UNIVERSITY OF MANNHEIM

Web Mining

Introduction to the Web Mining Projects (IE 684)

Christian Bizer / Simone Ponzetto

Keti Korini / Alexander Brinkmann

FSS 2024

Outline

1. Information about Final Exam (IE671)
2. Introduction to the Web Mining Projects (IE684)
3. Group Formation (IE684)

1. Information about Final Exam (IE671)

- **Date: June 7th; Duration: 60 minutes; 3 ECTS**

- **3 blocks of questions on Web Usage Mining, Web Structure Mining, Web Content Mining**
  - 10 points per block, several questions per block

- **Content: open questions that**
  - check whether you have understood the content of the lecture
    - we try to cover all major chapters of the lecture, including recommender systems, network metrics, community detection, machine learning on graphs, sentiment analysis, named entity recognition
  - require you to describe the ideas behind algorithms or apply the methods
    - What is the advantage or problem of X compared to Y?
    - How do methods react to this special pattern in the data?
    - Given the following data/graph. Please calculate ….
  - might require you to do some simple calculations
    - you need to be able to use the most relevant formulas
    - you are not allowed to use a calculator (so only simple formulas can be applied)

2. Introduction to the Student Projects

- **Goals**
  - Gain practical experience on the topics that we have covered in the lecture:
    1. **Web Usage Mining** (including Recommender Systems)
    2. **Web Structure Mining** (including Social Network Analysis, Machine Learning on Graphs)
    3. **Web Content Mining** (including Sentiment Analysis, Hate Speech Detection, Named Entity Recognition)
  - Get to know additional current tools and methods

- **What is expected from you**
  - To find an interesting Web mining problem of your choice
  - To find a solution for the problem using
    - any of the Web mining methods that we have seen so far plus some additional task-specific techniques
    - other Web mining methods which might be helpful for solving the problem and build on what we learned in class

Overview

- **Teams of five students**
  1. realize a Web mining project
  2. write 12-page report about the project and the methods employed in the project
  3. present the project results to the other students (10 minutes presentation + 5 minutes discussion)

- **Final mark for the course**
  - 70 % project report (including code)
  - 30 % oral presentation

Schedule

Week	Topic / Deadline
19.03.2024	Kickoff Session and Team Formation
12.04.2024, 23:59	Submission of project outlines
18.04.2024, 13:45	Feedback on the project outlines (if necessary)
30.04.2024	Coaching session
07.05.2024	Coaching session
14.05.2024	Coaching session
17.05.2024, 23:59	Submission of project reports
21.05.2024, 10:15	Presentation of project results
07.06.2024	Final exam

Step 1: Team Formation

- You can form a team with other students of your choice
  - Each team must consist of 5 students
- If you do not find a team yourself, we will assign you to a team in the kickoff session
- Process:

1. Find 5 fellow students you want to do the project with
2. Send Keti and Alex a mail with your preferred team or with a request that you are looking for a team till Thursday the 21st of March 2024.

- People who do not have a team
  - will be assigned to existing teams or grouped into new teams by Friday the 22nd of March 2024

Step 2: Project Outlines

- **Write 3 pages (sharp!) project outline**
  - include a project name and your team number on the first page
  - using Springer Computer Science Proceedings layout or Word

- **Send the project outline until 12.04.2024, 23:59 via mail to Keti and Alex**

- **The project outline needs to answer the following questions:**

1. **What is the problem you are solving?**
2. **What data will you use?**
   - Where will you get it?
   - How will you gather it?
3. **How will you solve the problem?**
   - What preprocessing steps will be required?
   - Which algorithms you plan to use? Be as specific as you can!
4. **How will you evaluate, measure success?**

Step 3: Feedback and Coaching Sessions

- After submitting your outline, we will give you feedback (if required) on Thursday, 18.04.2023
- Later, Keti and Alex will give you tips and answer questions concerning your projects during the coaching sessions.
- Coaching sessions are optional: please send Keti and Alex an email if you want to attend until Monday night including your questions
- They will afterwards inform you about your slot via email.
- You are required to attend at least one coaching session.

