Vietnamese-German University

FRANKFURT UNIVERSITY OF APPLIED SCIENCES

Thesis Defense

Comparative Analysis of Personalized Recommender Systems for Master's Programs: Exploring Different Approaches

---

Vinh Nguyen Phuoc Bao Minh
Mat. No. 1403541
Bachelor Thesis

Supervisor: Dr. Tran Hong Ngoc
Co-supervisor: Dr Dinh Hai Dung

Date of Presentation: 13/12/2024

Overview

- Motivation
- Challenges
- Research Question and Objectives
- Key Terminologies
- Methodology
- Experiments
- Results and Findings
- Conclusion and Future Work
- Q&A

Motivation

- *“On average, the adult human makes 35,000 decisions a day.”* (Bryan Robinson, Jul 2024)

- Emphasize the importance of Recommender Systems in daily life

Back to Overview

Bryan Robinson. Six Tips To Prevent Decision Fatigue From Short-Circuiting Your Mind And Career, July 2024.

MEng Applied AI for Digital Production Management

DEGGENDORF INSTITUTE OF TECHNOLOGY DIT

Deggendorf Institute of Tec

Deggendorf, Germany

18 months (90 ECTS) full-time

MSc Artificial Intelligence in Computer Science

kaunas university of

ersity of Technology

huania

ne

MSc MultiMediaTechnology

FH Salzburg

zburg University of Applied Sciences

lzburg, Austria

onths (120 ECTS) full-time

AI and Language

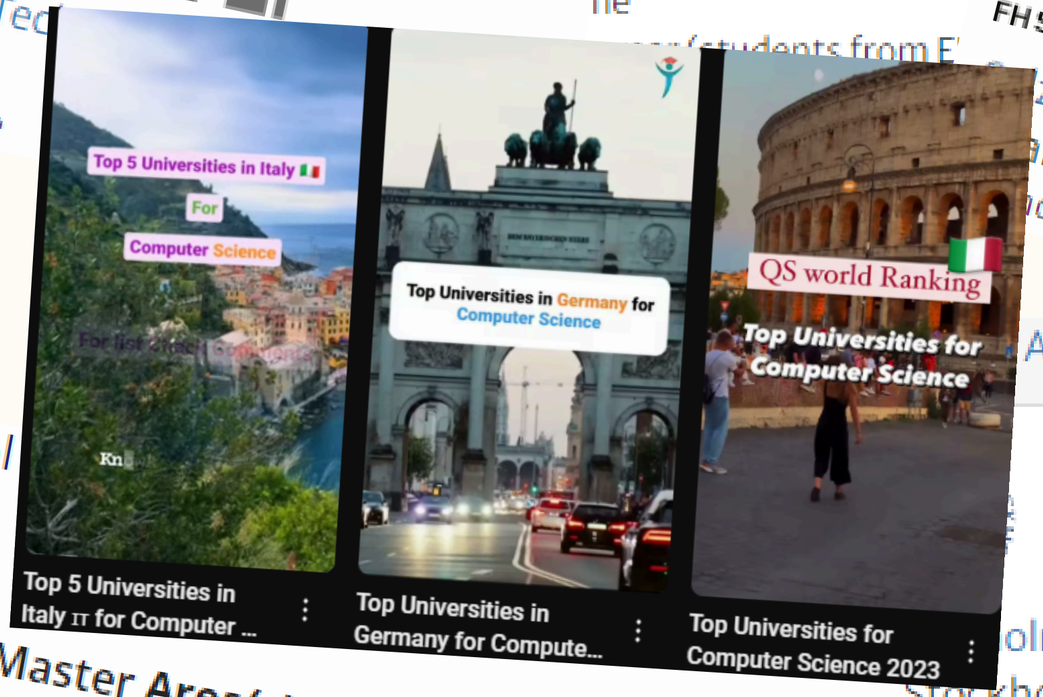
Stockholm University

olm University

Stockholm, Sweden

4 months (120 ECTS) full-time

uition-free for students from the EU/EEA



Results 1 to 10 of 812 total

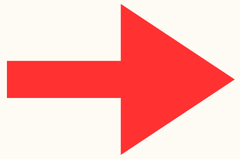
Your search criteria

Study Type: Both Degree: Master Area(s) of Study: Computer Science

Back to Overview

Motivation

Decision-making is time-consuming and prone to errors.



The need of recommendation techniques for Masters program.

Back to Overview

Challenges

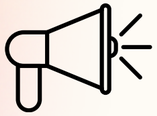
Back to Overview

Lack of variety

in user demographics and content variety

Sparsity

affect recommendation accuracy



Implicit Data

absence, not readily available on the internet.

Research Question

How can various recommender system methodologies be effectively designed, implemented, and evaluated under the absence of user interaction data?

