

Introduction to KDD Cup 2023



Updated Timetable

21.3.	New: KDD Cup Kick Off	Neural Networks & Deep Learning
28.3.	Hyperparameter Tuning	--
4.4.	<i>Easter Break</i>	<i>Easter Break</i>
11.4.	<i>Easter Break</i>	<i>Easter Break</i>
18.4.	KDD Cup	Hyperparameter Tuning
25.4.	Model Verification	--
2.5.	KDD Cup	Model Verification
9.5.	Anomaly Detection	--
16.5.	KDD Cup	Anomaly Detection
23.5.	KDD Cup	--

TASK 1: Next Product Recommendation




- Predict next product that a user is likely to engage with, given their session data and the product attributes
- Multilingual! (**UK, DE, JP**, FR, IT, ES)
- Data
 - *products_train.csv* (590MB)
 - *sessions_train.csv* (260MB)
 - *sessions_test_task1.csv* (20MB)distribution UK:DE:JP => 10:1:1

Language (Locale)	# Sessions	# Products (ASINs)
German (DE)	1111416	513811
Japanese (JP)	979119	389888
English (UK)	1182181	494409
Spanish (ES)	89047	41341
French (FR)	117561	43033
Italian (IT)	126925	48788

Timeline

- Start Date: **15th March 2023**
- End Date: **14th June 2023 00.00 UTC**
- Winner Announcement: **14th June 2023**

Prizes

-  First place: \$4,000
-  Second place: \$2,000
-  Third place: \$1,000

Dataset Details

Products

- locale: the locale code of the product (e.g., DE)
- id: a unique for the product. Also known as Amazon Standard Item Number (ASIN) (e.g., B07WSY3MG8)
- title: title of the item (e.g., "Japanese Aesthetic Sakura Flowers Vaporwave Soft Grunge Gift T-Shirt")
- price: price of the item in local currency (e.g., 24.99)
- brand: item brand name (e.g., "Japanese Aesthetic Flowers & Vaporwave Clothing")
- color: color of the item (e.g., "Black")
- size: size of the item (e.g., "xxl")
- model: model of the item (e.g., "iphone 13")
- material: material of the item (e.g., "cotton")
- author: author of the item (e.g., "J. K. Rowling")
- desc: description about a item's key features and benefits called out via bullet points (e.g., "Solid colors: 100% Cotton; Heather Grey: 90% Cotton, 10% Polyester; All Other Heathers ...")

Sessions

Input example:

locale	example_session
UK	[product_1, product_2, product_3]
DE	[product_4, product_5]

Output example:

next_item
[product_25, product_100,..., product_199]
[product_333, product_123,..., product_231]

Data Exploration

- Product frequencies?
- Session sizes?
- Products from multiple locales in a session?
- Products from FR/IT/ES relevant?

=> compare all this for **train and test set!**

- Price conversion necessary?
- Relevant external information?

The Evaluation: MRR

The evaluation metric for Task 1 is Mean Reciprocal Rank (MRR).

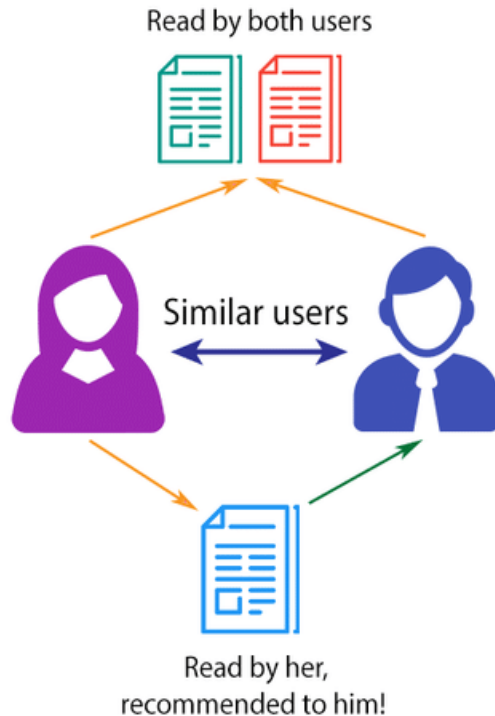
Mean Reciprocal Rank (MRR) is a metric used in information retrieval and recommendation systems to measure the effectiveness of a model in providing relevant results. MRR is computed with the following two steps: (1) calculate the reciprocal rank. The reciprocal rank is the inverse of the position at which the first relevant item appears in the list of recommendations. If no relevant item is found in the list, the reciprocal rank is considered 0. (2) average of the reciprocal ranks of the first relevant item for each session.