Step 4: Project Reports

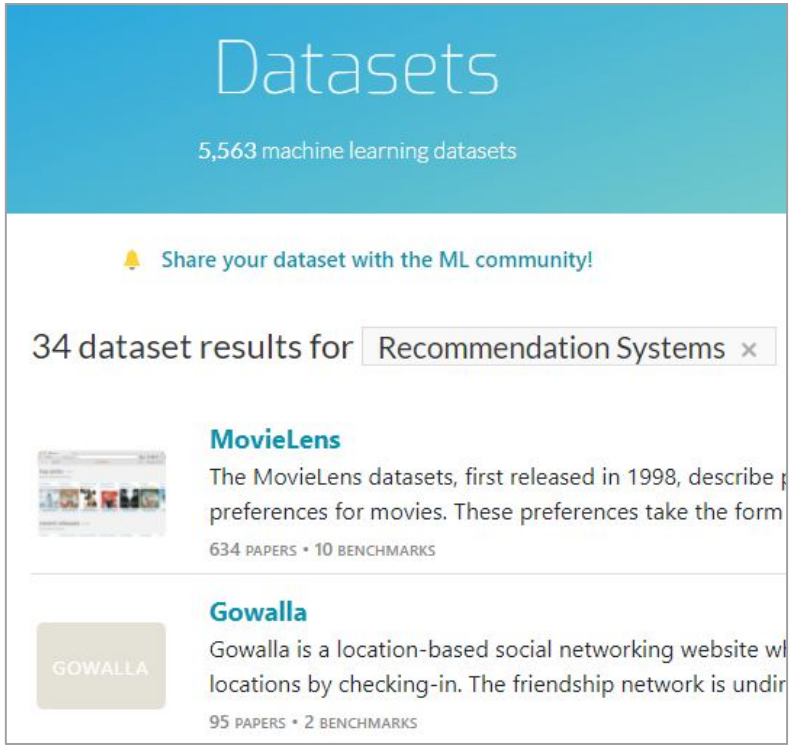
- Max. 12 pages (sharp!): title, toc or list of references do not count.
- Every additional page (including appendices) and every day of late submission downgrades your mark by 0.3
- Due Friday, 17.05.2023, 23:59. Send by mail to Chris, Simone, Keti and Alex.
- Outline for project summaries:
  1. Introduction: problem/task formulation, research questions and objective
  2. Methodology: describe the methods that you used and why you choose them
  3. Experimental setting: structure and statistics of the data set, evaluation measures
  4. Evaluation and discussion of the results: How do your results compare to existing solution?
  5. Conclusions (what can we learn from your work?) and future direction (what would you do differently, or additionally, why?)
- Requirements
  - You must use the Springer Computer Science Proceedings layout template.
  - Please cite sources properly. Preferred citation style [Author, year].
  - Also submit your code and links to the dataset. Alternatively, you can submit a link to a GitHub archive

Step 5: Project Presentations

- **Present your project in front of your fellow students**
- **Covers the contents of your report, this time in a “presentation” format**
- **Format**
  - 10 minutes presentation: each team member presents for 2-4 minutes
  - 5 minutes Question/Answer slot – everybody can (should) ask questions
- **Submit your slides via mail to Keti and Alex**
- **All students / project members must attend all sessions and presentations**

Where to find datasets for Web Usage Mining?

- **MovieLens**
  - 1M Dataset: 6.000 users, 3.900 movies, 1 million ratings
  - 10M Dataset: 71.000 users, 10.600 movies, 10 million ratings
- **Netflix Challenge**
  - 100M Dataset: 500.000 users, 18.000 movies, 100M ratings
- **Amazon Product Reviews**
  - 230M product reviews including star ratings
  - https://nijianmo.github.io/amazon/
- **Microsoft MIND**
  - 160k English news articles and
  - 15 million impression logs by 1 million users
  - https://msnews.github.io/
- **Papers with Code**
  - collects benchmark datasets
  - https://paperswithcode.com/datasets?task=recommendation-systems
- **Web 2.0 Platforms offer plenty of rating and usage data**
  - lastFM, Wikipedia



Datasets

5,563 machine learning datasets

Share your dataset with the ML community!

