Samir Karmacharya
CS 410 Course Project
Fall 2018
IMDB Review Sentiment Analysis
Code: https://github.com/samirk927/CS410

IMDB Review Sentiment Analysis

IMDB or Internet Movie Database provides diverse information on Movies, TV series and their associated cast, crew, biographies, trivia etc. One of the most interesting and important information that it provides is reviews and ratings. The objective of this project is to train the model and predict the sentiment on the testing dataset.

Dataset for this project has been gathered from:
Source: http://ai.stanford.edu/~amaas/data/sentiment/

The dataset contains 50,000 reviews from IMDB which are divided evenly into 25K training and 25K testing dataset. On the entire collection, no more than 30 reviews are allowed per movie. Any movie with a rating score of <=4 is considered as negative and >=7 as positive. Neutral ratings (rating 5 and 6) are not considered.

Files
There are two main directories (train/ and test/) which corresponds to training and testing dataset. They include both positive and negative reviews.

Overview of functions:

1) imdb_data_preprocess : Creates training dataset as part of preprocessing where the stop words are removed and stored in training dataset path:

   • training_dataset = "./aclImdb/train/" # source data

   Testing dataset is created as part of preprocessing where the stop words and polarity are removed and stored in testing dataset path:

   • testing_dataset = "./aclImdb/test/"

2) remove_stopwords : Removes the stopwords from the input sentence and returns the sentence.

3) unigram_process : Takes the data as the input and returns a vectorizer of the unigram as output

4) bigram_process : Takes the data as the input and returns a vectorizer of the bigram as output

5) tfidf_process : Takes the data as the input and returns a vectorizer of the tfidf as output

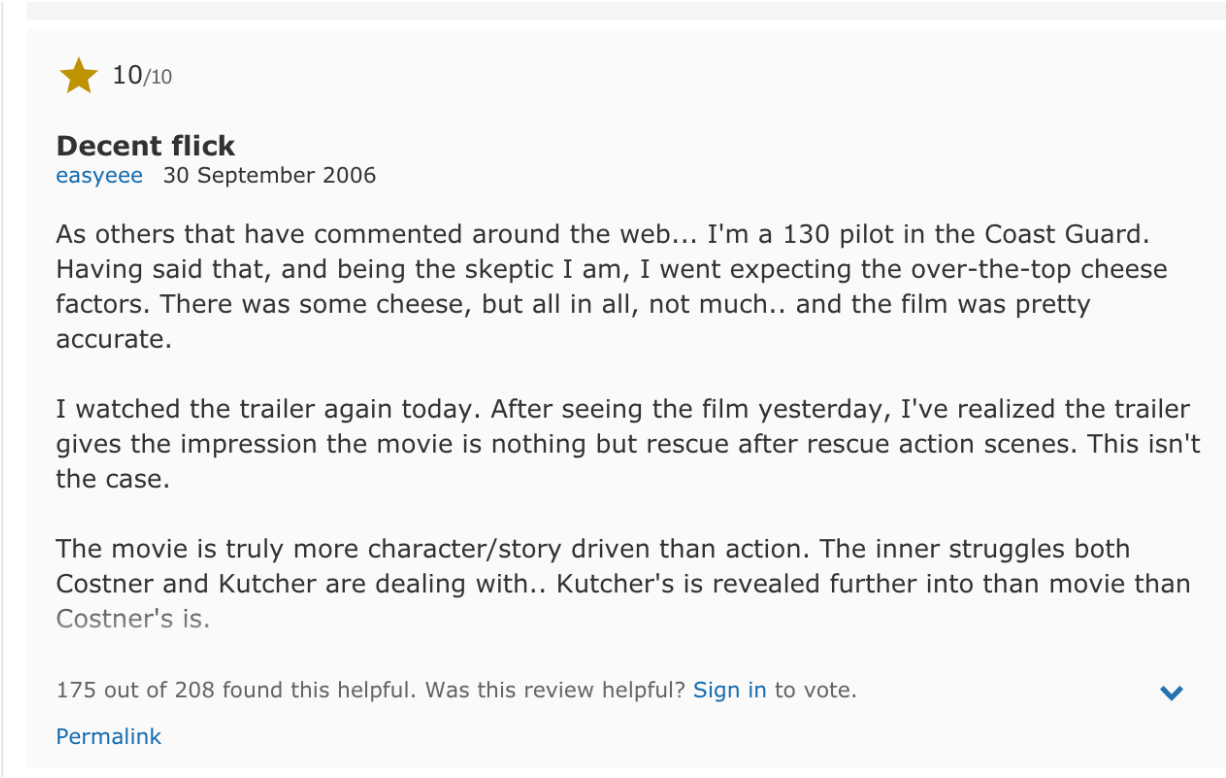
6) retrieve_data : Takes a CSV file as the input and returns the corresponding arrays of labels and data as output

