1: Sentiment Classification

Machine Learning and Real-world Data (MLRD)

Simone Teufel

Lent 2021

This course: Machine Learning and Real-world Data (MLRD)

Goals of the course:

- Three different types of machine learning
  - Naive Bayes
  - Hidden Markov Models
  - Clique finding / clustering
- Straightforward approaches you can implement quickly and then experiment with
- Emphasis on methodology: relevant for all approaches.
- Coupling with Algorithms and Data structures
- Practical-based, but each session contains a short lecture introducing the main concepts.

Topics and Real-world Data

- Three Topics:
  - Classification according to sentiment (7 sessions)
  - Sequence analysis of proteins (4 sessions)
  - Network analysis of social networks (5 sessions)
- Plenty of data:
  - thousands of movie reviews
  - hundreds of amino acid sequences
  - thousands of users and links between them

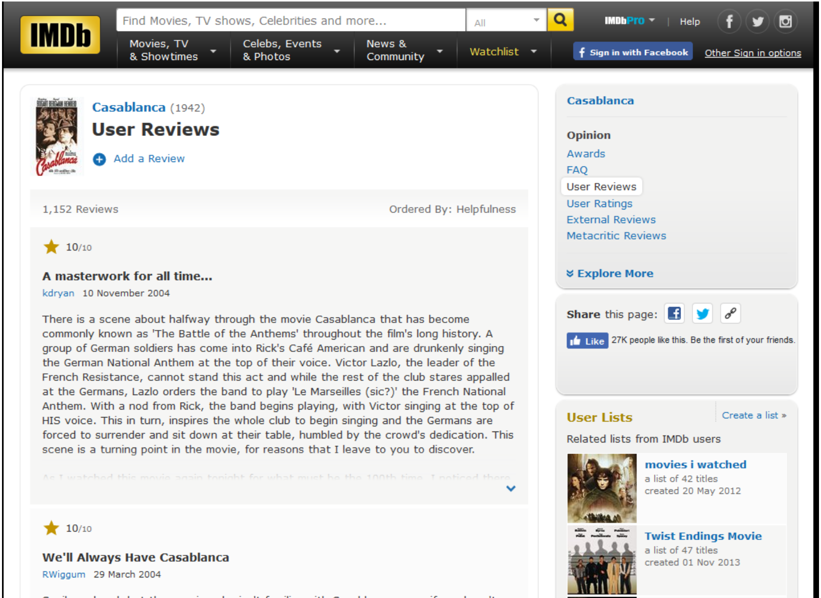
Computer Science as an empirical subject

- The style of solving tasks in this course is *empirical*.
- You will start from a hypothesis or an idea which you will test.
- Then you perform some manipulations on your data.
- You observe and record the results.
- You need a lab book to record your manipulations, observations and measurements.
    - physical book (many advantages) or electronic record
    - be prepared to show your lab book to your demonstrator

Topic 1: Sentiment classification

- IMDb (= Internet Movie Data Base) has about 4.7 million titles (`http://www.imdb.com/pressroom/stats/`).
- Reviews: written in natural language by the general public.
- **Sentiment classification** — the task of automatically deciding whether a review is positive or negative, based on the text of the review.
- Standard task in **Natural Language Processing (NLP)**.
- The evaluative language used is interesting from a linguistic viewpoint.

IMDb



Casablanca (1942)

User Reviews

Add a Review

1,152 Reviews | Ordered By: Helpfulness

★ 10/10

A masterwork for all time...

kdryan 10 November 2004

There is a scene about halfway through the movie Casablanca that has become commonly known as 'The Battle of the Anthems' throughout the film's long history. A group of German soldiers has come into Rick's Café American and are drunkenly singing the German National Anthem at the top of their voice. Victor Lazlo, the leader of the French Resistance, cannot stand this act and while the rest of the club stares appalled at the Germans, Lazlo orders the band to play 'Le Marseilles (sic?)' the French National Anthem. With a nod from Rick, the band begins playing, with Victor singing at the top of HIS voice. This in turn, inspires the whole club to begin singing and the Germans are forced to surrender and sit down at their table, humbled by the crowd's dedication. This scene is a turning point in the movie, for reasons that I leave to you to discover.

As I watched this movie again tonight for what must be the 100th time, I noticed there

★ 10/10

We'll Always Have Casablanca

RWiggum 29 March 2004

Casablanca

Opinion

Awards

FAQ

User Reviews

User Ratings

External Reviews

Metacritic Reviews

Explore More

Share this page:

Like 27K people like this. Be the first of your friends.