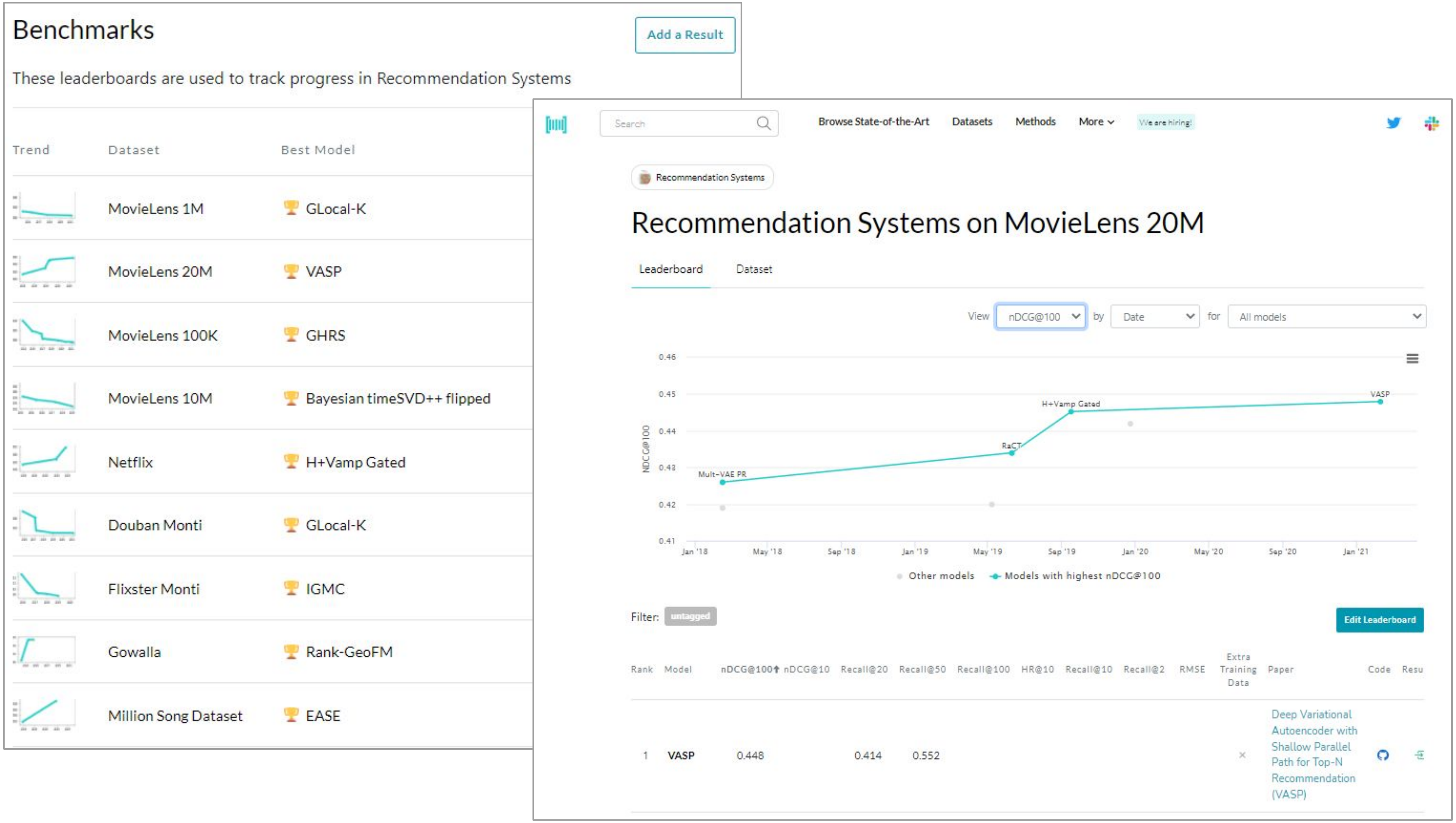
34 dataset results for Recommendation Systems ×

MovieLens
The MovieLens datasets, first released in 1998, describe p
preferences for movies. These preferences take the form
634 PAPERS • 10 BENCHMARKS

Gowalla
Gowalla is a location-based social networking website wh
locations by checking-in. The friendship network is undir
95 PAPERS • 2 BENCHMARKS

Benchmark Results: Recommender Systems

https://paperswithcode.com/task/recommendation-systems



Benchmarks

Add a Result

These leaderboards are used to track progress in Recommendation Systems

Trend	Dataset	Best Model
	MovieLens 1M	🏆 GLocal-K
	MovieLens 20M	🏆 VASP
	MovieLens 100K	🏆 GHRS
	MovieLens 10M	🏆 Bayesian timeSVD++ flipped
	Netflix	🏆 H+Vamp Gated
	Douban Monti	🏆 GLocal-K
	Flixster Monti	🏆 IGMC
	Gowalla	🏆 Rank-GeoFM
	Million Song Dataset	🏆 EASE