7) stochastic_descent : Applies Stochastic descent on the training data and returns the predicted labels

8) accuracy : Finds the accuracy in percentage given the training and test labels

9) write_txt : Takes training/testing data as input and writes to a csv file.

Sample positive review:



10/10

Decent flick
easyeee 30 September 2006

As others that have commented around the web... I'm a 130 pilot in the Coast Guard. Having said that, and being the skeptic I am, I went expecting the over-the-top cheese factors. There was some cheese, but all in all, not much.. and the film was pretty accurate.

I watched the trailer again today. After seeing the film yesterday, I've realized the trailer gives the impression the movie is nothing but rescue after rescue action scenes. This isn't the case.

The movie is truly more character/story driven than action. The inner struggles both Costner and Kutcher are dealing with.. Kutcher's is revealed further into than movie than Costner's is.

175 out of 208 found this helpful. Was this review helpful? Sign in to vote.

Permalink

Step 1: Reading and preprocessing the data:
Download the data from mentioned source and unzip.

Training dataset is created as part of preprocessing where the stop words are removed and stored in training dataset path:

- training_dataset = "./aclImdb/train/" # source data

	A	B	C	D	E	F
1	row_Numbe	text	polarity			
2	15776	It's sort hard	0			
3	20649	poor remake	0			
4	10663	night TV, del	1			
5	3413	movie dying	1			
6	6021	classy film pu	1			
7	15791	pains write s	0			
8	16231	Slim Slam Slu	0			
9	8801	It's sly 60's lc	1			
.0	17721	God, bored h	0			
.1	18015	Shirley Jacks	0			
.2	10794	Woo's maste	1			
.3	24181	WATCH SAD	0			
.4	5492	earlier film, "	1			
.5	1457	true measure	1			
.6	20065	Justifications	0			
.7	17831	Inconvenient	0			
.8	11454	Short synops	1			
.9	5913	Japan 1918.	1			
0	6642	Stephen King	1			

Testing dataset is created as part of preprocessing where the stop words and polarity are removed and stored in testing dataset path:

- testing_dataset = "./aclImdb/test/"

	A	B	C	D	E
1	row_Numbe	text			
2	342	say first film ever? You can't rate this, it's supposed entertaining. rate it, give 10. stunning see moving im			
3	20061	film Name Modesty based around episode takes one page 10th modesty Blaise novel called Night Morningstar. des			
4	14772	never like comment good film comes bad movie, gotta come really hard it. Talking Vivah, guy, Sooraj Badjatya, seen			
5	21347	Woman wig, "dyes" hair middle film (=takes wig) presumably see audience see miles away: *** begin sp			
6	20498	totally wall good way, totally stupid. "Killer Tongue" uneasy mixture sci-fi, horror, supposed comedy. equates mindl			
7	24425	two points need make clear right beginning. First know year's Oscar's REALLY year. Academy's way showing people			
8	3042	seen Deliverance, movies like Pulp Fiction don't seem extreme. Maybe today's blood bullets standards doesn't seen			
9	18329	Like watching neighbor's summer camp home movies, "Indian Summer" sleep inducing bore. Eight alumni campers			
.0	11473	pleasure seeing film Rhode Island International Film Festival impressed. Tess Nanavati clearly great screenwriter wc			
.1	10309	movie passage manhood one gay man, must deal everyone. mother depressed, yo inger sister pain, older sister son			
.2	13535	going say worst gay-themed film I've ever seen, honestly say worst film genre I've ever seen. You know			
.3	20774	single positive film? Critics knew nothing video games could spot gaming errors made. damage taken damage clearl			
.4	20249	saw movie way back first theatrical release, justifiably empty theater. Believe not, decades watching movies, one st			
.5	22148	saw movie much younger thought funny. saw last week, guess result. funny parts it, long. beginning thing funny ask			
.6	22618	movies Lifetime anemic titles? "An Unexpected Love" - ooh, provocative!! "This Much know" would better. film not			
.7	11539	Strange yet emotionally disturbing chiller fed middle-aged man (William H. Macy) finally decides leave family busine			
.8	6702	Many reviews comments read movie say rather stale film performance Clara Bow. Although story-line rather typical			
.9	23150	Buster Keaton fan heart broken regular basis. us first encounter Keaton one brilliant feature films great period inde			
0	4212	films manage survive almost originality alone - "Wonderland" certainly one films. script manages throw everything			
1	14372	BDSM "sub-culture" Los Angeles serves backdrop low budget shabbily constructed mess, plainly vanity piece top-bil			
2	10588	...this classic many great dialogs scenes nobody miss. Nice story, funny riches-to-rags situations, Mel Brooks bad lea			
3	11856	Secret Fury, many ways run-of-the-mill romantic suspense drama (directed Mel Ferrer) boasts top-notch principals			