Back to Overview

Objectives

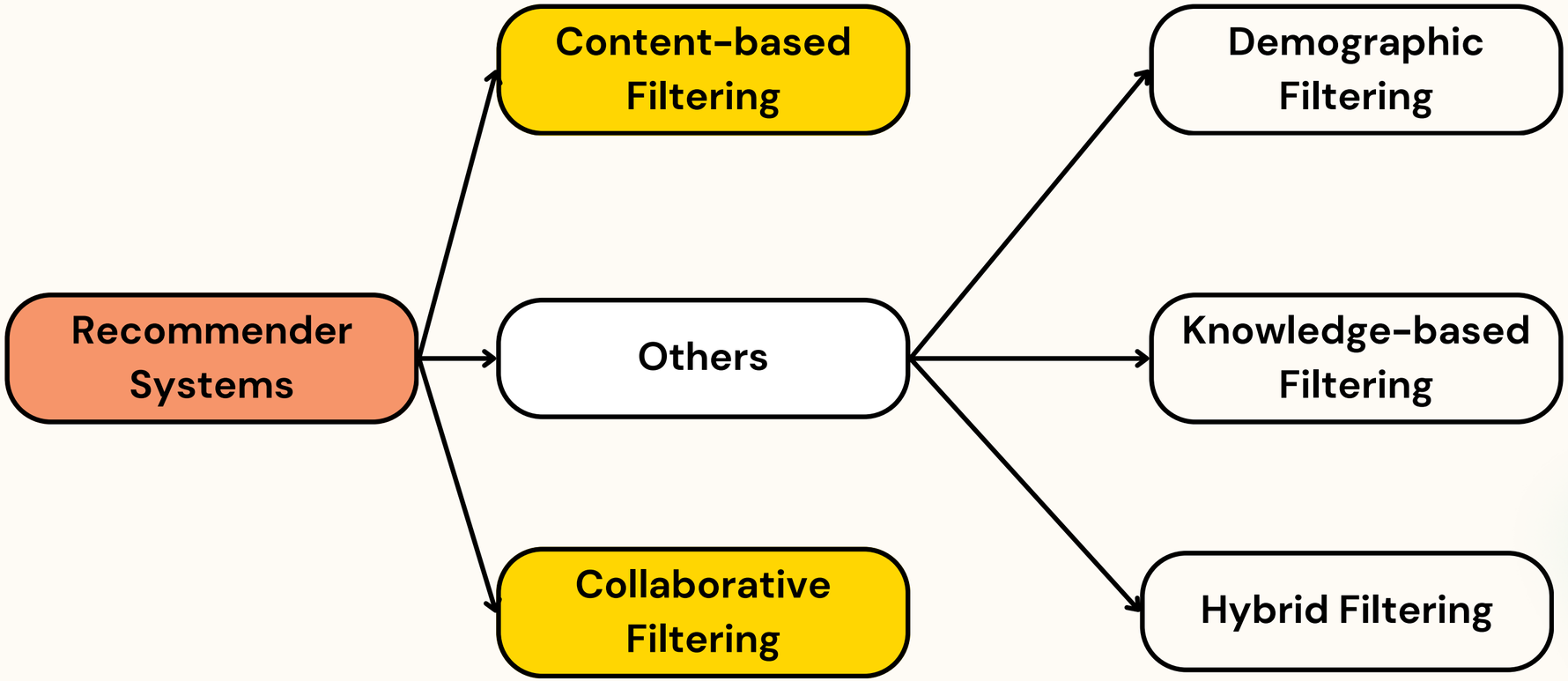
- Overview of Recommender Systems.
- Create a database of user's study preferences.
- Evaluation of our database and recommendation methodologies.

Key Terminologies

Back to Overview

Recommender Systems

- Applications that attempt to predict and “recommend” items in which users may be interested.



