Text Mining in Practice

Ralf Krestel, Julian Risch

Hasso Plattner Institute, University of Potsdam

11/04/2018

News Comments in the Wild

wow great insight bro, you should write for lonely planet. (/sarcasm)

What a well written response!

Well said!

A very mental ability? What was I drinking? I meant, of course, to write that I can't conceive of someone being allowed to join the ranks of our post-people without demonstrating great intellectual prowess, so Mr Berchmans' lack of intellectual rigour can only be a recent thing.

Correct. I means go fuck yourself.

These people work the long hours because each hour makes them richer. A medic or nurse works long hours so their patients do not die due to understaffing

[## Source or it isn't true##] Im comfortable with it...you prove me wrong. Sarka did. I can be beaten in argument .Just not often! :) B

Helen121I agree with you that ColinwithanM's comment is flawed. Honour killing is not specifically a phenomenon of migration....there's a huge amount of it (as Fisk shows) in Pakistan...probably proportionately more than among Pakistani-immigrants and their descendants in the UK. On its relationship to faith though, it's clear that Christian communities in the ME are not exempt, but I rather doubt that the recorded higher rates among Muslims in Jordan can really mean that Christians are hugely better at completely concealing the crimes than Muslims are. A West Bank Palestinian sociologist speaking on the World Service last year said a) that basically the same honour values were comon to both Muslim and Christian Palestinian families - i.e. this was common Arab culture not specifically faith culture, but that the Muslims tended to be more likely to beat and murder (as opposed to e.g. ostracising) offending women As a Muslim himself, he found this dismaying, and was at a loss to explain it...I wouldn't want to jump in with any crass explanations where he held back...It's obviously not a single issue if faith, but a very complicated one of interaction between religious ethos and the rest of the culture. Another interesting example is FGM in Egypt. Christians traditionally practice it as do Muslims. Muslim and Christian campaigners against the practice have noted, though, that it is much harder to persuade Musims to give it up long-term than to persuade Christians. This suggests that at present - for whatever reasons - Egyptian Muslims feel that the practice is more intensely bound up with their religious communal identity than Egyptian Christians do.... which is BTW by no means the same thing as saying that Islam essentially mandates it in a way that Christianity does not. Another important thing about honour killing is that (as far as I know), perpetrators are not more common among the more religious. Perpetrators can easily be personally not specially religious at all (not enthusiastic mosque attenders, not religiose, lax) while some highly religiose Muslims, including local imams, have been whistle-blowers against the practice. Sense of obligation to kill for honour is very much a social thing, not a conscience thing in the strictly religious sense.

https://www.theguardian.com/

What you will Learn in this Seminar

- How to do text mining, obviously
  - Design, **implement**, evaluate software in a small team
  - Sell your work:
    Present problems and solutions to your peers
  - Document your (research) work:
    Write a scientific paper

https://c1.staticflickr.com/5/4492/37089792263_714e73ae24_b.jpg
https://upload.wikimedia.org/wikipedia/commons/2/26/Microsoft_Surface_presentation.jpg
https://upload.wikimedia.org/wikipedia/commons/1/1b/Aale_Tynni_writing_with_a_typewriter.jpg

Organization

- Grading
  - 20% Mid-Term Presentation
    - Present your problem and ideas how to solve it
  - 20% Final Presentation
    - Present your final solution and your evaluation
  - 30% Implementation
    - Exciting text mining methods in Python
  - 30% Paper (4 pages, ACM Template)
    - Prepare for your Master‘s thesis

http://blog.newrelic.com/wp-content/uploads/shutterstock_174281696_python-e1420856395184.jpg

Schedule

11.04.18	Introduction I	Ralf Krestel, Julian Risch
18.04.18	Deep Learning Tutorial (optional)	Julian Risch
25.04.18	Introduction II	Julian Risch
02.05.18		
09.05.18	Discussion	Students
16.05.18		
23.05.18		
30.05.18	Mid-Term Presentations	Students
06.06.18		
13.06.18		
20.06.18		
27.06.18	Writing Research Papers	Ralf Krestel
04.07.18		
11.07.18		
18.07.18	Final Presentations	Students
01.08.18	Paper Submission Deadline	

https://hpi.de/naumann/teaching/teaching/ss-18/text-mining-in-practice-ps-master.html