Recommendation Systems on MovieLens 20M

View nDCG@100 by Date for All models

Filter: untagged

Edit Leaderboard

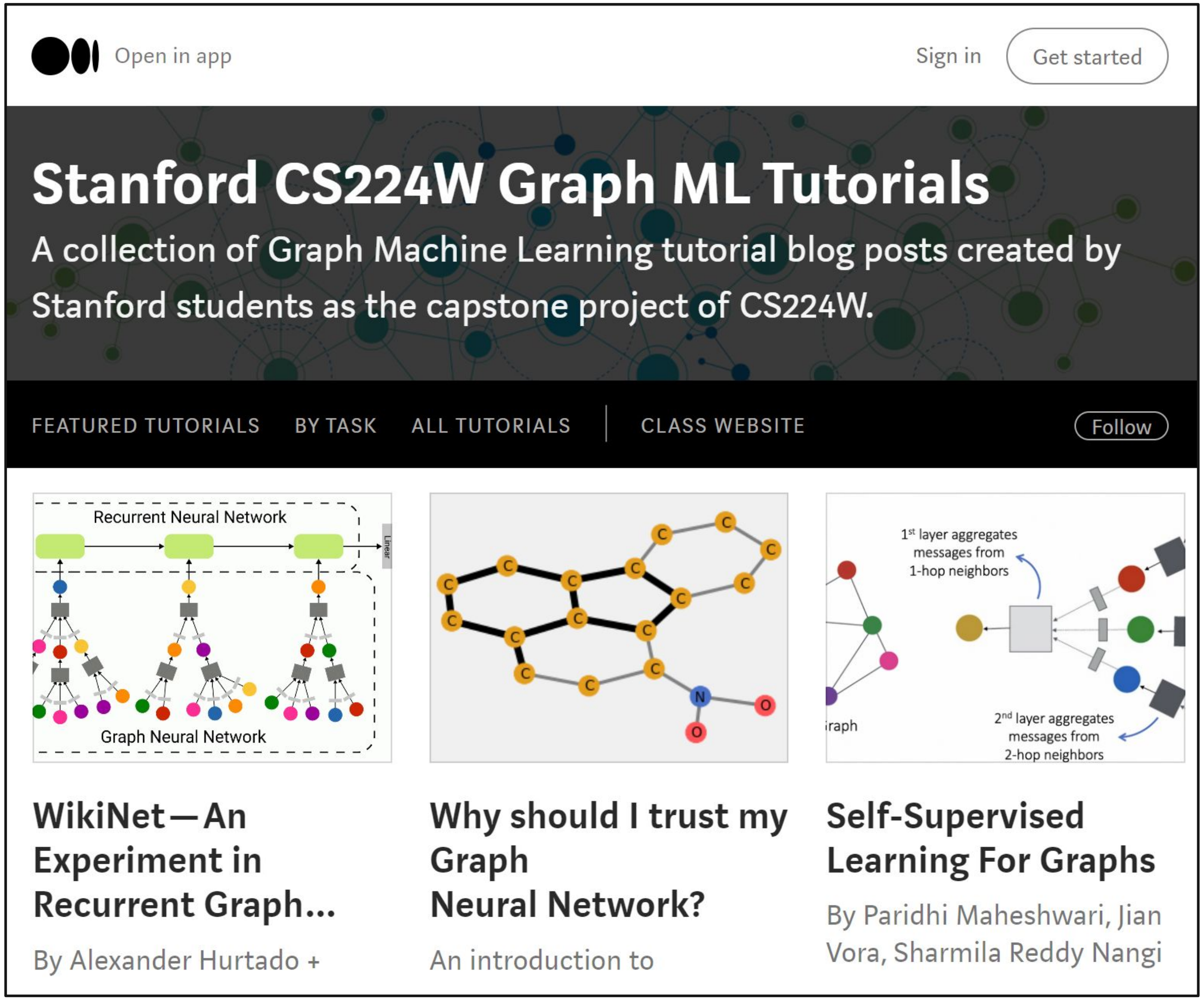
Rank	Model	nDCG@100↑	nDCG@10	Recall@20	Recall@50	Recall@100	HR@10	Recall@10	Recall@2	RMSE	Extra Training Data	Paper	Code	Resu
1	VASP	0.448		0.414	0.552						×	Deep Variational Autoencoder with Shallow Parallel Path for Top-N Recommendation (VASP)		

Where to find datasets for Web Structure Mining?

