УДК 004.9+616.24

V. A. Kovenko, I. V. Bogach, M. V. Baraban

ITEM-BASED COLLABORATIVE FILTERING BASED ON NLP TECHNIQUES

Vinnytsia National Technical University, Vinnytsia

Abstract. The benchmark approach to content-based recommendation systems is exposed in this article. The usage of Word2Vec embeddings made by Google is unleashed. The opportunity of using additional business logic is considered.
Keywords: NLP techniques, Word2Vec, CountVectorizer, cosine similarity, embeddings, content based system, content based recommendation, user based recommendation.
Аннотация. В данной работе представлен подход к рекомендательным системам на основе каталога. Показано использование Word2Vec эмбеддингов, представленных Google. Описана возможность использования дополнительной бизнес логики.
Ключевые слова: NLP подходы, Word2Vec, CountVectorizier, косинусная похожесть, ембеддинги, система основаная на каталоге, рекомендации основанные на каталоге, рекомендации основанные на пользователях.
Анотація. У цій роботі представлено алгоритм для вирішення проблеми рекомендацій на основі каталога. Показано використання Word2Vec ембеддінгів, представленных Google.
Ключові слова: NLP підходи, Word2Vec, CountVectorizier, косинусна схожість, ембеддинги, система основана на каталозі, рекомендації основанні на каталозі, рекомендації основані на користувачах.
DOI: https://doi.org/10.31649/1999-9941-2021-51-2-17-22.

Introduction

Understanding users' preferences and proposing them the most relevant products is essential for every commercial business which involves the process of interacting with users. As nowadays the web infrastructure is developing rapidly, lot's of commercial activities move to the space of the internet. Thus, the demand for recommendation systems arises. Recommendation system is an engine which goal is to recommend relevant items to users. Many world famous companies like Netflix, Amazon, YouTube, etc, use them to attract more people to their websites and increase their income. The recommendation systems can be divided into two groups: content based [1] and user based [2]. Content based recommendation systems focus on the content, its taxonomy and metadata for making predictions, while the user based ones require user interactions like clicks or ratings the user left for items. Nevertheless user based recommendation systems are much more powerful than content based, they require lot's of computational power that can afford working with big data. On the opposite, when building a content based recommendation system, the one is interested only in the catalog of items, and as a rule the number of items is always smaller than the number of users in the system. Content based recommendation system is a nice start for a small company that just appeared on the market of web products. With the development of the sphere of Natural Language Processing (NLP), new opportunities for content-based recommendation systems appeared. The new approach to recommendation systems is proposed in this article, the problem is stated as measuring the similarities between items' metadata and is addressed as an NLP task. The system uses a hybrid algorithm based on counting words in a statement and Word2Vec model provided by Google. The possibility of using an additional business logic is considered. Finally, the results are viewed with respect to Movielens dataset.

Data preparation

Constructed system was validated using the famous Movielens [1] dataset, which contains catalogues of movies, their metadata and intersections of users with a catalog. The columns that were used are the following: title, movie_id and genres (Fig. 1).

As the constructed algorithm is based on content metadata – it was decided to enrich Movielens dataset with IMDB database to get more relative columns.

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

Figure 1 – Original Movielens data

As genres columns were already in our dataset, this field was enriched with more data from IMDB, other columns were just added (Fig. 2).

	title	year	titleType	genres	characters	directors	writers	cast
0	Toy Story	1995	movie	[Children, Fantasy, Adventure, Animation, Comedy]	["Mr. Potato Head", "Slinky Dog", "Buzz Lighty...	[John Lasseter]	[John Lasseter, Pete Docter, Andrew Stanton, J...	[Joel Cohen, Tom Hanks, Tim Allen, Don Rickles...
1	Jumanji	1995	movie	[Fantasy, Children, Adventure, Family, Comedy]	["Sarah Whittle", "Alan Parrish", "Van Pelt","...	[Joe Johnston]	[Jonathan Hensleigh, Greg Taylor, Jim Strain, ...	[Scott Kroopf, Robin Williams, Kirsten Dunst, ...
2	Grumpier Old Men	1995	movie	[Romance, Comedy]	["John Gustafson", "Maria Sophia Coletta Raget...	[Howard Deutch]	[Mark Steven Johnson]	[Tak Fujimoto, Walter Matthau, Jack Lemmon, An...
3	Waiting to Exhale	1995	movie	Drama, Romance, Comedy]	["Bernadine Harris", "Robin Stokes", "Savannah...	[Forest Whitaker]	[Terry McMillan, Ronald Bass]	[Kenneth 'Babyface' Edmonds, Whitney Houston, ...
4	Father of the Bride Part II	1995	movie	[Comedy, Family, Romance]	["George Banks", "Annie Banks-MacKenzie", "Nin...	[Charles Shyer]	[Albert Hackett, Frances Goodrich, Nancy Meyer...	[Elliot Davis, Steve Martin, Diane Keaton, Mar...