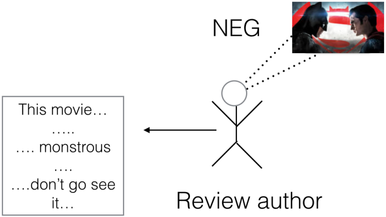
User Lists Create a list »

Related lists from IMDb users

movies i watched
a list of 42 titles
created 20 May 2012

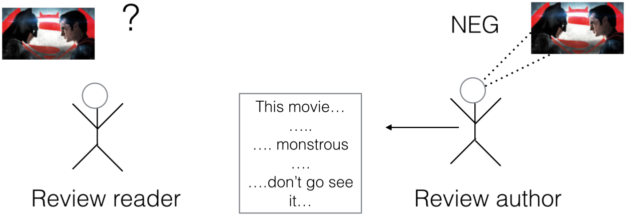
Twist Endings Movie
a list of 47 titles
created 01 Nov 2013

Review sentiment



This movie…
…..
…. monstrous
….
….don't go see it…

Review sentiment



?

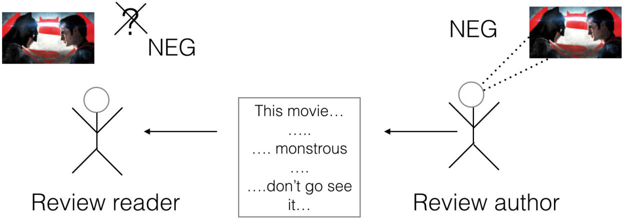
NEG

This movie…
.....
…. monstrous
….
….don't go see it…

Review reader

Review author

Review sentiment



~~?~~ NEG

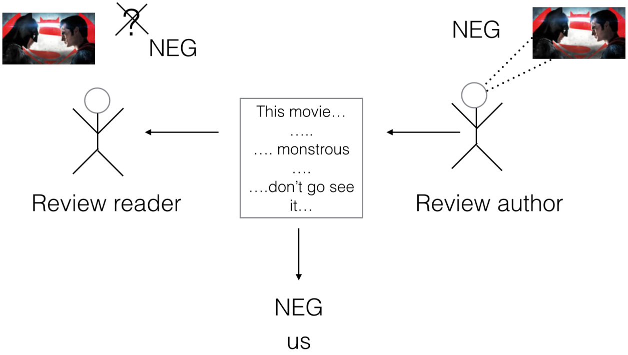
NEG

This movie…
…..
…. monstrous
….
….don't go see
it…

Review reader

Review author

Review sentiment



~~?~~ NEG

NEG

This movie…
…..
…. monstrous
….
….don't go see it…

Review reader

Review author

NEG

us

From a good review

... He’s incredible in fights. ... Also his relationship with Irons, who plays Alfred, is just wonderful in general. Irons was exceptional in the role.

A bad review

This movie tries so hard... It completely fails on every single level. The movie is tedious and boring with characters that I just did not care about at all. ...

Experiments with movie reviews

- Lots of possible NLP experiments . . .
- Today: use data about individual words to find sentiment.
  - Sentiment **lexicon** lists over 8000 words as positive or negative.
  - Hypothesis: a review that contains more positive than negative words is positive overall.

Experiments with movie reviews

- Lots of possible NLP experiments . . .
- Today: use data about individual words to find sentiment.
  - Sentiment **lexicon** lists over 8000 words as positive or negative.
  - Hypothesis: a review that contains more positive than negative words is positive overall.

```
word=foul intensity=weak polarity=negative
word=mirage intensity=strong polarity=negative
word=aggression intensity=strong polarity=negative
word=eligible intensity=weak polarity=positive
word=chatter intensity=strong polarity=negative
```

Note: a lexicon is a list of words with some associated information.

Sentiment lexicon words in the good review

... He’s incredible in fights. ... Also his relationship with Irons, who plays Alfred, is just wonderful in general. Irons was exceptional in the role.

- incredible positive
- wonderful positive
- exceptional positive

Sentiment lexicon words in the bad review

This movie tries so hard... It completely fails on every single level. The movie is tedious and boring with characters that I just did not care about at all. ...

- try negative
- fail negative
- tedious negative
- boring negative
- care positive

But it doesn’t always work . . .

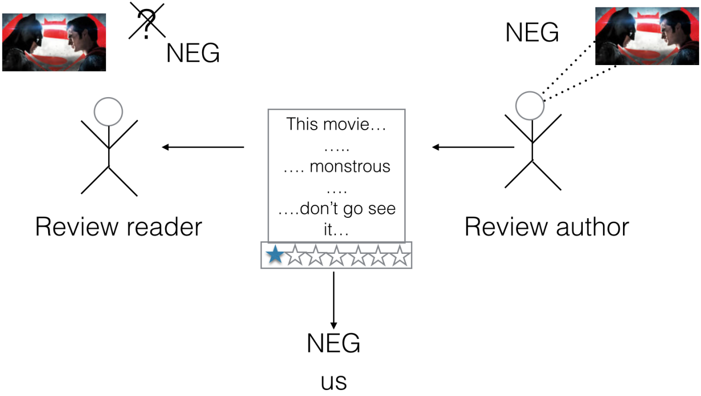
This movie tries so hard... The ending should be exciting and fun and amazing.. and it just... wasn’t. It completely fails on every single level. The movie is tedious and boring with characters that I just did not care about at all. ...

