they were sitting around one day in film school and said, \"Hey, let's pool our r
3	45057_0	I saw this film about 20 years ago and remember it as being particularly nasty. I believe it is based on a true incident: a young man breaks into a nurses' home a
4	15561_0	Minor Spoilers In New York, Joan Barnard (Elvire Audrey) is informed that her husband, the archeologist Arthur Barnard (John Saxon), was mysterious
5	7161_0	I went to see this film with a great deal of excitement as I was at school with the director, he was even a good friend of mine for a while. But sorry mate, this film
6	43971_0	Yes, I agree with everyone on this site this movie is VERY VERY bad. To even call this a movie is an insult to all movies ever made. It's 40 minutes long. <br
7	36495_0	Jennifer Ehle was sparkling in \"Pride and Prejudice.\" Jeremy Northam was simply wonderful in \"The Winslow Boy.\" With actors of this caliber, this film had to
8	49472_0	Amy Poehler is a terrific comedian on Saturday Night Live, but her role in this movie doesn't give her anything to work with. Her character, a publisher's represen
9	36693_0	A plane carrying employees of a large biotech firm--including the CEO's daughter--goes down in thick forest in the Pacific Northwest. When the search and rescu
10	316_0	A well made, gritty science fiction movie, it could be lost among hundreds of other similar movies, but it has several strong points to keep it near the top. For on
11	32454_0	"37128_0