Previously on... HPI Season 2016-2018

- Master's theses:
  - Classification of German Newspaper Comments (Christian Godde)
  - Detection of Inappropriate Content in Online Comments (Dustin Gläser)
  - Automatically Managing News Comments (Carl Ambroselli)
- Master's project: Hate Speech Detection
- Lecture: Information Retrieval and Web Search (News Comment Search Engine)
- Project seminar: **Text Mining in Practice**

open.hpi.de/assets/startpage_visual.png

Different Tasks of News Comment Analysis

- Data-centric: news articles and their received comments

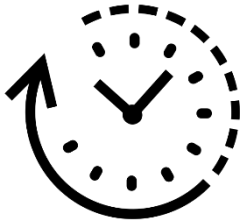
hate speech detection

volume prediction

comment ranking

modeling user behavior

discussion summarization



modeling linguistic change

Hate Speech Detection

- Semi-automatic comment moderation for the newsroom
- Cooperation with an online newspaper: labeled dataset
- Need for explanations
- Recall more important
- Logistic regression trained on
  - Comment features
  - User features
  - Article features

Number of Comments	3,132,917	100.00%
Appropriate	3,034,125	96.85%
Inappropriate	98,792	3.15%
Number of Users	59,377	
Number of Articles	26,391	

Hate Speech Detection
Replies to Moderators

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

This newspaper more and more degrades to a fun fair of personal sensitivities, which have nothing to do with journalism.

Hate Speech Detection
Replies to Moderators

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

edited

“please comment on the article topic”

Hate Speech Detection
Replies to Moderators

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

edited

“please comment on the article topic”

If the article only had a topic, then I could discuss it.

Hate Speech Detection
Replies to Moderators

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

edited

"please comment on the article topic"

If the article only had a topic, then I could discuss it.

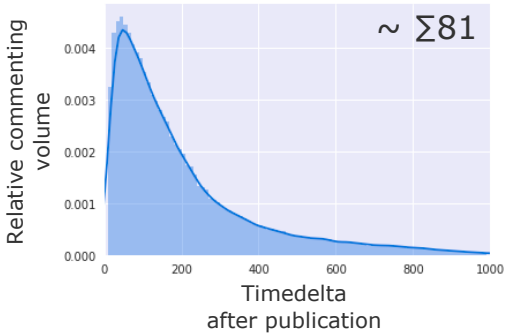
Toxic Comment Classification

- Toxic comment classification challenge on Kaggle
- What's toxic? "A rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."
- Six classes
  - toxic: Bye! Don't look, come or think of coming back!
  - severe toxic: You should die from cancer.
  - obscene: …
  - threat: I am going to … you
  - insult: You are a ...
  - identity hate: All the … are …

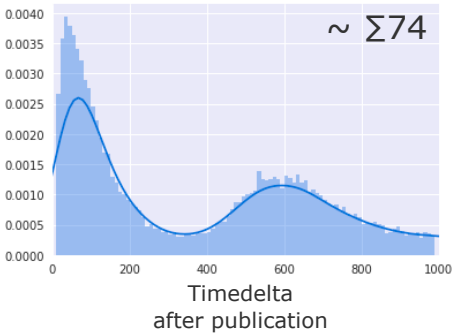
Comment Volume Prediction

- Predict how many comments a news article will receive
- Helps moderators to schedule their work
- Transfer learning approach on English and German comments

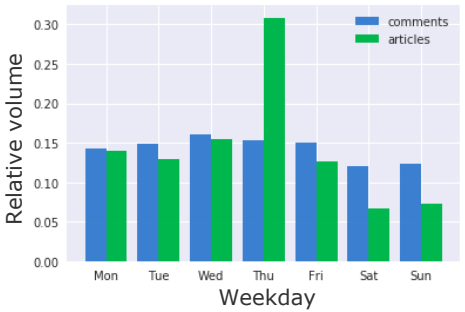
Article published
08:00-09:00



Article published
21:00-22:00



Article published
weekday



Comment Ranking

- Pagination favors the first ten users.
- First comment has strong influence on the direction of the discussion
- What is the goal of re-ranking the comments?
  - engaging, respectful, informative discussions (Napoles et al.)