Green - enriched Red - added

Figure 2 – Data after enriching with IMDB

Data transformation

The MovieLens data is made of strings which describe items' metadata, but for an algorithm to work the transformation of relative columns to the matrix of numbers is needed. For this purpose a hybrid transformation, made of CountVectorizer [3] and Word2Vec [4] model, is used. CountVectorizer is a technique of counting words in a sentence or text corpus. Each word is then represented by its frequency of appearance in a sentence (Fig. 3).

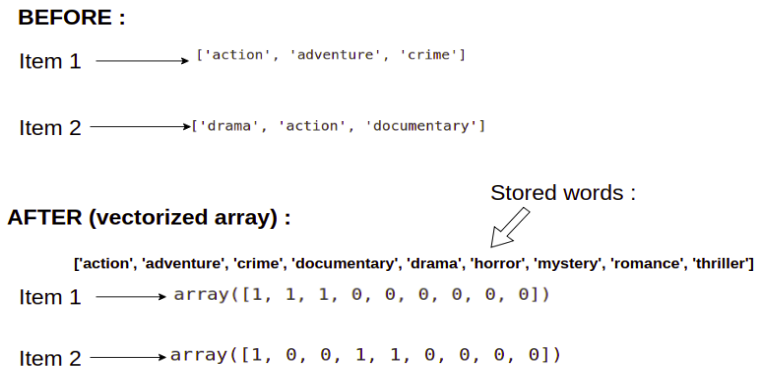


Figure 3 – The usage of CountVectorizer algorithm

That gives a possibility to represent word sequences as vectors of numbers. CountVectorizer is used only on columns that don't have a strong semantic context (directors, writers, cast, characters, etc.) For the fields with semantic context the other algorithm, known as Word2Vec is used. Word2Vec is a powerful machine learning algorithm that is used in the NLP sphere for semantic text classification, finding similarity between words, etc. Word2Vec learns the embedding space of words in which similar words (the ones that tend to appear in the same context) are closer to each other. Thus, the learned Word2Vec embeddings can be used to represent our columns in a numerical way and also save a semantic context of them (Fig. 4).

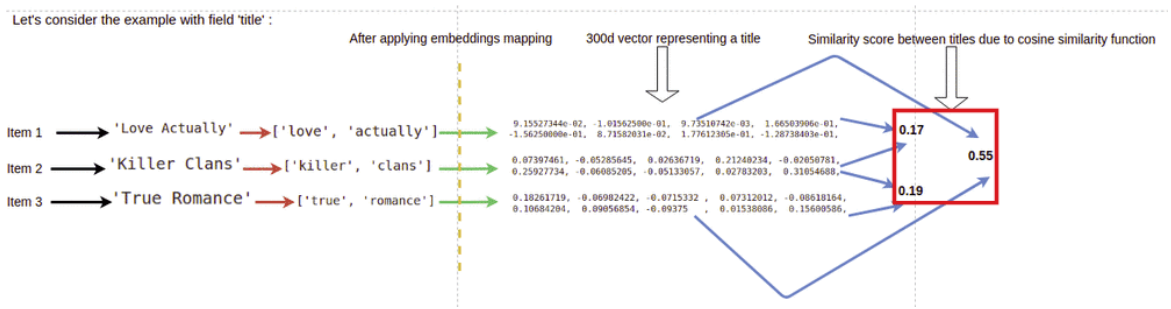


Figure 4 – Example of usage of Word2Vec model

For the purpose of representing not just a single word with Word2Vec, but the whole sentence – the embeddings of sentence's words are averaged. All the highlighted transformations give a possibility to represent content as nested matrices of number related to its metadata, thus the opportunity of using mathematical methods for finding similarity between content can be applied.