20000+ reddit comments

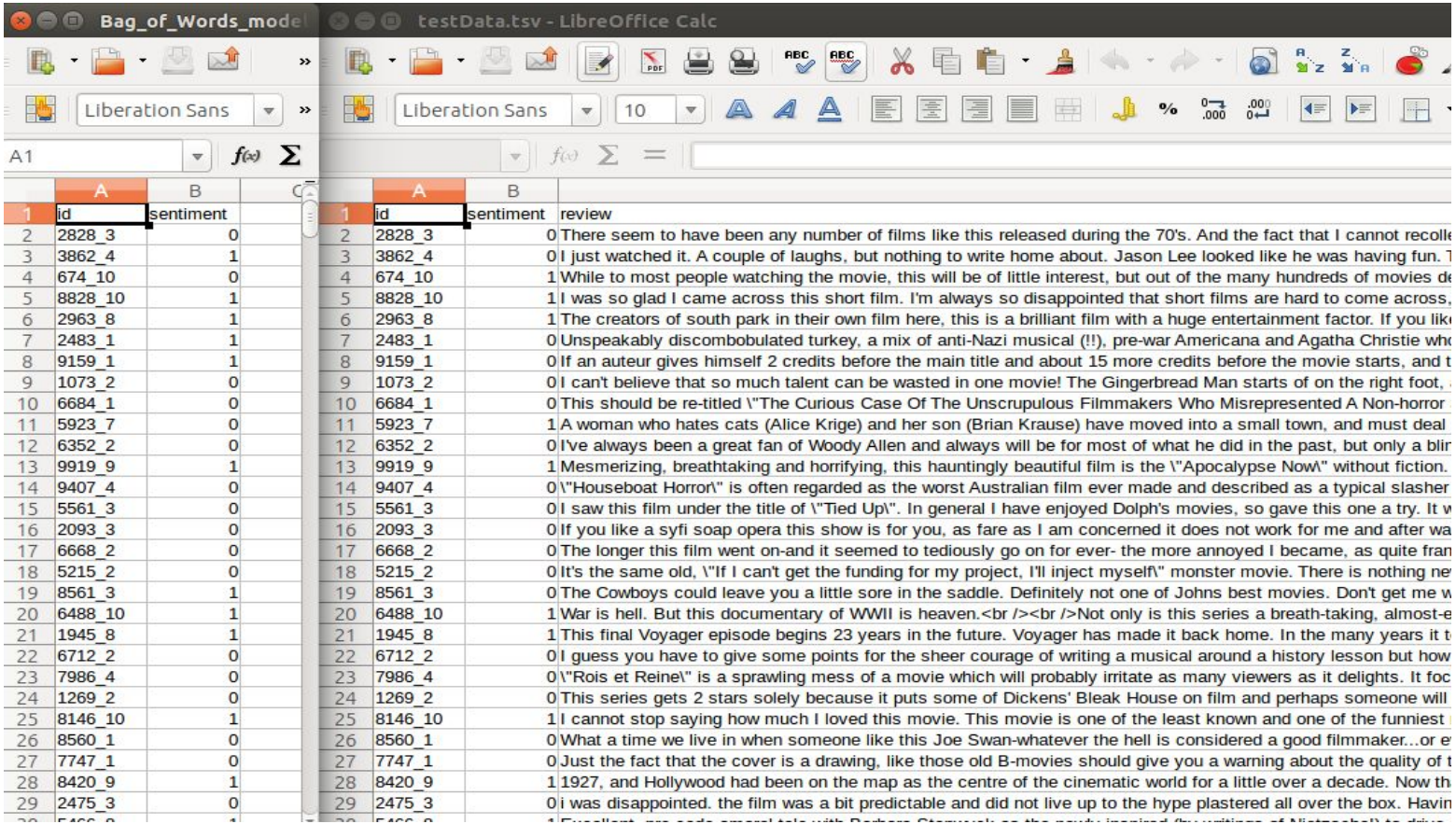
Id of reddit comment

Review of movies

Methodology

- Tokenization
- Feature Extraction
- Classification using classifiers
  - Random Forest Classifier
  - Naive bayes

Comparison



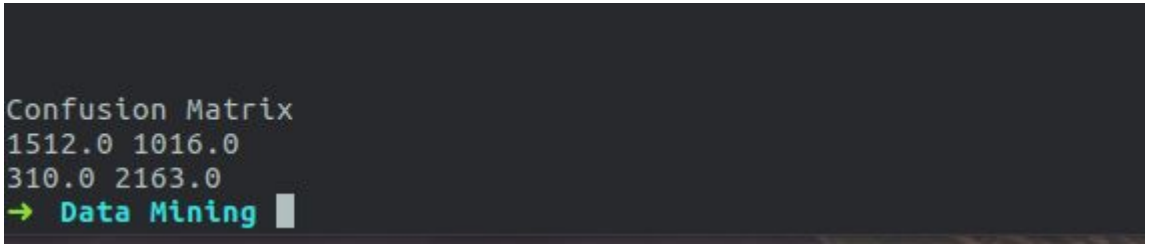
Bag_of_Words_model

	A	B
1	id	sentiment
2	2828_3	0
3	3862_4	1
4	674_10	0
5	8828_10	1
6	2963_8	1
7	2483_1	1
8	9159_1	1
9	1073_2	0
10	6684_1	0
11	5923_7	0
12	6352_2	0
13	9919_9	1
14	9407_4	0
15	5561_3	0
16	2093_3	0
17	6668_2	0
18	5215_2	0
19	8561_3	1
20	6488_10	1
21	1945_8	1
22	6712_2	0
23	7986_4	0
24	1269_2	0
25	8146_10	1
26	8560_1	0
27	7747_1	0
28	8420_9	1
29	2475_3	0