Discussion Summarization

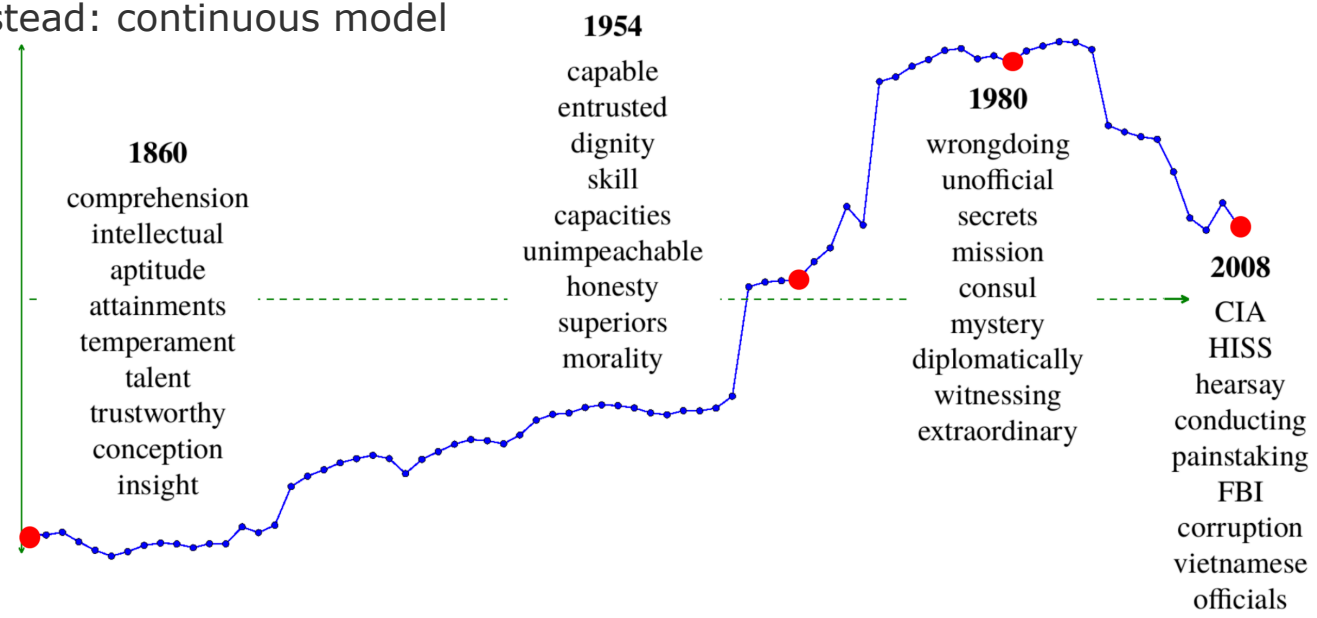
- Summarize not news stories but their discussions
  - discussions‘ subtopics
  - different arguments
  - comments with many replies
  - user‘s known for high-quality comments
  - upvotes / downvotes

Modeling User Behavior

- Detect users who team up to attack a discussion
- Prediction:
  - comment volume
  - sentiment of a particular user‘s comment
  - weekday and time