$$\text{MRR@K} = \frac{1}{N} \sum_{t \in T} \frac{1}{\text{Rank}(t)},$$

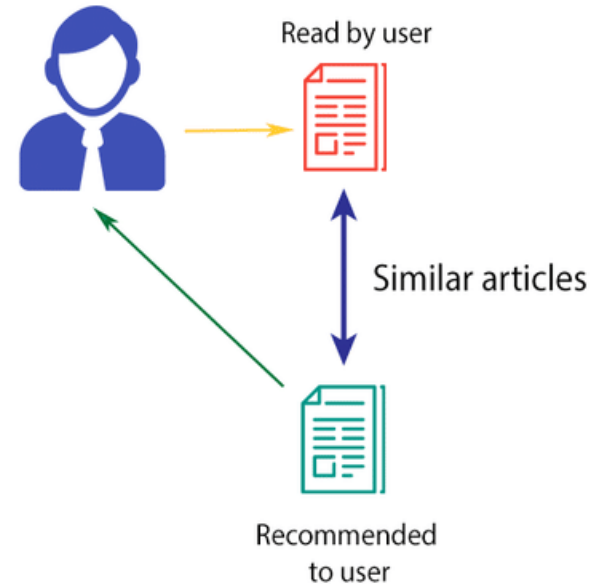
where $\text{Rank}(t)$ is the rank of the ground truth on the top K result ranking list of test session t , and if there is no ground truth on the top K ranking list, then we would set $\frac{1}{\text{Rank}(t)} = 0$. MRR values range from 0 to 1, with higher values indicating better performance. A perfect MRR score of 1 means that the model always places the first relevant item at the top of the recommendation list. An MRR score of 0 implies that no relevant items were found in the list of recommendations for any of the queries or users.

Recommender Systems in a Nutshell

COLLABORATIVE FILTERING



CONTENT-BASED FILTERING



https://www.researchgate.net/publication/323726564_Web_Recommender_System_for_Job_Seeking_and_Recruiting

Recommender Systems in a Nutshell

- Familiarize yourselves with the most common techniques
 - e.g., https://www.youtube.com/watch?v=v_mONWiFv0k
- Think about both collaborative and content-based methods
- In the end, you will most likely use a mix
 - aka hybrid recommender system
 - E.g.: generating candidates with one method, re-rank with another

Starting Points for Collaborative Filtering

- Start simple: find sets of cooccurring product pairs
 - E.g.: a significant amount of sessions with product A also contain product B
- More general:
 - Frequent itemset mining
- Investigate:
 - Does order matter? (e.g., is product A added after product B, but not vice versa?)
 - > maybe use GSP as an alternative

Starting Points for Content-based Recommenders

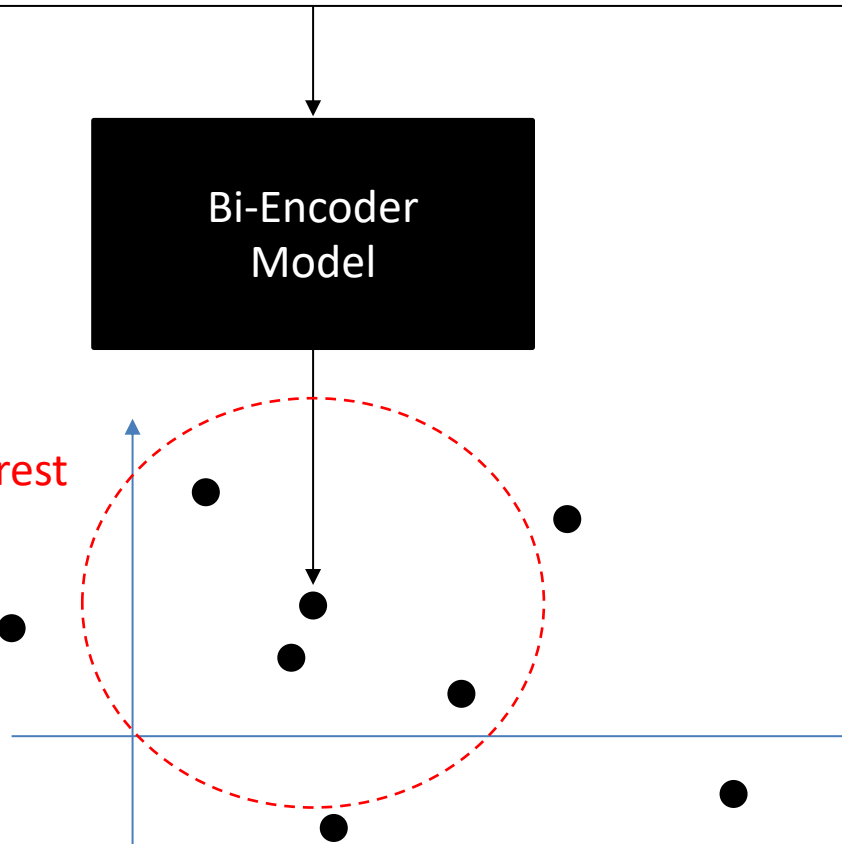
- Requires a similarity of products
- Start simple:
 - E.g., cosine on TF-IDF of product titles
- More complex methods
 - E.g., using transformers (see later)