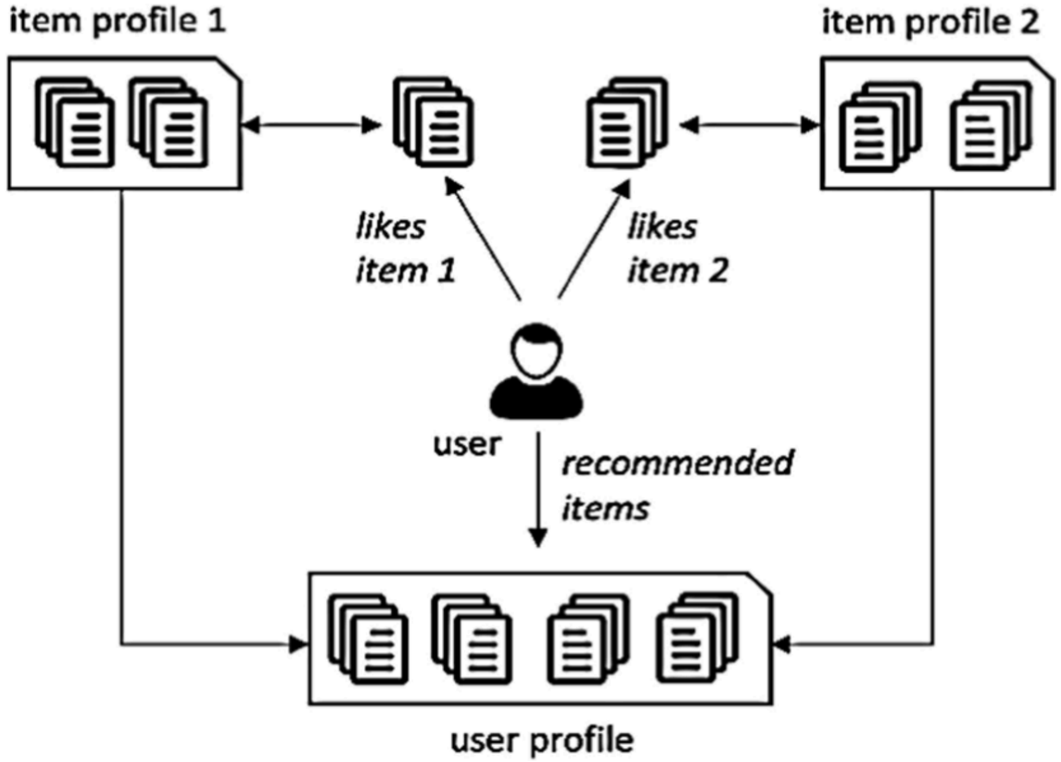
Key Terminologies

Back to Overview

Content-based Filtering

- Analyzes the characteristics of items, in terms of keywords, text, or metadata, and then matches them with a user's preferences.

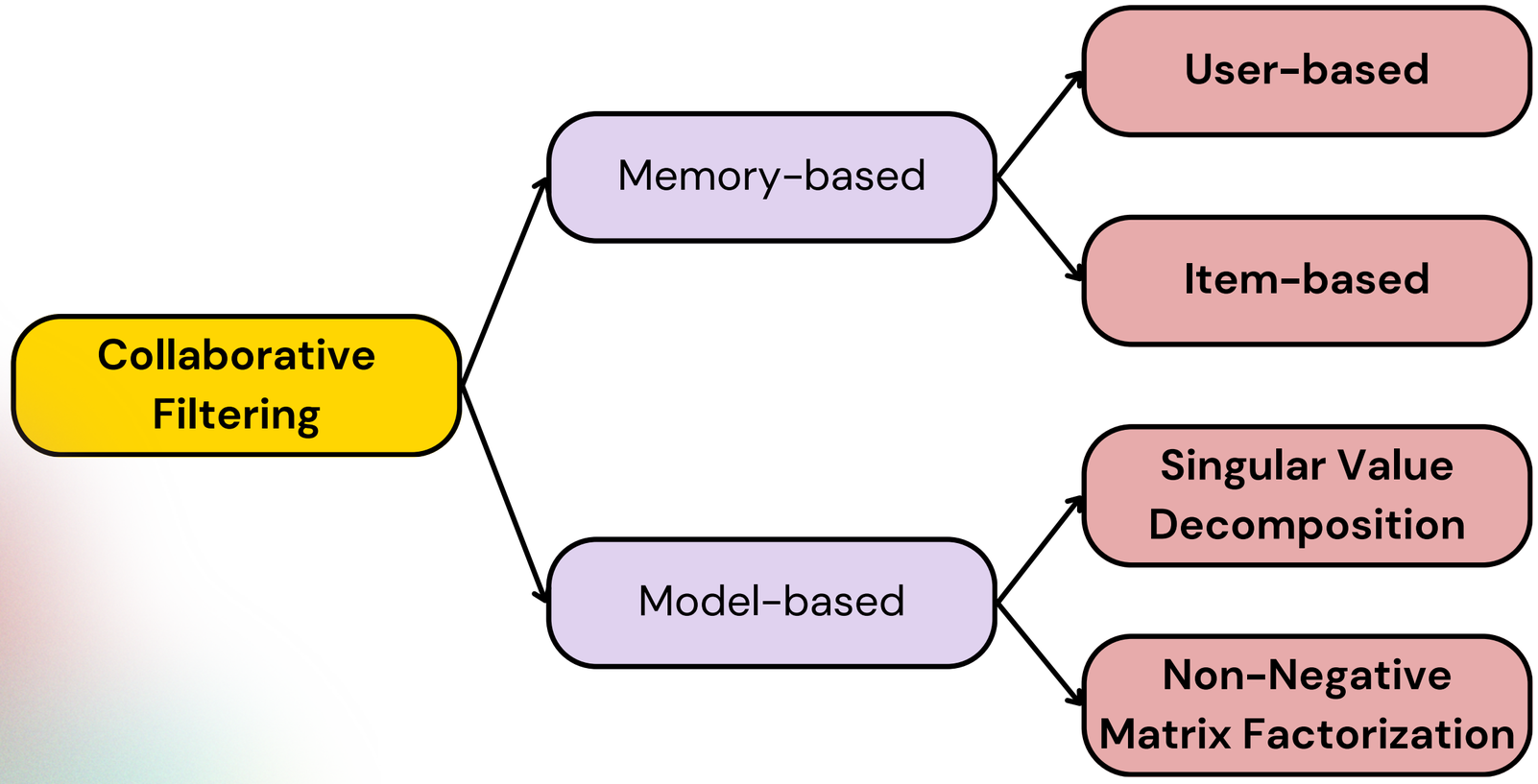
-> Compare user preferences with program features