Modeling Linguistic Changes

- No separate embeddings on slices of the data
- Instead: continuous model

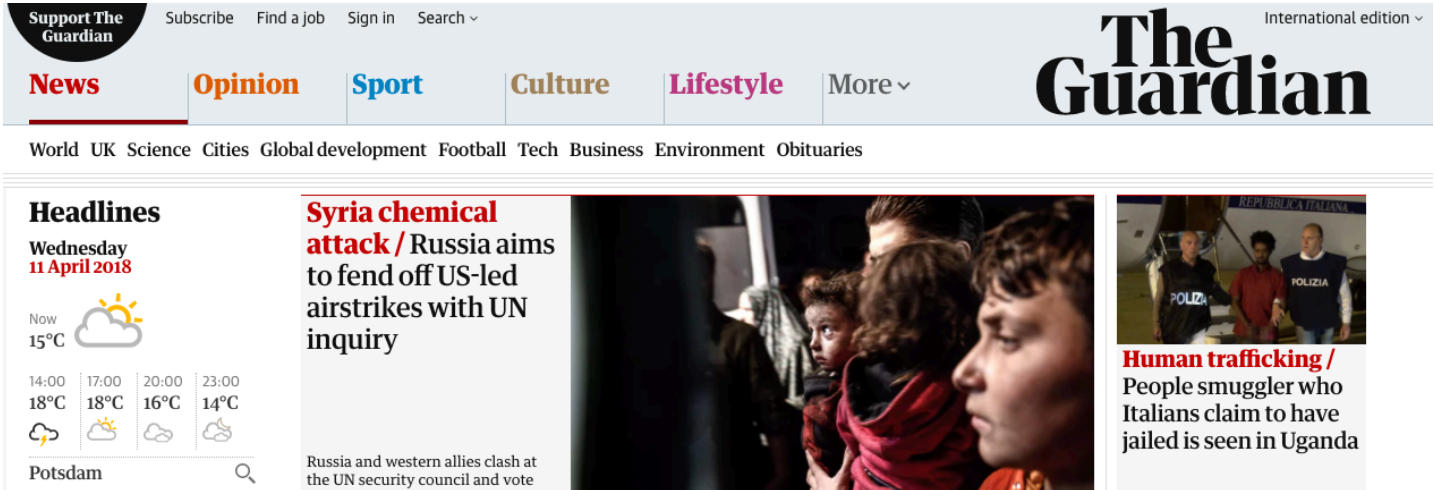


(b) INTELLIGENCE in U.S. Senate speeches (1858–2009)

Rudolph and Blei: Dynamic Bernoulli Embeddings for Language Evolution

Dataset
Guardian Online Comment Corpus

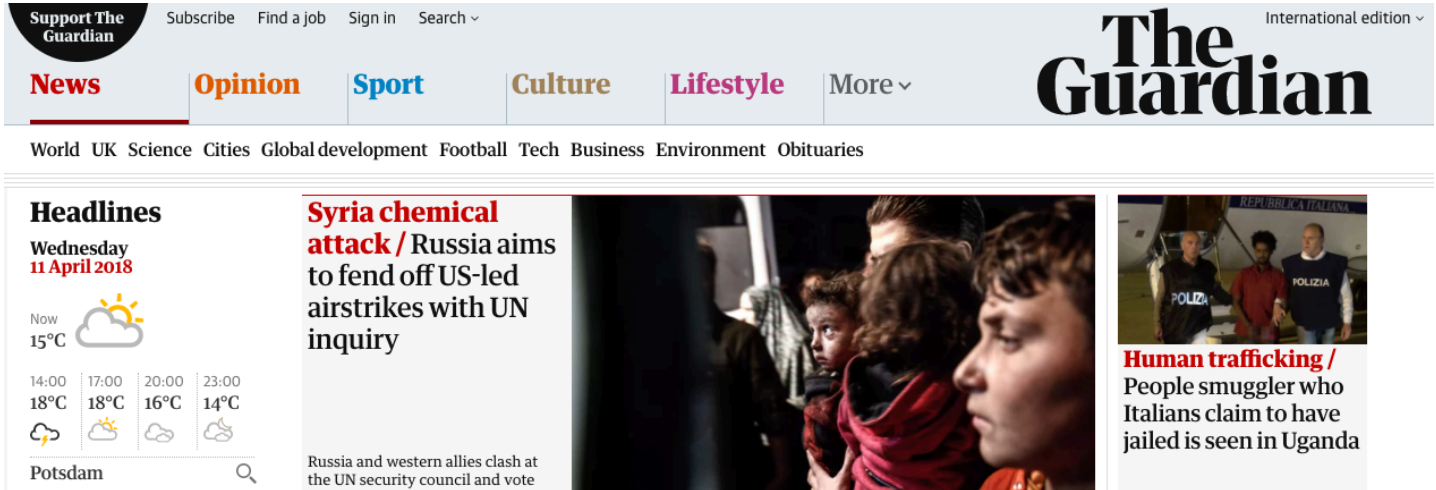
- 60 million user comments
- by 1 million users
- to 600,000 articles



https://www.theguardian.com/

Dataset
Guardian Online Comment Corpus – 100,000

- 100,000 user comments
- 100,000 user names
- 100,000 article URLs
- available here: https://owncloud.hpi.de/index.php/s/UfRMLZksVugnwky



https://www.theguardian.com/

Dataset
Guardian Online Comment Corpus – 100,000

- Articles

```
article_id,article_url
1,https://www.theguardian.com/commentisfree/2012/feb/09/women-at-work-equa
2,article_url
3,https://www.theguardian.com/business/2012/feb/09/britain-boardrooms-wome
4,https://www.theguardian.com/travel/2012/feb/08/copenhagen-best-restaurar
```

