1: Sentiment Classification
Machine Learning and Real-world Data (MLRD)

Simone Teufel
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 (and later ML courses)

- 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 strongly recommended
  - be prepared to show your lab book to your demonstrator
Simple Classifier Virtual Programming Lab 16/1/18

Dataset: 900 positive, 900 negative reviews

Simple Classifier
Method:
- Read a lexicon for positive & negative words
- Ignore neutral words
- Loop through dataset, incrementing/deducing a sentiment value
- Assign sentiment to review if sentiment value ≥ threshold, otherwise negative
- Set threshold to 0 as default

Improving Classifier Options
- Change threshold to account for a natural bias to use more +/−ive words
- Weighted lexicon relative to strong/weak subj.

Results
Simple Classifier Accuracy: 63.5%

Improved Classifier

<table>
<thead>
<tr>
<th>Option 1</th>
<th>Option 2 (Threshold=0)</th>
<th>Option 1 (Threshold=10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>Weight</td>
<td>Accuracy</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>63.5%</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>64.3%</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>65.8%</td>
</tr>
</tbody>
</table>

Observation/Discussion
- Possible Academy to use more positive words in reviews
  (Example, the negative review #652 starts with "Perhaps best remembered..."
Practicalities

- Lectures (approx 25 minutes) – in LT1 at 2:05 [Mo, Fr]
- 16 demonstrated lab sessions in Intel Lab: from 2:30pm to 4:30pm [Mo, Fr]
- 12 tasks and 4 catch-up sessions
- You must do all tasks
- This means passing the automatic tester for those tasks where there is one
- There is also additional ticking for some of you
Ticks

- Random generator decides who gets additionally ticked.
- If you have been selected...
  - Getting a tick means passing the automatic tester and then having a personal ticking session with a demonstrator
  - Pass automated tester before booking ticking sessions
- You can also ask to be ticked, even if you are not selected.
- Best learning effect: get each tick as soon as possible
- Normal expectation: get each tick triple by the deadline.
- Ticks are bundled into triples
- You can get up to three ticks in one ticking session
Lab sessions

- Lab sessions are there for help, questions and ticking
- Online ticking session for those with valid reasons (eg sickness), talk to student administration; your may need your DoS to support you.
- All info on practical is on Moodle . . .
Soft deadlines

- Deadline for each triple of tick: 1 week after announcement of last tick in triple
- We will use Moodle to announce who has been selected
- Announcements on the day when tasks 3, 6, 9 and 12 are released.
- Deadlines are soft: Your DoS can check to see your progress
- People who are sufficiently late are going to get ticked (irrespective of random selection).
- Everybody needs to get each task pretested or ticked, so no consequences of being late (as long as you catch up)
<table>
<thead>
<tr>
<th>Session</th>
<th>Date</th>
<th>Tick</th>
<th>Task</th>
<th>Soft Deadline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>F 20/01</td>
<td>T1</td>
<td>Sentiment Lexicon</td>
<td>03/02</td>
</tr>
<tr>
<td>S2</td>
<td>M 23/01</td>
<td>T2</td>
<td>NB</td>
<td>03/02</td>
</tr>
<tr>
<td>S3</td>
<td>F 27/01</td>
<td>T3</td>
<td>Zipf</td>
<td>03/02</td>
</tr>
<tr>
<td>S4</td>
<td>M 30/01</td>
<td>T4</td>
<td>Sign Test</td>
<td>13/02</td>
</tr>
<tr>
<td>S5</td>
<td>F 03/02</td>
<td>T5</td>
<td>CrossVal</td>
<td>13/02</td>
</tr>
<tr>
<td>S6</td>
<td>M 06/02</td>
<td>T6</td>
<td>Kappa (catch up)</td>
<td>13/02</td>
</tr>
<tr>
<td>S7</td>
<td>F 10/02</td>
<td>–</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>Topic 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>M 13/02</td>
<td>T7</td>
<td>HMM Training</td>
<td>27/02</td>
</tr>
<tr>
<td>S9</td>
<td>F 17/02</td>
<td>T8</td>
<td>Viterbi</td>
<td>27/02</td>
</tr>
<tr>
<td>S10</td>
<td>M 20/02</td>
<td>T9</td>
<td>Proteins (catch up)</td>
<td>27/02</td>
</tr>
<tr>
<td>S11</td>
<td>F 24/02</td>
<td>–</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>Topic 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S12</td>
<td>M 27/02</td>
<td>T10</td>
<td>Network Properties</td>
<td>13/03</td>
</tr>
<tr>
<td>S13</td>
<td>F 03/03</td>
<td>T11</td>
<td>Brandes’ Algo</td>
<td>13/03</td>
</tr>
<tr>
<td>S14</td>
<td>M 06/03</td>
<td>T12</td>
<td>Clustering</td>
<td>13/03</td>
</tr>
<tr>
<td>S15</td>
<td>F 10/03</td>
<td>–</td>
<td>(catch up)</td>
<td>–</td>
</tr>
<tr>
<td>S16</td>
<td>M 13/03</td>
<td>–</td>
<td>(catch up)</td>
<td>–</td>
</tr>
</tbody>
</table>
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.
Casablanca (1942)

User Reviews

1,152 Reviews    Ordered By: Helpfulness

⭐️ 10/10

A masterwork for all time...

kdrayn    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.

⭐️ 10/10

We'll Always Have Casablanca

RWiggum    29 March 2004

So cliched but the movie lives on. Great score, great acting, great story. It can't be beat.
Review sentiment

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

NEG

Review author
Review sentiment

Review reader

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

Review author

NEG
Review sentiment

This movie...
..... monstrous
..... don’t go see it...

Review reader

Review author
Review sentiment

This movie…
   … monstrous
   … don’t go see it…

Review reader

Review author

NEG

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*
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

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

Review reader

Review author

NEG

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 always keep a record of what you do