Idea: Multilingual Transformer

Using the predictions for ranking

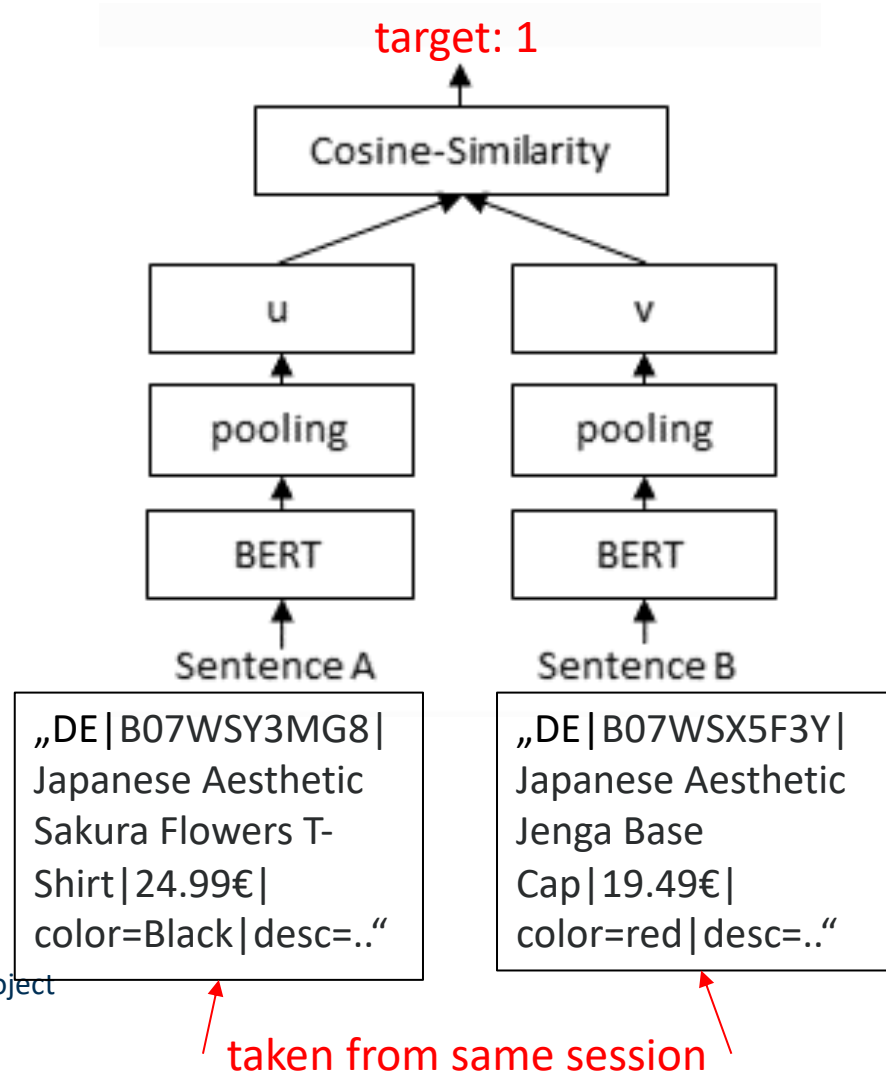
„DE|B07WSY3MG8|Japanese Aesthetic Sakura Flowers T-Shirt|24.99€|color=Black|desc=..“