- Authors

```
author_id,comment_author
1,rosieh2
2,asgoodasitgets
3,Mendoza
4,IndependentBrain
```

Dataset
Guardian Online Comment Corpus – 100,000

- Comments

```
article_id,author_id,comment_id,comment_text,parent_comment_id,timestamp,upvotes
1,1,14606180,"So you are saying that the demonisation of feminists is their own
y father out of his burning house is a woman or a man as long as that person is
tion, it being fundamental to the idea of equality. who no matter what they say
 no-one was suggesting it here. ...and that i only want ot be with my children be
f you accept that women want to be with their children more than men do, in gener
o you; neither entails calling you stupid. People will demonise anything that app
 not demonised for calling men rapists. They are demonised for holding perfectly
hat women and man should be treated equally. Men are criticised for not conformi
ghts' . How that any different? It is not different. Why should anyone have to co
er and not wanting to pursue a career. Why should you not respect the position th
 can be the primary caregiver just as easily as a woman can?  What caring roles
 evasive. Again, you are misrepresenting what I said. Nobody is forced into a car
t to positions in influential spheres, e.g. law, banking, political office, neuro
tled to. That is my problem.",,2012-02-11T12:49:42Z,0
```

Dataset
Explorative Data Analysis

- Pandas

```
>>> import pandas as pd
>>> df = pd.read_csv("sorted_comments-1000.csv")
>>> print df.head(3)
   article_id  author_id comment_id  \
0           1          1   14606180
1           1          1   14605967
2           1          1   14605172

                                        comment_text parent_comment_id  \
0  So you are saying that the demonisation of fem...               NaN
1  So you are explicitly saying that it is not ok...          14605336
2  So we have given up on the reasoned debate the...          14605044

              timestamp upvotes
0  2012-02-11T12:49:42Z       0
1  2012-02-11T12:29:29Z       0
```

Dataset
Explorative Data Analysis

- Pandas

```
>>> test = df.groupby(['article_id'])
>>> test.describe()
            author_id
                count
article_id
1               115.0
2                 1.0
3                97.0
4                21.0
5                41.0
6                59.0
7               126.0
8                10.0
9               119.0
10              316.0
11               94.0
```

Dataset
Explorative Data Analysis

- Pandas

```
>>> df[df['comment_text'].str.len() < 10]
     article_id  author_id comment_id comment_text parent_comment_id  \
220           4        132   14564533       Hello!               NaN
368           7        229   14491840    EXACTLY !          14491064
401           7        242   14489071    Here here          14488402
535           9        336   14476706            .          14475792
559           9        355   14479252         Why?          14475821
571           9        365   14476773         Amen          14475586
635          10        390   14478187     /shudder               NaN
```

Dataset
Explorative Data Analysis

- Pandas

```
>>> df[df['comment_text'].str.contains('hate speech')]
     article_id  author_id comment_id  \
436           7        267   14490363
912          11        392   14447379

                                          comment_text parent_comment_id  \
436  A couple of years ago, when I first became awa...               NaN
912  [ ..especially every time some new Muslim coun...          14447298

                timestamp upvotes
436  2012-02-04T12:00:07Z       4
912  2012-02-01T18:02:11Z       7
```