Key Terminologies

Back to Overview

Collaborative Filtering



Key Terminologies

Back to Overview

Collaborative Filtering

- **User-Item Matrix** instead of contents of users and items
- **Cold-start and Data Sparsity Problem**

	Item 1	Item 2	Item 3	Item 4	Item 5
User A	5	3	4	4	?
User B	3	1	2	3	3
User C	4	3	4	3	5
User D	1	3	1	?	2

Key Terminologies

Back to Overview

Cold-start and Data Sparsity

- missing too much data to relate user/item to another

	Item 1	Item 2	Item 3	Item 4	Item 5
User A	5	3	4	4	?
User B	3	1	2	3	?
User C	4	3	4	3	?
User D	1	3	1	?	?

Key Terminologies

Back to Overview

Memory-based Collaborative Filtering

- User-based technique finds users with similar interaction patterns to a target user and recommends items with whom these similar users have engaged.

Item-based

	Item 1	Item 2	Item 3	Item 4	Item 5
User A	5	3	4	4	?
User B	3	1	2	3	3
User C	4	3	4	3	5
User D	1	3	1	?	2

User-based

Key Terminologies

Back to Overview

Model-based Collaborative Filtering

- Uses mathematical techniques to reduce **user-item matrix** to **lower-dimension matrices.**
- **Example: Singular Value Decomposition (SVD)**

and **Non-Negative Matrix Factorization (NMF)**

Next to Methodology

Methodology

Back to Overview

Synthesizing User Preference

- using a **GPT-based generative model** with a realistic prompt

NLP-based Content-based Filtering

- **similarity matching** by transforming program and user preferences into **vector representations** using **TF-IDF** and **word embeddings**

Collecting Universities' Programs

- using **web scraping** techniques, extracting information about fields of study, course descriptions, and other relevant details

Collaborative Filtering Techniques

- **evaluating** CF techniques

Methodology

Back to Overview

- **Synthesizing User Preference**

Why?

- No available database.
- Synthetic user preferences can replicate real user input.
- Time effectiveness.
- Ethical compliance.
- Diverse user preferences.

Methodology

Back to Overview

- **Synthesizing User Preference**

- To generate a list of user preference data for testing the recommendation system.

No.	User preferences
1	I want to study Data Science in Germany for two years.
2	Looking for an online Master's in Business Administration with a focus on entrepreneurship.
3	Interested in a Master's program in Renewable Energy Engineering taught in English.

- Prompt message: ensure diverse, relatable and clear range statements

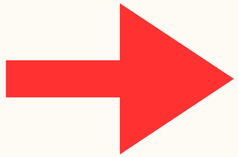
```
Task:
Generate N statements that align with these guidelines. Ensure diversity,
  clarity, and relatability in the statements, focusing on popular and
  emerging options. Introduce ambiguity by expressing openness or
  flexibility in any preferences, including program names, by sometimes
  using general fields of interest or unspecified areas. Avoid overly
  niche or obscure topics while maintaining a realistic and personalized
  tone.
```

Methodology

Back to Overview

- **Collecting Universities' Programs**

- Web scraping techniques were used
- Extensive Python libraries used: BeautifulSoup4 and Selenium WebDriver.



Programs' data was **automatically** retrieved from the university academic portal.