- **Stanford Large Network Dataset Collection**
  - Social networks: Facebook, Google+
  - Citation networks: Arxiv, US Patents
  - Product co-purchasing network: Amazon
  - http://snap.stanford.edu/data/index.html
- **Scientific Network Data Repository**
  - networks from 30+ categories ranging from biology to social networking
  - https://networkrepository.com/
- **Web Data Commons and Common Crawl Hyperlink Networks**
  - Different aggregation levels
  - http://webdatacommons.org/hyperlinkgraph/
  - https://commoncrawl.org/connect/blog/
- **The Koblenz Network Collection**
  - hundreds of networks about various topics
  - http://konect.cc/

Project Ideas for Machine Learning with Graphs

- **see term projects of Stanford CS224W students**



Open in app Sign in Get started

Stanford CS224W Graph ML Tutorials

A collection of Graph Machine Learning tutorial blog posts created by Stanford students as the capstone project of CS224W.

FEATURED TUTORIALS BY TASK ALL TUTORIALS | CLASS WEBSITE Follow

Recurrent Neural Network
Linear
Graph Neural Network
C
N
O
1st layer aggregates messages from 1-hop neighbors
raph
2nd layer aggregates messages from 2-hop neighbors

WikiNet—An Experiment in Recurrent Graph...

By Alexander Hurtado +

Why should I trust my Graph Neural Network?

An introduction to

Self-Supervised Learning For Graphs

By Paridhi Maheshwari, Jian Vora, Sharmila Reddy Nangi

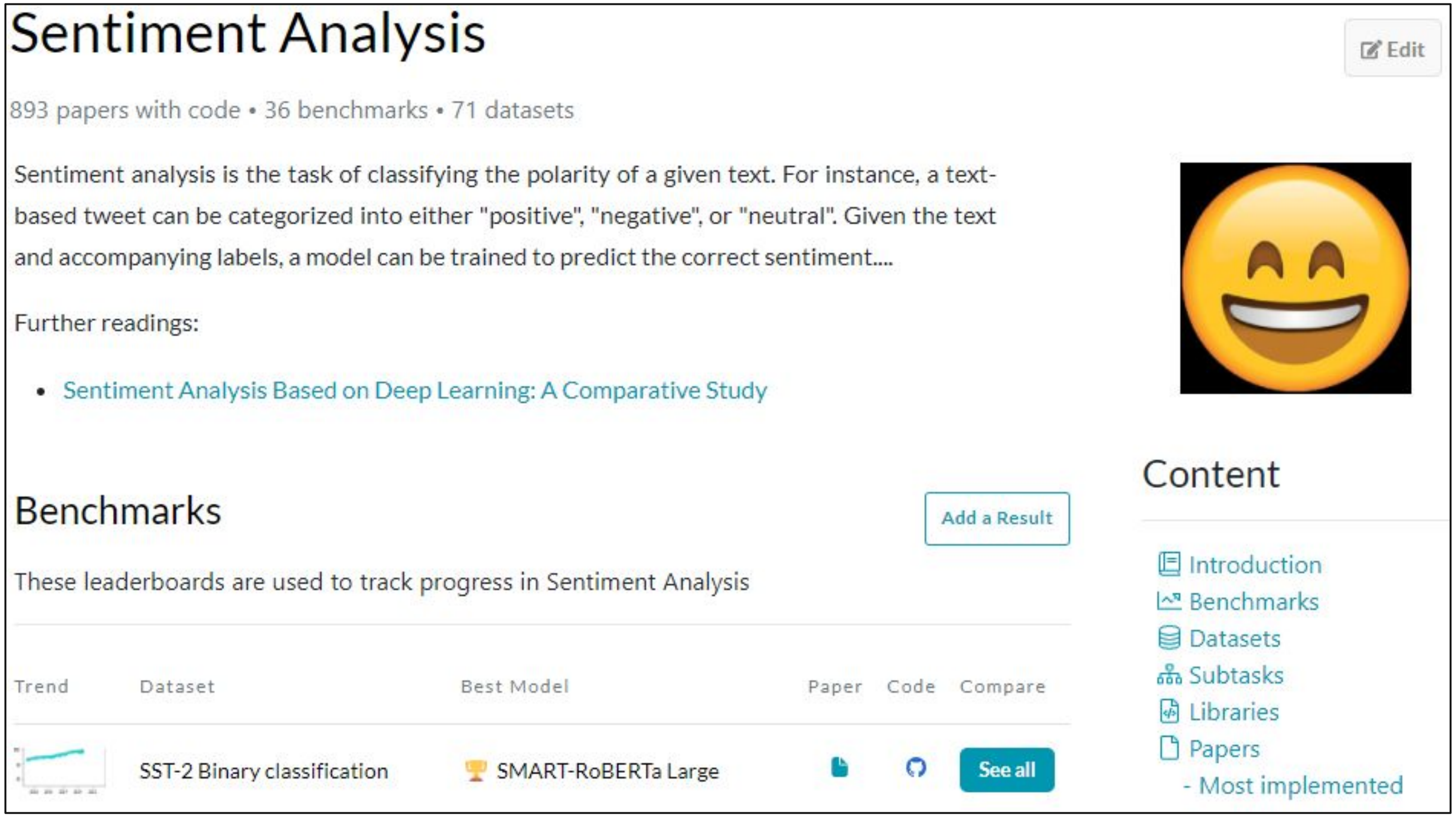
- **https://medium.com/stanford-cs224w**