testData.tsv - LibreOffice Calc

	A	B	
1	id	sentiment	review
2	2828_3	0	There seem to have been any number of films like this released during the 70's. And the fact that I cannot recolle
3	3862_4	0	I just watched it. A couple of laughs, but nothing to write home about. Jason Lee looked like he was having fun. T
4	674_10	1	While to most people watching the movie, this will be of little interest, but out of the many hundreds of movies de
5	8828_10	1	I was so glad I came across this short film. I'm always so disappointed that short films are hard to come across,
6	2963_8	1	The creators of south park in their own film here, this is a brilliant film with a huge entertainment factor. If you lik
7	2483_1	0	Unspeakably discombobulated turkey, a mix of anti-Nazi musical (!!), pre-war Americana and Agatha Christie who
8	9159_1	0	If an auteur gives himself 2 credits before the main title and about 15 more credits before the movie starts, and t
9	1073_2	0	I can't believe that so much talent can be wasted in one movie! The Gingerbread Man starts of on the right foot,
10	6684_1	0	This should be re-titled \"The Curious Case Of The Unscrupulous Filmmakers Who Misrepresented A Non-horror
11	5923_7	1	A woman who hates cats (Alice Krige) and her son (Brian Krause) have moved into a small town, and must deal
12	6352_2	0	I've always been a great fan of Woody Allen and always will be for most of what he did in the past, but only a blir
13	9919_9	1	Mesmerizing, breathtaking and horrifying, this hauntingly beautiful film is the \"Apocalypse Now\" without fiction.
14	9407_4	0	\"Houseboat Horror\" is often regarded as the worst Australian film ever made and described as a typical slasher
15	5561_3	0	I saw this film under the title of \"Tied Up\". In general I have enjoyed Dolph's movies, so gave this one a try. It w
16	2093_3	0	If you like a syfi soap opera this show is for you, as fare as I am concerned it does not work for me and after wa
17	6668_2	0	The longer this film went on-and it seemed to tediously go on for ever- the more annoyed I became, as quite fran
18	5215_2	0	It's the same old, \"If I can't get the funding for my project, I'll inject myself\" monster movie. There is nothing ne
19	8561_3	0	The Cowboys could leave you a little sore in the saddle. Definitely not one of Johns best movies. Don't get me w
20	6488_10	1	War is hell. But this documentary of WWII is heaven. Not only is this series a breath-taking, almost-e
21	1945_8	1	This final Voyager episode begins 23 years in the future. Voyager has made it back home. In the many years it t
22	6712_2	0	I guess you have to give some points for the sheer courage of writing a musical around a history lesson but how
23	7986_4	0	\"Rois et Reine\" is a sprawling mess of a movie which will probably irritate as many viewers as it delights. It foc
24	1269_2	0	This series gets 2 stars solely because it puts some of Dickens' Bleak House on film and perhaps someone will
25	8146_10	1	I cannot stop saying how much I loved this movie. This movie is one of the least known and one of the funniest
26	8560_1	0	What a time we live in when someone like this Joe Swan-whatever the hell is considered a good filmmaker...or e
27	7747_1	0	Just the fact that the cover is a drawing, like those old B-movies should give you a warning about the quality of t
28	8420_9	1	1927, and Hollywood had been on the map as the centre of the cinematic world for a little over a decade. Now th
29	2475_3	0	i was disappointed. the film was a bit predictable and did not live up to the hype plastered all over the box. Havin

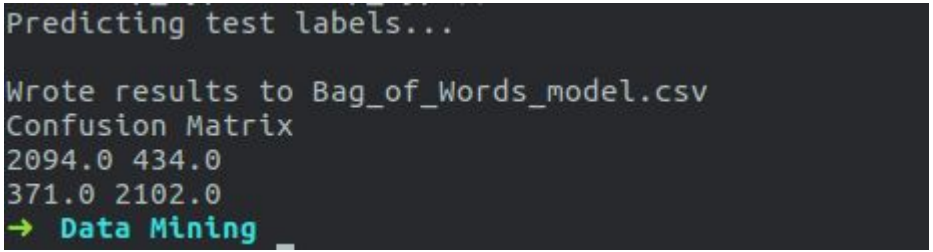
Confusion Matrix using Naive Bayes



N = 5000	Predicted Yes	Predicted No
Actual Yes	1512	1016
Actual No	310	2163

Accuracy : 73.52 %

Confusion Matrix using Random Forest



N = 5000	Predicted Yes	Predicted No
Actual Yes	2094	434
Actual No	371	2102

Accuracy : 83.92 %

Conclusion

This approach has helped us draw conclusions about the sentiments of the users. It helped us understand how much a person can be influenced by reading an online review of general public.

We were able to:

- Deeply Understand algorithms and implementing them
- Familiarize ourselves with aspects of data analysis

References

- http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf
- https://citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics/
- http://www.saedsayad.com/decision_tree.htm
- A. Go, L. Huang and R. Bhayani, "Twitter Sentiment Classification using Distant Supervision," The Stanford Natural Language Processing Group, 2008/2009.

Thank You