Methodology

Back to Overview

- **Collecting Universities' Programs**

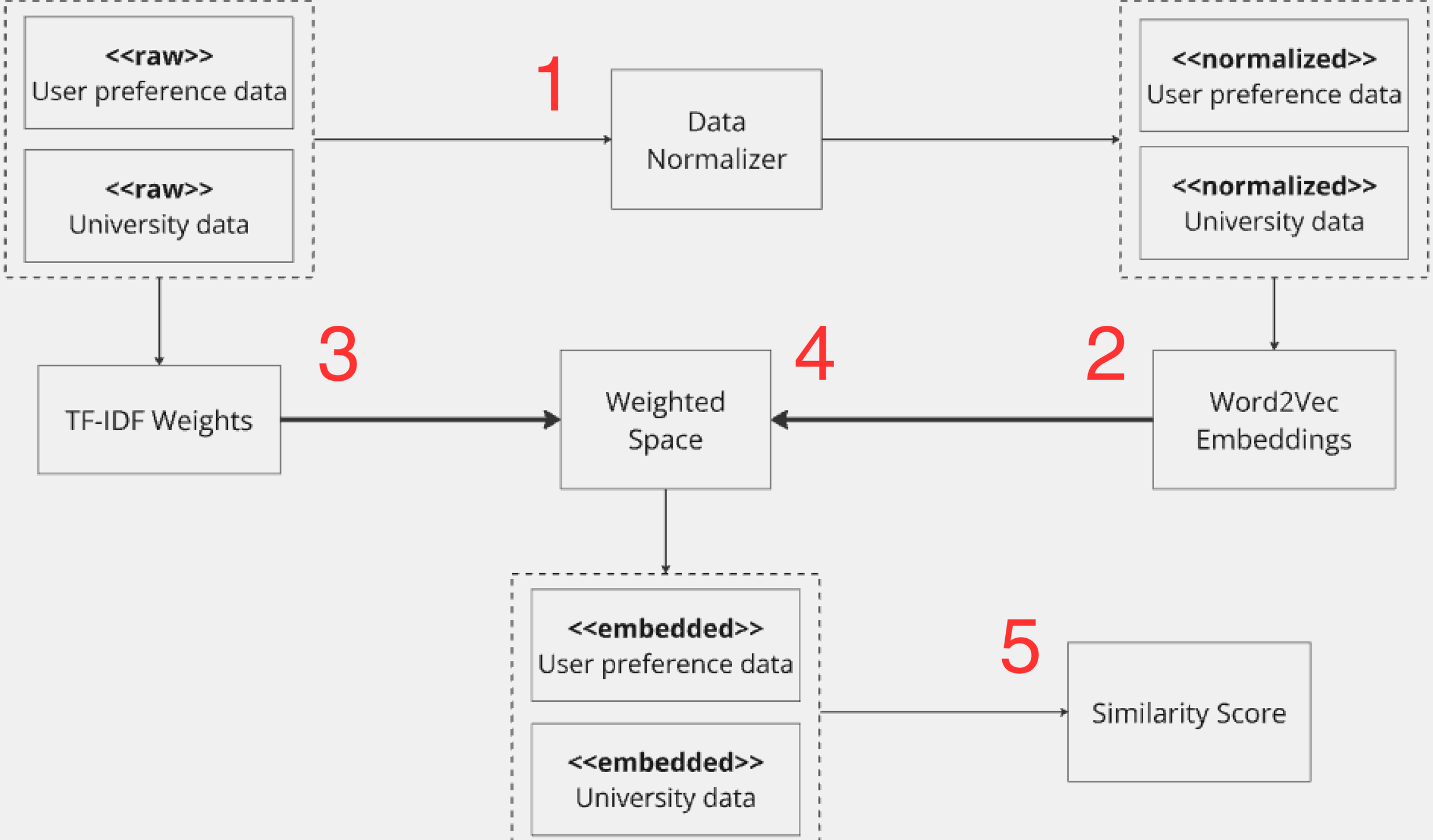
Attribute	Description and Examples
Program Name	Title of the academic program. *Examples:* Computer Science, Urban Planning.
University	Institution offering the program. *Examples:* Technical University of Berlin, University of Cologne.
Location	Geographical location of the university. *Examples:* Berlin, Cologne.
Duration	Length of the program. *Examples:* 4 semesters, 32 months.
Degree Type	Type of degree awarded upon completion. *Examples:* Master of Science, Master of Ar
Language	Language of instruction. *Examples:* English, German.
Subject	Focus of the academic program. *Examples:* Statistics, Software Engineering
Subject	Focus of the academic program. *Examples:* Statistics, Software Engineering.
Study Mode	Whether the program is online, on-campus, or hybrid. *Examples:* Online, On-campus, Hybrid.
Overview	A detailed overview of the university or program. *Example:* offers a strong focus on interdisciplinary studies; ing cutting-edge research opportunities; state-of-the-art f
Teaching & Studying	Description of teaching methods and program structure. *Example:* program combines lectures, hands-on worksho collaborative projects; ensure a balance of theoretical and cal skills.
Researching	Information about current research projects and facilities *Example:* research focuses on a certain field/discipline.

Methodology

Back to Overview

- **NLP-based Content-based Filtering**

1. Data Normalization
2. Word2Vec Embeddings
3. Term Frequency - Inverse Document Frequency Weights
4. Weighted Sentence Embeddings
5. Similarity Scoring


<<raw>>
User preference data
<<raw>>
University data
1
Data
Normalizer
<<normalized>>
User preference data
<<normalized>>
University data
3
4
2
TF-IDF Weights
Weighted
Space
Word2Vec
Embeddings
<<embedded>>
User preference data
<<embedded>>
University data
5
Similarity Score

Methodology

Back to Overview

- **NLP-based Content-based Filtering**

Process	Descriptions & Examples
Decontraction	return contractions to its full form. • can't ➡ cannot
Lowercasing	• PROGRAM ➡ program
Tokenization	break down phrase to individual words, called **tokens** • "my name is Minh" ➡ ['my', 'name' ,'is', 'Minh']
POS Tagging	each token is assigned Verb, Noun...