Retrieval through
Approximate Nearest
Neighbour Search



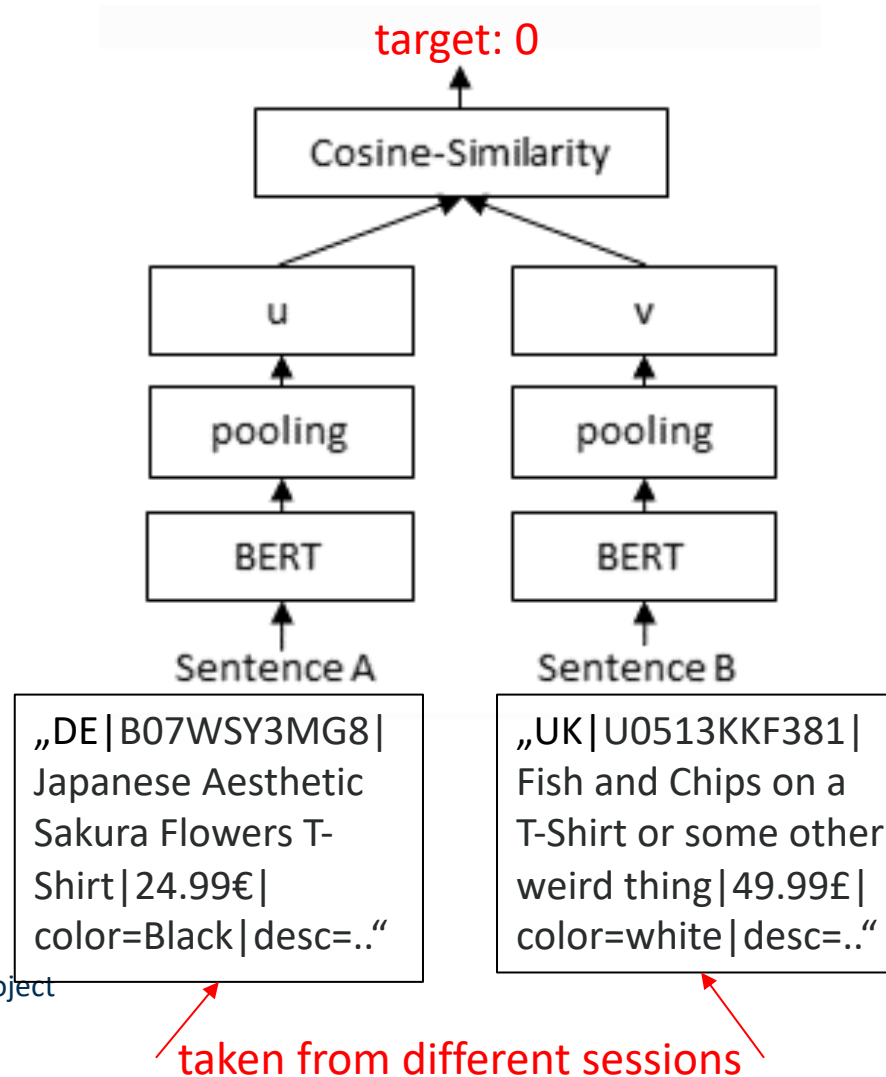
Idea: Multilingual Transformer

Training the Model with Sessions



Idea: Multilingual Transformer

Training the Model with Sessions



Idea: Multilingual Transformer

Tools and Documentation

- Implementation of Bi-Encoder in S-Bert: [sbert.net](https://www.sbert.net)
- Introduction to Transformers and Models: huggingface.co
- You can try out various models and see what works
- There are models that are already capable of multiple languages => make sure to try out some of those!
- Relevant publications for this topic:
 - [BERT background](#)
 - [S-BERT](#)

Google Colab and bwUniCluster

- Google Colab provides computing resources (CPU/GPU) to a certain extent => **accessible and easy to use**
- In case you need more resources: try **BwUniCluster**
 - Run Jupyter notebooks or computing jobs on the BW cluster
 - More info: <https://wiki.bwhpc.de/e/BwUniCluster2.0>
 - How to register: <https://wiki.bwhpc.de/e/Registration/bwForCluster>

Team Formation

- Use [Google Doc](#) to look for team members or register team
- **Deadline: March 28th, 3:30pm**
- If you want to participate in the project:
 - Make sure that your name is in the Google Doc (either in **TEAM REGISTRATION** or in **LOOKING FOR A TEAM**)
 - If you are still looking for a team after the deadline, we will assign you to an existing team or create a new one



Q & A

Your Tasks until after Easter

- Fill out the Google Doc for team formation
- After the deadline, get in contact with your teammates
- Register your team at the KDD Cup website
- In our session after Easter, every team will have roughly 10 minutes to present their first ideas and findings
 - Analyses, insights, questions concerning the data
 - A first approach of recommending products
 - don't forget to include the performance on the leaderboard!
 - Anything else they find interesting about the task
 - Problems they encountered (and solved)
 - Unclear details of the task
 - ...