The Application of Deep learning algorithms to Computational Problems - Group 1

- STUDENTS / UNIVERSITIES
  İlker GÜL / Sabancı University
  Metehan KOÇ / Sabancı University
  Tahir TURGUT / Sabancı University
- SUPERVISOR(S)
  Dr. Mehmet KESKİNÖZ

ABSTRACT

- ✓ This research focuses on which method would provide more successful book recommendations to the people.
- ✓ Initially, research examines cosine similarity, KNN, matrix factorization(MF) methods' background and implementation and explains other methods with those.
- ✓ Afterwards, these methods are compared by their error rate.
- ✓ Finally, successful methods of comparison are represented and further developments about that methods and possible future work is explained.

First 5 lines of the Goodbooks-10k data with significant attributes

	user_id	book_id	rating	best_book_id	work_id	books_count	authors	original_publication_year	title	language_code	tag_name
0	1	258	5	1232	3209783	279	Carlos Ruiz Zafón, Lucia Graves	2001.0	The Shadow of the Wind (The Cemetery of Forgot...	eng	to-read fantasy favorites currently-reading fi...
1	11	258	3	1232	3209783	279	Carlos Ruiz Zafón, Lucia Graves	2001.0	The Shadow of the Wind (The Cemetery of Forgot...	eng	to-read fantasy favorites currently-reading fi...
2	143	258	4	1232	3209783	279	Carlos Ruiz Zafón, Lucia Graves	2001.0	The Shadow of the Wind (The Cemetery of Forgot...	eng	to-read fantasy favorites currently-reading fi...
3	242	258	5	1232	3209783	279	Carlos Ruiz Zafón, Lucia Graves	2001.0	The Shadow of the Wind (The Cemetery of Forgot...	eng	to-read fantasy favorites currently-reading fi...

OBJECTIVES

- ➢ The research is made on a dataset called **Goodbooks-10k**
- ➢ Various algorithms are applied in **Python** to have a recommender system
- ➢ These models are evaluated by **Root mean square error and similarity scores**

CONTENT-BASED FILTERING

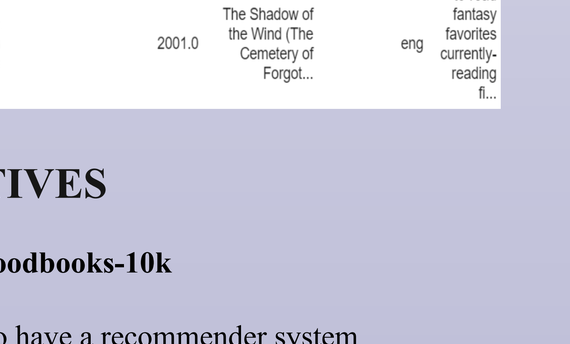
There are several methods to build content-based filtering systems. In this study, cosine similarity is used on vector space of book's title and tags which is created by TF-IDF and Count vectorizer to find similarities between items and recommend them with cosine similarity accordingly.

General Formula of TF-IDF

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{10}{df_i}\right)$$

- ➢ $f_{i,j}$ = number of occurrences of i in j
- ➢ df_i = number of documents containing i
- ➢ $N = total\ number\ of\ documents$

CountVectorizer (Edpresso Editor, 2020)



TF-IDF based recommendation with similarity scores

```
Recommending 5 products similar to Harry Potter and the Sorcerer's Stone (Harry Potter, #1)
-------
Recommended: Harry Potter and the Prisoner of Azkaban (Harry Potter, #3) (score:0.9840073445393256)
Recommended: Harry Potter and the Chamber of Secrets (Harry Potter, #2) (score:0.9820423548270265)
Recommended: Harry Potter and the Deathly Hallows (Harry Potter, #7) (score:0.9657235269252039)
Recommended: Harry Potter and the Half-Blood Prince (Harry Potter, #6) (score:0.9650833898658011)
Recommended: Harry Potter and the Goblet of Fire (Harry Potter, #4) (score:0.9488388043776768)
```