- **Data Normalization**

Remove punctuations	remove all punctuation marks: **, . ? !**
Remove stopwords	remove all commonly used words. • **'the', 'a', 'is', 'and'...**
Lemma -tization	different grammatical forms to one type. • studying ➡ study • studies ➡ study • studied ➡ study

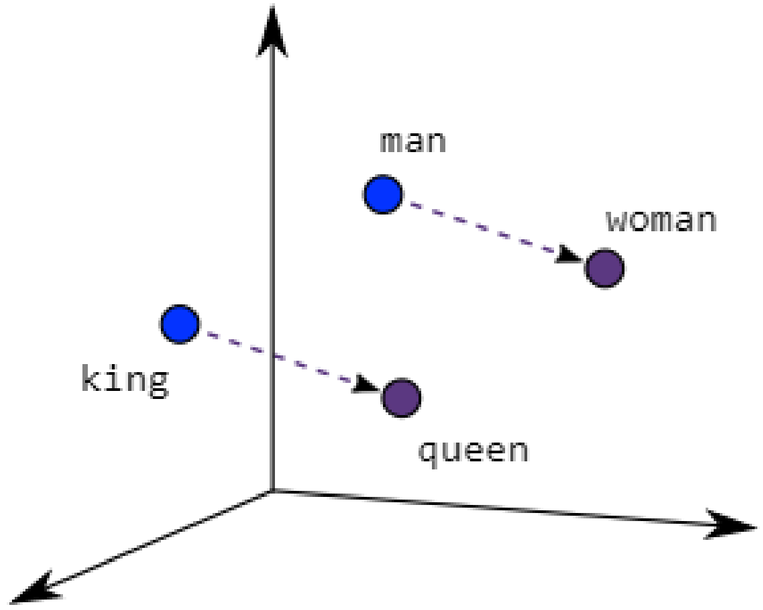
Methodology

Back to Overview

- **NLP-based Content-based Filtering**

Word Embeddings - Word2Vec:

- **Purpose**: represent words as vectors in a continuous vector space.
- Create **custom** Word2Vec model: **‘artificial’** and **‘intelligence’** appear closer
  → **More domain-specific**
- Using Gensim library with built-in Word2Vec functions



Methodology

Back to Overview

- **NLP-based Content-based Filtering**

Term Frequency – Inverse Document Frequency

$$\text{TF\_IDF}(t, d) = \text{TF}(t, d) \cdot \text{IDF}(t)$$

where

- $\text{TF}(t, d)$ is the frequency of term t in document d, and
- $\text{IDF}(t) = \log\left(\frac{N}{n_t}\right)$, where N is the total number of documents in the corpus, and n_t is the number of documents *containing the term* t.

Methodology

Back to Overview

- **NLP-based Content-based Filtering**

Weighted Sentence Embeddings

$$\text{SentenceEmbedding} = \sum_{i=1}^{n} \text{TF\_IDF}(t_i) \cdot \text{Word2Vec}(t_i)$$

Methodology

Back to Overview

- **NLP-based Content-based Filtering**

Similarity Scoring – Cosine Similarity

- Calculate cosine similarity for every user-item pair.
    - Ranging 0 to 1: 0 - completely dissimilar, 1 - perfectly similar
- Ranking for each user.
- Output **user-item similarity** matrix .

	Item 1	Item 2	Item 3	Item 4	Item 5
User A	5	3	4	4	?
User B	3	1	2	3	3
User C	4	3	4	3	5
User D	1	3	1	?	2

Methodology

Back to Overview

- **Collaborative Filtering Techniques**

KNN User-based CF

- select the top K most similar **items** between **users**
- use the average users' ratings on these similar items to predict the rating for the target item.

KNN Item-based CF

- same basis as User-based, but switch roles of users and items

```
sim_options = {
    'name': 'cosine',
    'user_based': True,
    'min_support': 5,
}
user_cf = KNNBasic(k=20, min_k=1,sim_options=sim_options)
```