Algorithm

The algorithm is based on a cosine similarity, that is a mere choice for NLP tasks (1).

$$k(x, y)=xy^T / ||x|| ||y|| \quad (1)$$

The cosine similarity is computed between each column of each item in the dataset, thus for each item we have a matrix of similarities with others by a particular column. Because of the fact that one data column regarding an item can be much more important for the final recommendation than the other one, the additional business logic is added. The additional set of columns' weights that can be configured manually was added. Each attribute/column matrix is multiplied by a related weight parameter that gives an opportunity to decrease/increase contribution of it to the final similarity calculation. This makes the overall system more flexible and extendable towards new logic. Finally, the similarity matrices by columns are averaged to produce a final similarity matrix for an item (Fig. 5). Then, top N recommendations can be retrieved using a similarity score.

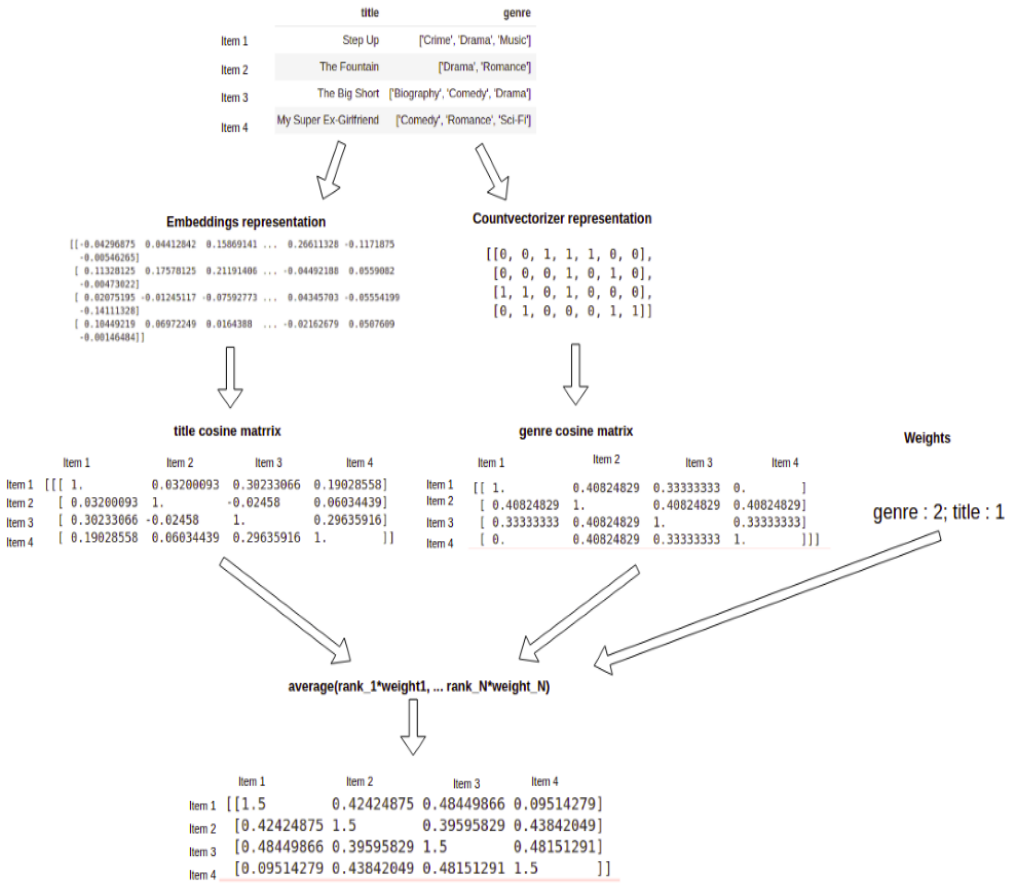


Figure 5 – Example of similarity calculation on two columns of data

Experiments

Experiments were conducted with respect to different weights sets. For each set of weights two experiments were done. For each experiment the item was randomly chosen and the top-4 related recommendations by similarity score were shown. From the first experiment (Fig. 6) it's obvious that given and predicted movies have similar titles by semantics and overlapping genres, as the highest weights were set to genre and title columns.

With different weights, the other picture appears. From the second experiment (Fig. 7), it's obvious that items have the biggest similarity by such fields as: characters, cast, directors, which is not a surprise if to look at weights.