CountVectorizer based recommendation with similarity scores

	book_title	sim_books	scores	tags
0	Allegiant (Divergent, #3)	Insurgent (Divergent, #2)	1.0	[adult , adult fiction , fiction , young , you...
1	Allegiant (Divergent, #3)	Harry Potter and the Chamber of Secrets (Harry...	1.0	[adult , adult fiction , fiction , young , you...
2	Allegiant (Divergent, #3)	The Lightning Thief (Percy Jackson and the Oly...	1.0	[adult , adult fiction , fiction , young , you...
3	Allegiant (Divergent, #3)	Paper Towns	1.0	[adult , adult fiction , fiction , young , you...
4	Allegiant (Divergent, #3)	City of Ashes (The Mortal Instruments, #2)	1.0	[adult , adult fiction , fiction , young , you...
5	Allegiant (Divergent, #3)	The Maze Runner (Maze Runner, #1)	1.0	[adult , adult fiction , fiction , young , you...

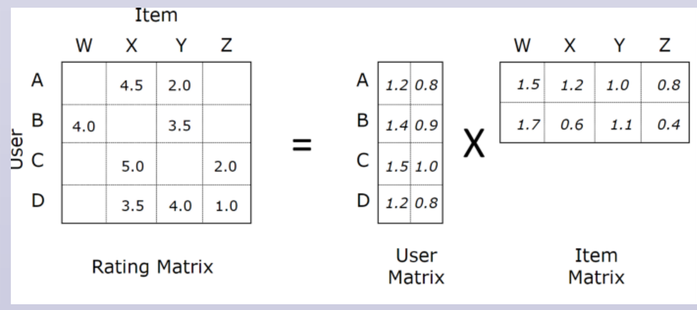
COLLABORATIVE FILTERING and CLUSTERING

- ➢ Collaborative filtering methods use the ratings in the form of a matrix called the rating matrix.
- ➢ The K-nearest Neighbors (KNN) algorithm is a simple, easy-to-implement algorithm to solve both classification and regression problems (Harrison, 2019).
- ➢ Matrix factorization (MF) is found to be effective in reducing the sparsity problem.
- ➢ Clustering is a common procedure that is done while having an exploratory data analysis.

KNN Algorithms Results

Algorithms	RMSE Results
KNNBaseline	0.805
KNNBasic	0.832
KNNWithMeans	0.807

Matrix Factorization Example



KNN Recommendation

```
[64] recommendBook(book_id=13)

Top recommendations for 1984 are:
Animal Farm
Animal Farm / 1984
Brave New World
Fahrenheit 451
Brave New World / Brave New World Revisited
Lord of the Flies
A Clockwork Orange
Slaughterhouse-Five
Darkness at Noon
Essays and Poems
```

K-Means Clustering

CONCLUSION and FUTURE WORK

- ✓ Both cosine similarity approaches produced a proper recommendation list. However, CountVectorizer offered books that are not as much related as the tf-idf vectorizer.
- ✓ Moreover, KNN based algorithms' RMSE scores are very close and have the best RMSE scores. RMSE scores are between 0.83 and 0.86.
- ✓ It should be noted that ratings are between 1 and 5. Hence, it is needed to decrease RMSE scores between 0.3 and 0.5 to get more accurate recommendations
- ✓ Deep learning techniques , language processing and filtering techniques will help recommender systems to better capture the user's needs and user's satisfaction
- ✓ Hybrid filtering composed of content-based and collaborative filtering will enable developers to reach people more in real life applications.

REFERENCES

- ➢ Edpresso Editor. (2020, February 28). *CountVectorizer in Python*. Educative: Interactive Courses for Software Developers. https://www.educative.io/edpresso/countvectorizer-in-python
- ➢ Harrison, O. (2019, July 14). *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. Medium. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761