Methodology

Back to Overview

- **Collaborative Filtering Techniques**

Singular Value Decomposition

$$R \approx U \Sigma_{m \times n} I_{n \times n}$$

Non-Negative Matrix Factorization

$$R \approx U.I$$

Experiment Setup

Synthetic Data Survey

Research Question

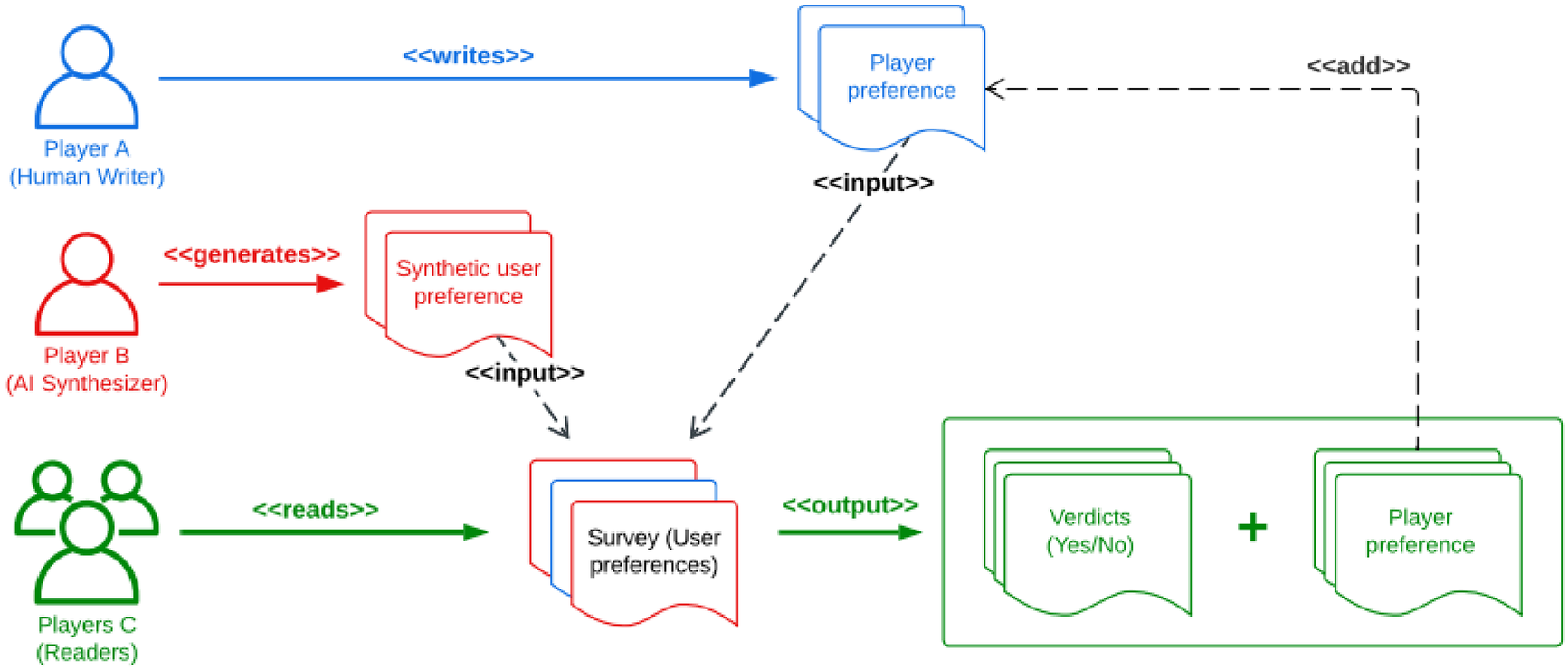
- *Given a set of unlabeled, text-only statements designed to appear human-like, to what extent can evaluators accurately distinguish between Human- and AI-generated text?*

Hypothesis

- *Evaluators are **not** able to distinguish between Human- and AI-generated text, and their classification accuracy will not differ significantly from random chance.*