Step 2- Algorithms used:

For the purposed of this project unigram, bigram and tfidf algorithms are used to analyze the data.

Step 3- Classifier

Stochastic Gradient Descent classifier is used in this project to minimize the processing expense instead of gradient descent, as gradient descent tends to be expensive on large dataset.

Step 4-Analysis on training data:

Accuracy of unigram model is 92.604
Accuracy for the Bigram Model is 93.952
Accuracy for the Unigram TFIDF Model is 88.344
Accuracy for the Bigram TFIDF Model is 86.16

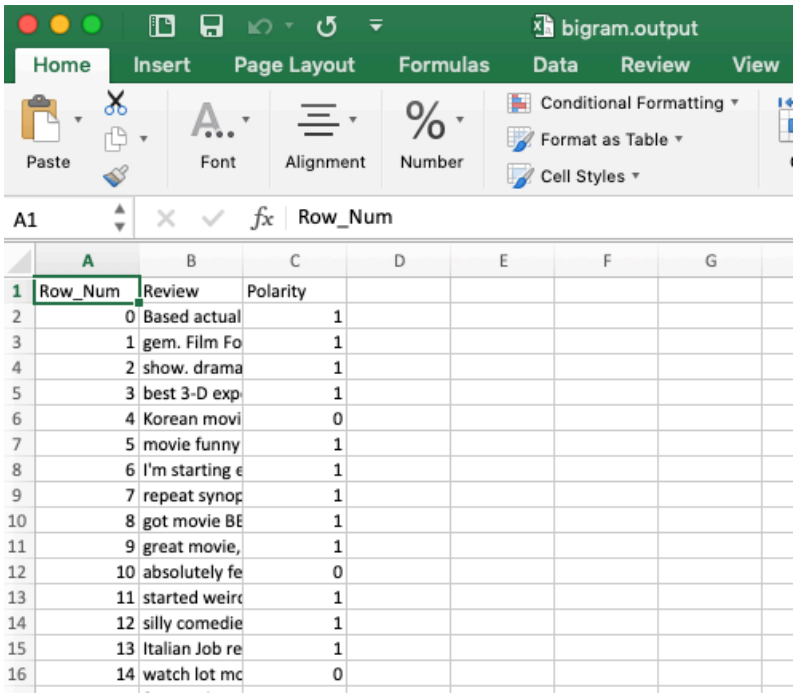
Step 5- Applying the classifier on testing data:

Four files are outputted:

1. bigram.output.csv
2. bigramtfidf.output.csv
3. unigram.output.csv
4. unigramtfidf.output.csv

Step 6- Reviewing the results:

Sample from bigram.output.csv, where the polarity has been predicted based on the testing data.



Row_Num	Review	Polarity
0	Based actual	1
1	gem. Film Fo	1
2	show. drama	1
3	best 3-D exp	1
4	Korean movi	0
5	movie funny	1
6	I'm starting e	1
7	repeat synop	1
8	got movie BE	1
9	great movie,	1
10	absolutely fe	0
11	started weird	1
12	silly comedie	1
13	Italian Job re	1
14	watch lot mo	0

Step 7: The code file and performance:

Code file:

https://github.com/samirk927/CS410/blob/master/IMDBSentimentAnalysis.py

Performance:

In my testing with 25K training and 25K testing dataset, the code ran for ~100 seconds.