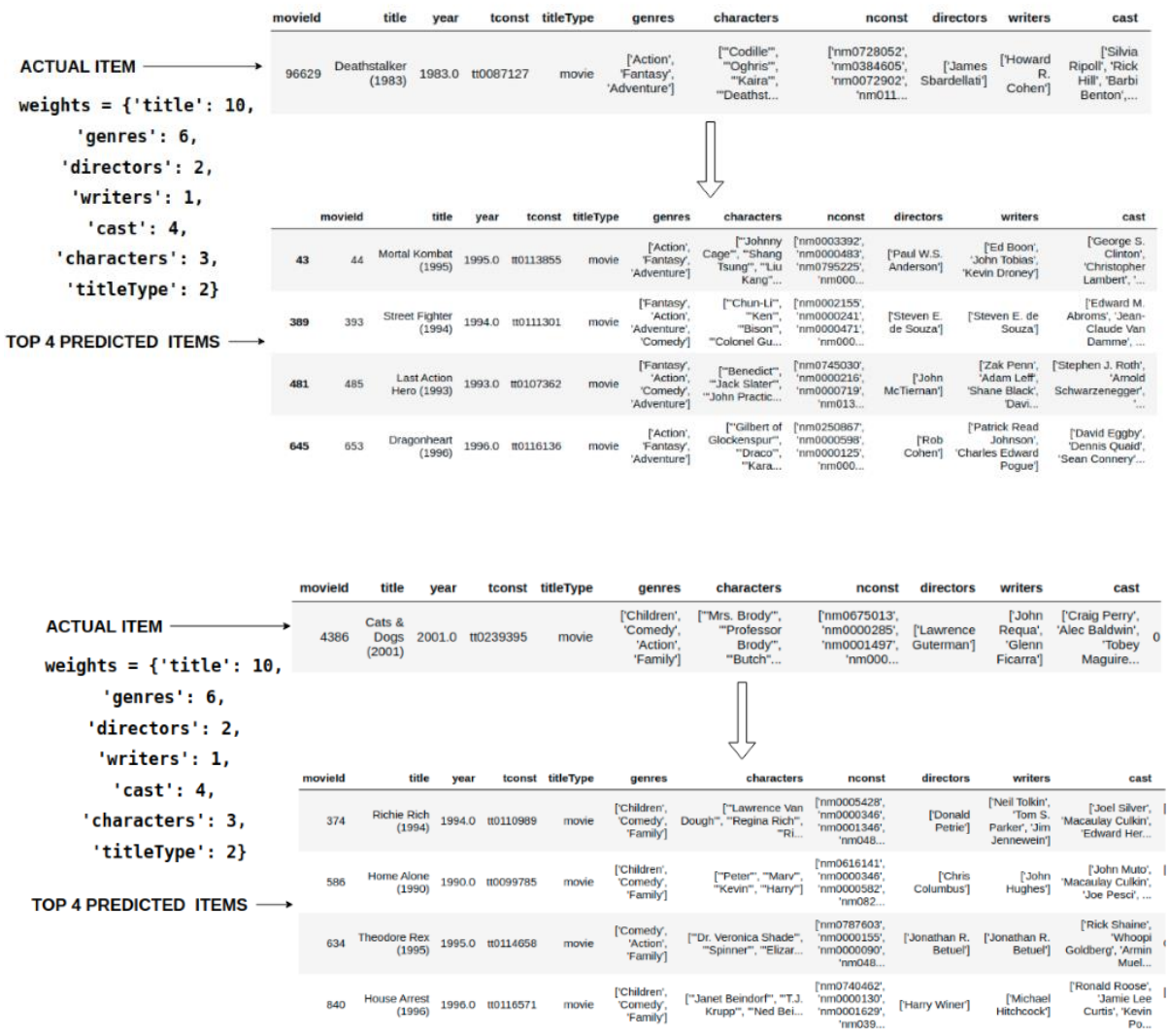
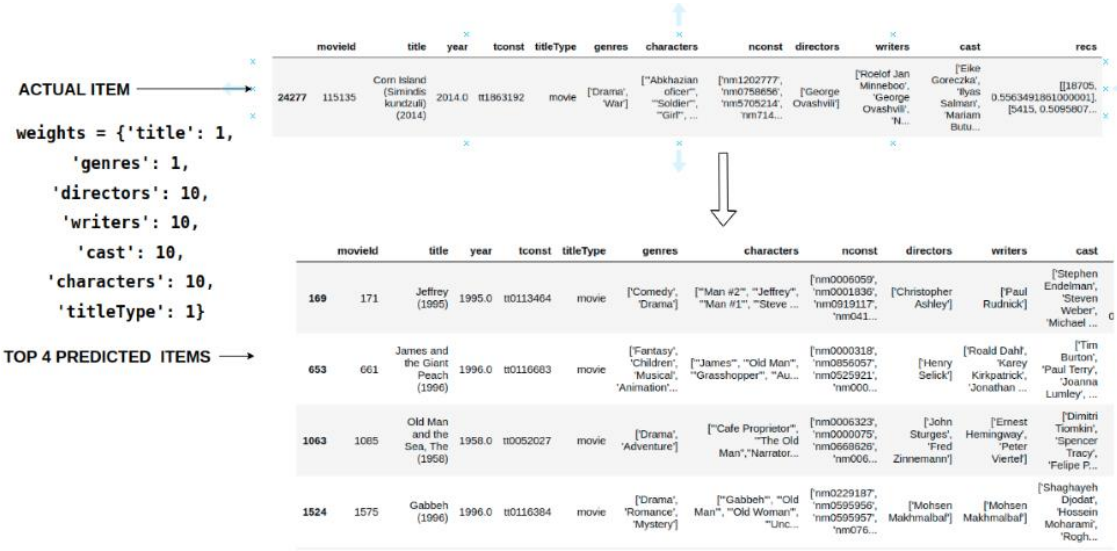