- try negative
- exciting positive
- fun positive
- amazing positive
- fail negative
- tedious negative
- boring negative
- care positive

Evaluation

- No system predicts sentiment perfectly.
- How do we know the extent to which we've got it right?
- The author of the review told us the truth explicitly via a star rating (that's why NLP researchers like movie reviews).
- The rating has been extracted along with the review text.
- We will calculate a metric called A (accuracy).

Star rating



~~?~~ NEG

NEG

Review reader

This movie…
.....
…. monstrous
….
….don't go see it…

★☆☆☆☆☆☆

Review author

NEG

us

Accuracy

- The number of correct decisions c divided by total decisions (correct plus incorrect (i)):

$$A = \frac{c}{c+i}$$

- This metric is called A (accuracy).
- We know which decisions are “correct” because we can use the star rating as our definition of truth.

Tokenisation: getting the words out

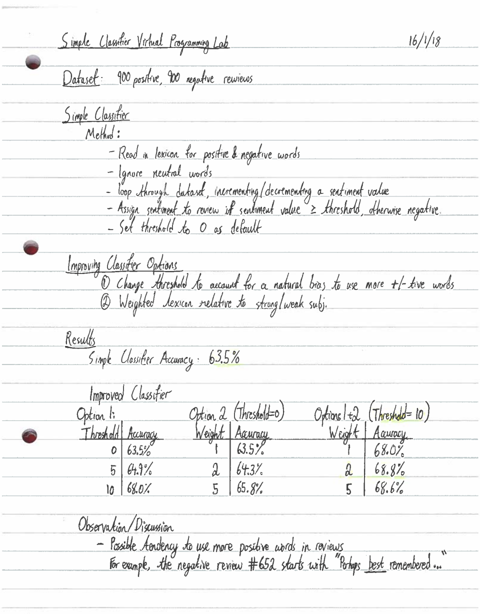
- Your code will look up words from your review document in the lexicon.
- So it needs to divide the text into words.
- Splitting on whitespace is not enough.
  - Words at the beginning of a sentence appear in upper case.
  - Words occurring before and after punctuation may be directly attached to the punctuation.
  - and many other things . . .
- Your code will use a well-known basic **tokeniser** to split the text into individual words.
- Note: **type** vs **token** (see ‘Further notes’ in Session 2)

Your tasks for today

Task 1:

- explore the review data (1800 documents)
- make judgment about sentiment of 4 reviews
- explore the sentiment lexicon
- guess 10 sentiment-indicating words
- write a program that tests the sentiment lexicon approach
- write a program for using the star ratings to evaluate how well your program is doing
- and keep a record of what you do

Example lab book page



Simple Classifier Virtual Programming Lab 16/1/18

Dataset: 900 positive, 900 negative rewiews

Simple Classifier

Method:

- Read in lexicon for positive & negative words
- Ignore neutral words
- loop through dataset, incrementing/decrementing a sentiment value
- Assign sentiment to review if sentiment value ≥ threshold, otherwise negative.
- Set threshold to 0 as default

Improving Classifier Options

① Change threshold to account for a natural bias to use more +/- tive words

② Weighted lexicon relative to strong/weak subj.

Results

Simple Classifier Accuracy: 63.5%

Improved Classifier

Option 1:

Threshold	Accuracy
0	63.5%
5	64.9%
10	68.0%

Option 2 (Threshold=0)

Weight	Accuracy
1	63.5%
2	64.3%
5	65.8%

Options 1+2 (Threshold= 10)

Weight	Accuracy
1	68.0%
2	68.8%
5	68.6%

Observation/Discussion

- Possible tendency to use more positive words in reviews
  For example, the negative review #652 starts with "Perhaps best remembered..."

Practicalities

- 16 lectures (approx 25 minutes) – live at 2:05 [Mo, Fr] or watch video before 2:30pm
- 16 demonstrated Zoom sessions: from 2:30pm to 4:30pm [Mo, Fr]
- 12 tasks and 4 catch-up sessions
- 12 ticks: you should get them all
- Interactive sessions for help, questions and ticking
- Most tasks have automated tester: pass this first!
- Ticking session with some screensharing
- Tester will close on midnight of deadline. Unlock via DOS.
- Students out of Timezone – arrange ticking and help via email.
- Lots more on Moodle . . .

Sessions, Ticks, Deadlines

Session	Date	Tick	Task	Tester Deadline
S1	22/1	T1	Sentiment Lexicon	11/2
S2	25/1	T2	NB	15/2
S3	29/1	T3	Zipf	11/2
S4	1/2	T4	Sign Test	15/2
S5	5/2	T5	CrossVal	19/2
S6	8/2	T6	Kappa	22/2
S7 (catch up)	11/2	–	–	–
S8	15/2	T7	HMM Training	1/3
S9	19/2	T8	Viterbi	5/3
S10	22/2	T9	Proteins	8/3
S11 (catch up)	26/2	–	–	–
S12	1/3	T10	Network Properties	15/3
S13	5/3	T11	Brandes' Algo	15/3
S14	8/3	T12	Clustering	15/3
S15 (catch up)	12/3	–	–	–
S16 (catch up)	15/3	–	–	–

Last Chance Session	30/4			