Where to find datasets for Web Content Mining?

- **SemEval datasets**
  - Multiple datasets on text understanding task like sentiment analysis (e.g., from Twitter)
  - http://alt.qcri.org/semeval{2014-2021}/
- **Amazon Review Data**
  - Amazon product metadata and reviews
  - https://nijianmo.github.io/amazon/index.html
  - https://s3.amazonaws.com/amazon-reviews-pds/readme.html
- **Web Data Commons**
  - Product/hotel/restaurant reviews as part of Microdata dataset
  - http://www.webdatacommons.org/structureddata/
- **Academictorrents.com**
  - Various large data sets
  - e.g. Enron Email Bag of Words, Arizona State University Twitter Data Set
- **Kaggle**
  - Tons of datasets on a variety of topics
  - https://www.kaggle.com/datasets
- **Crawl your own data**

Benchmark Results: Sentiment Analysis

- **Papers with code**
  - https://paperswithcode.com/task/sentiment-analysis



Sentiment Analysis

893 papers with code • 36 benchmarks • 71 datasets

Sentiment analysis is the task of classifying the polarity of a given text. For instance, a text-based tweet can be categorized into either "positive", "negative", or "neutral". Given the text and accompanying labels, a model can be trained to predict the correct sentiment....

Further readings:

- Sentiment Analysis Based on Deep Learning: A Comparative Study

Benchmarks

These leaderboards are used to track progress in Sentiment Analysis

Trend	Dataset	Best Model	Paper	Code	Compare
	SST-2 Binary classification	🏆 SMART-RoBERTa Large			See all

Content

- **Huggingface Datasets Hub - Task Sentiment Analysis**
  - https://huggingface.co/datasets?task_ids=task_ids:sentiment-classification&sort=downloads

Where to Find Information about Additional Methods?

Check out the solutions to your task that other people have tried.

- by investigating the state-of-the-art for your task on Papers with Code
- by looking through the discussion groups and code of related Kaggle competitions
- search for survey papers about your task on Google Scholar: “task name + survey”. Select recent and frequently cited ones.

Get Additional Advice from a Stanford Professor

Christopher Potts

- **How to evaluate your model?**
  - **https://www.youtube.com/watch?v=TxTblROT9lY**
- **How to structure your project report?**
  - **https://www.youtube.com/watch?v=DZNwO-p5PGY**
- **How to present the results of your project?**
  - **https://www.youtube.com/watch?v=GGx7klcahzY**

Questions?

4. Team Formation and Next Steps

1. Anybody without a team?

2. People with teams:
   - Meet in your team now!
   - Agree on use case
   - Decide on or collect data
   - Write project outline