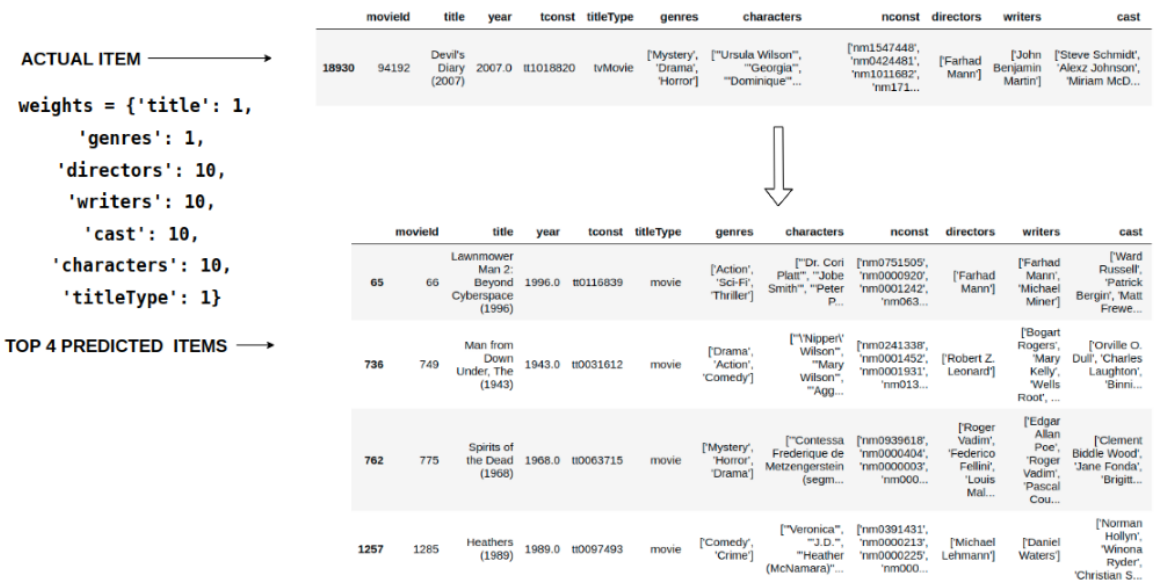