Dataset
Explorative Data Analysis

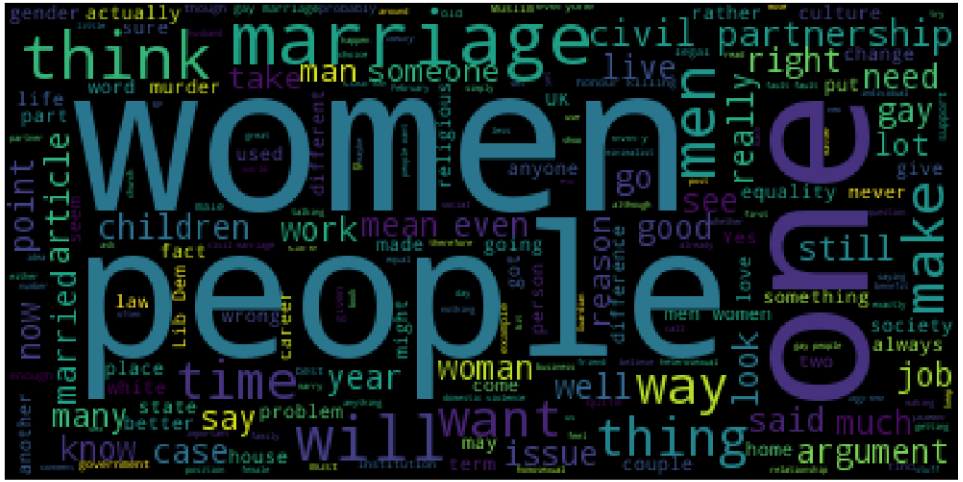
- Pandas

```
>>> import pandas as pd
>>> df = pd.read_csv("sorted_comments-1000.csv")
>>> from wordcloud import WordCloud
>>> wordcloud2 = WordCloud().generate(' '.join(df['comment_text']))
>>> import matplotlib.pyplot as plt
>>> plt.imshow(wordcloud2)
<matplotlib.image.AxesImage object at 0x114951510>
>>> plt.axis("off")
(-0.5, 399.5, 199.5, -0.5)
>>> plt.show()
```