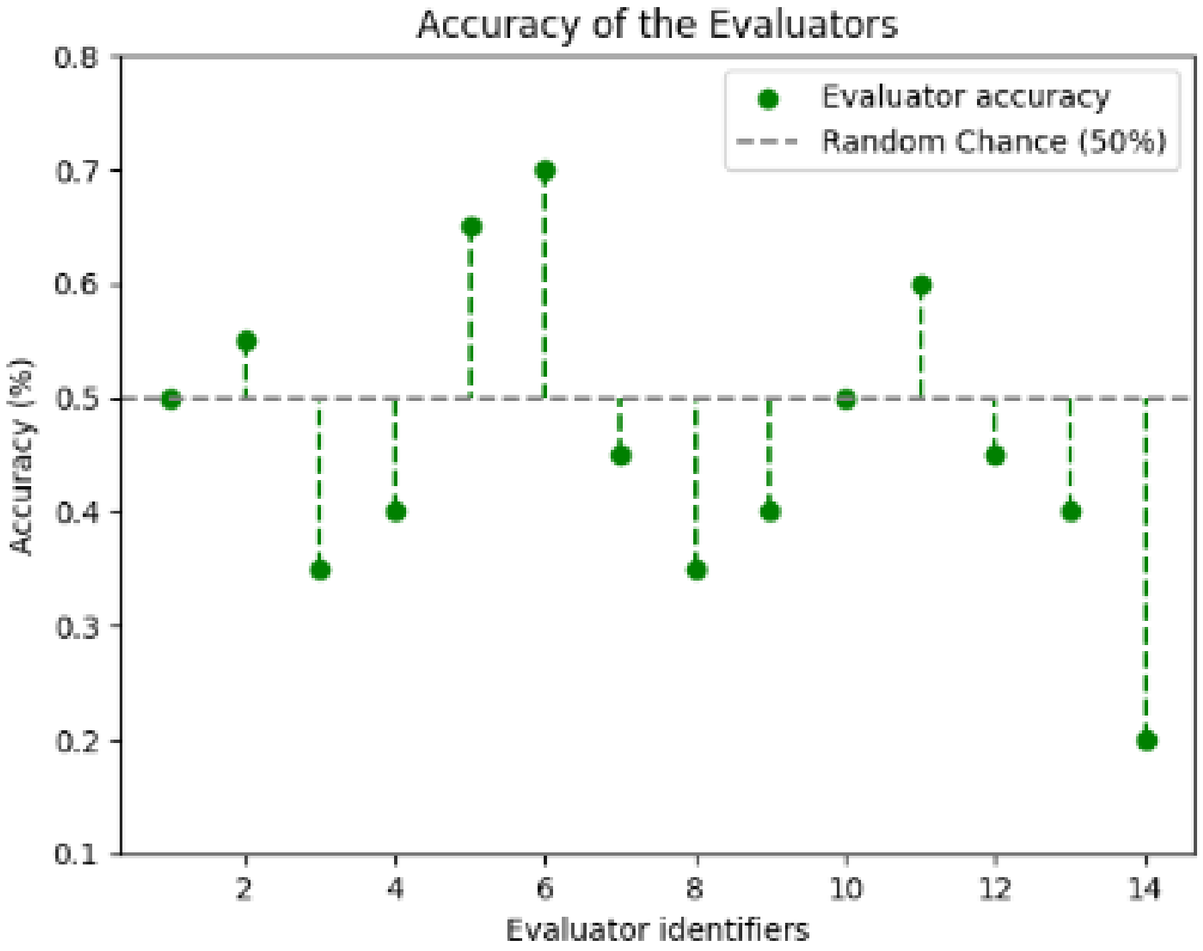
Back to Overview

Experiment Design



Back to Overview

Back to Overview



Experiment Results

Synthetic User Data

Average accuracy: 46.43%

- Mean accuracy result is close to random chance (50%)

Results & Findings

Back to Overview

- NLP-based Content-based Filtering

  Qualitative assessment of NLP-based CBF

- Collaborative Filtering Techniques

  Quantitative assessment of CF Techniques

Back to Overview

User:

- *A Master of Science in Energy Engineering in Stuttgart would be an excellent choice, focusing on renewable and efficient energy systems.*

Ranking	Program Details	Degree Type
1	Sustainable Energy Competence (SENCE) Stuttgart University of Applied Sciences Stuttgart, Germany	Master
2	Green Energy Westcoast University of Applied Sciences Heide, Germany	Master of Science
3	Renewable Energy Systems Hochschule Nordhausen Nordhausen, Germany	Master

Results & Findings

Qualitative assessment of NLP-based CBF

- Sample recommendations from a user preference statement.

Back to Overview

Results & Findings

Quantitative assessment of CF Techniques

- evaluate four Collaborative techniques using RMSE and MAE metrics across different data sparsity levels

Algorithm	RMSE	MAE
User-based CF	0.6769	0.5336
Item-based CF	0.7034	0.5596
NMF	0.5920	0.4738
SVD	0.4982	0.4020

25% sparsity

99% sparsity

Algorithm	RMSE	MAE
User-based CF	0.7653	0.6130
Item-based CF	0.7220	0.5749
NMF	0.7631	0.6109
SVD	0.6692	0.5383

Conclusion

Back to Overview

Synthetic User Preference

Use ChatGPT's OpenAI model to create first synthetic database of users' preferences

Literature Review of RecSys

common recommendation techniques, in theory and in practice

NLP-based Content-based Filtering

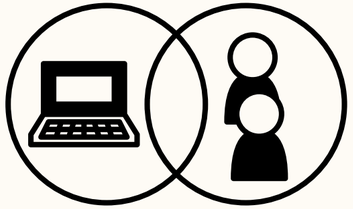
TF-IDF and **Word2Vec** helps user preferences with recommendations.

Collaborative Filtering

how data sparsity affect the performance through accuracy metrics such as **RMSE** and **MAE**

Future Work

Back to Overview

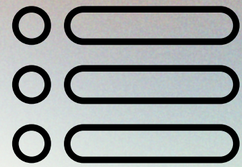


Hybrid RecSys

weighted hybrids or switching models

Knowledge-based RecSys

more domain-specific recommendations



Implicit Data

a database of user interaction data

Back to Overview

Q&A Session 

Thank you for listening!