Figure 6 – The first experiment with different weights sets





	movieId	title	year	tconst	titleType	genres	characters	nconst	directors	writers	cast
18930	94192	Devil's Diary (2007)	2007.0	tt1018820	tvMovie	['Mystery', 'Drama', 'Horror']	['"Ursula Wilson"', '"Georgia"', '"Dominique"'...	['nm1547448', 'nm0424481', 'nm1011682', 'nm171...	['Farhad Mann']	['John Benjamin Martin']	['Steve Schmidt', 'Alexz Johnson', 'Miriam McD...

	movieId	title	year	tconst	titleType	genres	characters	nconst	directors	writers	cast
65	66	Lawnmower Man 2: Beyond Cyberspace (1996)	1996.0	tt0116839	movie	['Action', 'Sci-Fi', 'Thriller']	['"Dr. Cori Platt"', '"Jobe Smith"', '"Peter P...	['nm0751505', 'nm0000920', 'nm0001242', 'nm063...	['Farhad Mann']	['Farhad Mann', 'Michael Miner']	['Ward Russell', 'Patrick Bergin', 'Matt Frewe...
736	749	Man from Down Under, The (1943)	1943.0	tt0031612	movie	['Drama', 'Action', 'Comedy']	['"\'Nipper\' Wilson"', '"Mary Wilson"', '"Agg...	['nm0241338', 'nm0001452', 'nm0001931', 'nm013...	['Robert Z. Leonard']	['Bogart Rogers', 'Mary Kelly', 'Wells Root', ...	['Orville O. Dull', 'Charles Laughton', 'Binni...
762	775	Spirits of the Dead (1968)	1968.0	tt0063715	movie	['Mystery', 'Horror', 'Drama']	['"Contessa Frederique de Metzengerstein (segm...	['nm0939618', 'nm0000404', 'nm0000003', 'nm000...	['Roger Vadim', 'Federico Fellini', 'Louis Mal...	['Edgar Allan Poe', 'Roger Vadim', 'Pascal Cou...	['Clement Biddle Wood', 'Jane Fonda', 'Brigitt...
1257	1285	Heathers (1989)	1989.0	tt0097493	movie	['Comedy', 'Crime']	['"Veronica"', '"J.D."', '"Heather (McNamara)"...	['nm0391431', 'nm0000213', 'nm0000225', 'nm000...	['Michael Lehmann']	['Daniel Waters']	['Norman Hollyn', 'Winona Ryder', 'Christian S...

Figure 7 – The second experiment with different weights sets

Summary and further work

Nevertheless, the proposed system is not based on user activity, it has lots of advantages. It's flexible, not power consuming, easy to extend and flexible. Addressing the issue of recommendation systems as an NLP task, gives a lift to usage of novel NLP techniques like BERT [5]. On the opposite side, it's much less powerful than a user based system and is heavily dependent on the quality of catalog metadata. The particular system can also be used in ensemble with a user based recommendation system to construct a hybrid system. To sum up, the proposed algorithm can be used as an alternative to user based system and is an adequate choice for companies which just started their activity on the web market.

References

[1] Content-Based Recommendation System [Electronic resource]. URL: https://www.researchgate.net/publication/236895069_Content-Based_Recommendation_Systems − Title from the screen.

[2] Recommendation Systems : User-based Collaborative Filtering using N Nearest Neighbors. [Electronic resource]. URL: https://medium.com/sfu-big-data/recommendation-systems-user-based-collaborative-filtering-using-n-nearest-neighbors-bf7361dc24e0 − Title from the screen.

[3] 10+ Examples for Using CountVectorizer [Electronic resource]. URL: https://kavita-ganesan.com/how-to-use-countvectorizer/#.XlF2wHUzaV4 − Title from the screen.

[4] Distributed Representations of Words and Phrases and their Compositionality [Electronic resource]. URL: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf − Title from the screen.

[5] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Electronic resource]. URL: https://arxiv.org/abs/1810.04805 − Title from the screen.

Стаття надійшла: 29.08.2021.

Information about the authors

Kovenko Volodymyr Andriyovych – student of the Department r of Automation and Intelligent Information Technologies.

Bogach Ilona Vitalievna – Ph.D., Associate Professor, Associate Professor of the Department of Automation and Intelligent Information Technologies.

Baraban Maria Volodymyrivna – Ph.D., Associate Professo of the Department r of Automation and Intelligent Information Technologies.

В. А. Ковенко, І. В. Богач, М. В. Барабан

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ НА БАЗЕ КАТАЛОГА С ИСПОЛЬЗОВАНИЕМ NLP ТЕХНОЛОГИИ

Винницкий национальный технический университет, Винница

В. А. Ковенко, І. В. Богач, М. В. Барабан

КОЛАБОРАТИВНА ФІЛЬТРАЦІЯ НА ОСНОВІ КАТОЛОГУ З ВИКОРИСТАННЯМ NLP ТЕХНОЛОГІЇ

Вінницький національний технічний університет, Вінниця