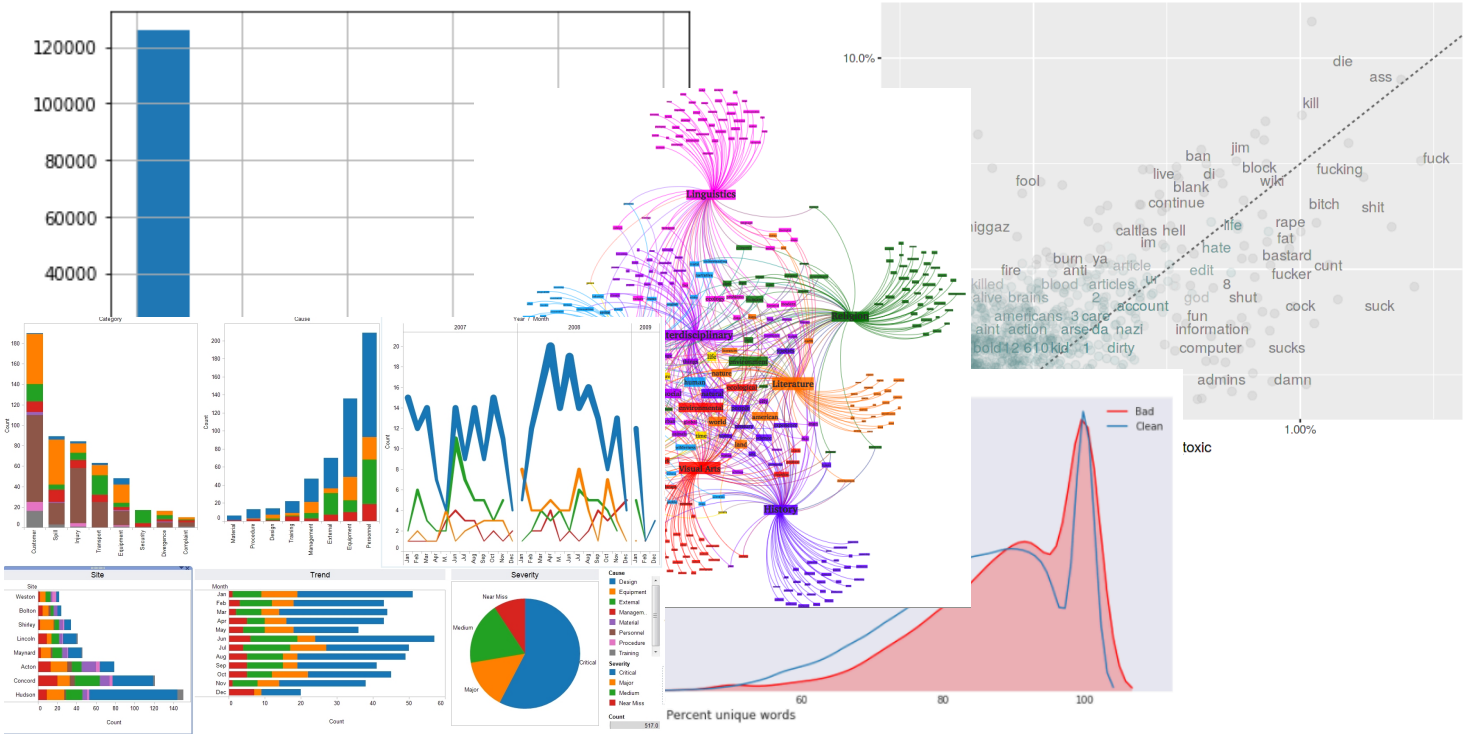
Dataset
Explorative Data Analysis

- Pandas

```
>>> import pandas as pd
>>> df = pd.read_csv("sorted_comments-1000.csv")
>>> from wordcloud import WordCloud
>>> wordcloud2 = WordCloud().generate(' '.join(df['comment_text']))
>>> import matplotlib.py
>>> plt.imshow(wordcloud
<matplotlib.image.AxesIm
>>> plt.axis("off")
(-0.5, 399.5, 199.5, -0.
>>> plt.show()
```



Your Task:
Explorative Data Analysis



Text Mining in Practice

Ralf Krestel, Julian Risch 11/04/2018

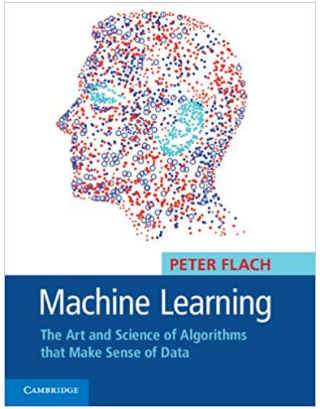
What do you know about…
Machine Learning and Deep Learning?

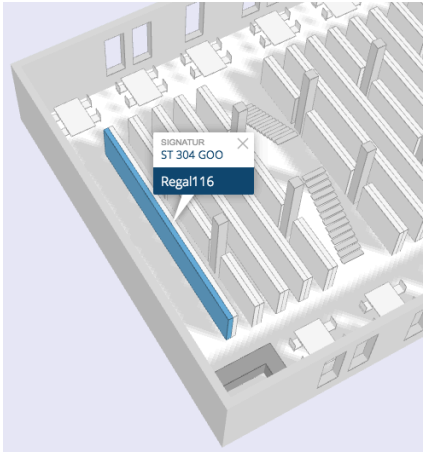
- Training set, validation set, test set
- Logistic Regression

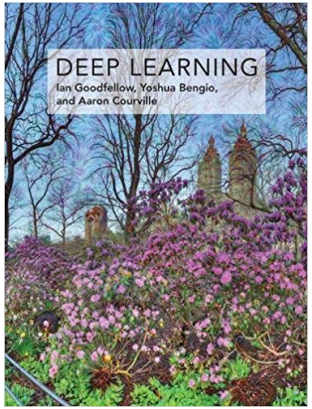
- Tensorflow, Keras, PyTorch
- Word2Vec, FastText
- Convolutional Neural Nets, Recurrent Neural Nets, Attention Layers

References

- Machine Learning: The Art and Science of Algorithms that Make Sense of Data by Peter Flach
- Deep Learning by Ian Goodfellow, Yoshua Bengio, Aaron Courville







What's next?

- **18.04.** optional Deep Learning Tutorial
- **20.04.** "Belegungsfrist"
- **25.04.** Team Matching, Topic Matching,
  Explorative Data Analysis (share your results)

Text Mining in Practice

Ralf Krestel, Julian Risch

Hasso Plattner Institute, University of Potsdam

11/04/2018