UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

E-learning text analysis for automated knowledge map creation

*Author:*Danylo Sahaidak

*Supervisor:*Kateryna Bondar

A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2022

Declaration of Authorship

I, Danylo Sahaidak , declare that this thesis titled, “E-learning text analysis for automated knowledge map creation” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

E-learning text analysis for automated knowledge map creation

by Danylo Sahaidak

Abstract

The online learning process is constantly integrating into our lives. The pandemic and global lockdown only speed up this process. E-learning content is spread all around the internet, and it is often available for free. At the same time, it lacks structure, and it is hard to find essential materials. This thesis will discuss the possible ways of e-learning content analysis and categorization for the following integration in the knowledge map.

Acknowledgements

Firstly, I want to thank my supervisor Kateryna Bondar for opportunity to be part of Classe project. I am deeply grateful to the Clasee team who were working on this project with me. These people are Evgeniy Mashyna and Dmytro Fedoriuk....

Contents

List of Figures

List of Tables

For a person who can write 30+ pages from nothing, Ganna Mazurak…

Chapter 1

Introduction

1.1 Motivation

Education is one of the essential parts of modern people's lives. More and more people are getting primary education, and it is estimated that 30% of the world population will be receiving higher education by the year 2100 [Roser and Ortiz-Ospina, 2016]. From a young age, we are studying. On average, people spend 8.5 years in school and may spend even more to receive higher education and become specialists in a particular sphere. Such an investment comes at a high cost, and this is not the end. Our world is changing very fast. A person cannot get a degree and be confident that their profession will be in demand throughout life. So, we have to continue studying and improving our skills and expertise for the whole life.

On the other hand, the COVID-19 pandemic forced the educational system to change in the direction of online learning. After the lockdown, students lost the ability to study together in big groups; many people had to move back home and study there. Universities were not ready for such a change but had to implement the remote learning techniques in the short term, and not all went well. Online education results in poorer student learning compared to face-to-face education in general, but it is constantly evolving [Altindag, Filiz, and Tekin, 2021]. The pandemic had pushed e-learning to a new level.

Due to the increasing demand for efficient education, we got a new interest in online learning. The easiest way to get the knowledge is through platforms such as Coursera, edX, and Udemy. Their advantages are:

- ease of access to information;
- flexible schedule;
- ease of finding the community;
- ability to get the official certificates for course completion.

Besides the general online educational platforms, the e-learning content is well spread on social media. With their enormous auditory and openness, YouTube, Facebook, and even TikTok have become popular places to share knowledge. Such platforms attract young people with short texts or videos and the possibility to learn something new in a common and convenient way. But such medias often lack content and structuring. It's hard to follow the progress and find additional material.

In this work, we want to explore possible solutions that may help analyze the e-learning content.

1.2 Goals of the master thesis

This thesis aims to:

1. Categorize the e-learning text content.
2. Discover the topics described in the material.
3. Prepare the content for creation knowledge map based on it.

Chapter 2

Background and Related works

2.1 E-learning content analysis

In resent years e-learning became popular topic for researchers. This was facilitated by the Covid-19 pandemic and following popularity growth of online learning platforms.

There were different approaches to analyze e-learning content for further integration into recommendation systems [Khanal et al., 2020] or getting analytics about content quality or student satisfaction with his experience [Chanaa et al., 2021].

2.2 Content analysis techniques

We can distinguish three primary techniques for content analysis: text categorization, text summarization, and keyword extraction. Now let's consider each one and determine which will suit e-learning texts analysis the most.

Text categorization is a great technique that is highly efficient and well-researched. The methods include SVM (Support vector machines), Logistic regression, decision Tree, KNN, K-Means clustering, Hierarchical clustering, and others [Thangaraj and Sivakami, 2018]. In recent years we got a new step forward with the creation of huge pre-trained models like BERT [Devlin et al., 2018] and similar such as RoBERTa [Liu et al., 2019] or AlBERT [Lan et al., 2019], which can be used to do the classification task. It shows decent performance, great flexibility, and ease of configuration. The main drawback of such techniques is the requirement of predefined categories as well as the not-so-good flexibility of categorization models in case of categories change. But nested categories models may help with new categories addition. To sum up, categorization models are highly efficient in case of a known set of the categories and their immutability.

Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. [Allahyari et al., 2017]. Such techniques help shorten the amount of data needed to be processed, but it still requires automated or manual post-processing to determine content topics. It may lead to increased inaccuracy and reduced efficiency.

Keyword extraction is the automated process of extracting the most relevant words and expressions from the text. Such a technique can work well to predict the topics of the text based on the keywords. There are different methods for keyword extraction, including Linguistic-Based, Statistical-Based, Graph-Based Methods or other models [Sun et al., 2020]. The disadvantage of keyword extraction is the lack of context. Sometimes a word can mean completely different things. For example, in Astronomy and Astrophysics, the term metals may be considered for elements heavier than hydrogen and helium, while in Chemistry, it refers to a substance with high electrical conductivity, luster, and malleability which readily loses electrons to form positive ions (cations). So keywords will not work to classify text on its own.

2.3 Knowledge map

Mapping is a form of structuring the material and creating the visual representation. It helps get a clear idea about the information we have by giving visual context through maps or graphs. A knowledge map is an example of such representation. It helps understand connections and relationships between topics and their inheritance. Examples of such projects are "Galaxy OpenSyllabus" and "Open Knowledge Map".

Chapter 3

Methodology

3.1 E-learning content

To start, let's establish an understanding of e-learning content. It may be represented in different forms: animated videos, lectures, podcasts, posts in the news feed, and others. But for proper analysis, we need to reduce the diversity to a standard form. In this work, we will use plain text representation. Such text can be easily analyzed by machine learning techniques. In the case of video or audio, we can use the transcription. This may lead to the loss of the information presented in the video as an image like writings or slide text, but taking such information into account will make the analysis way more complicated.

3.2 Suggested model

For our purpose, we will consider a two-step text analysis process. In the first step, the text will be classified into base categories representing the subjects and global topics of learning. The categories may be nested and divided into different levels. We have wide categories like Culture, Arts, or Natural Science on the first level. Those wide categories may be narrowed to particular science or art on the second level. There may be multiple narrowing to achieve the necessary exactness. Such an approach can help add more categories after the model is created and trained.

After the categorization of a text, we will get some basic information about the discussed subject. Following this, we will run the keyword extraction model. It will help to explore the topics discussed in the text. We will consider such keywords as tags that may represent it. And with them, we can compare two texts and distinguish if they are on the same topic or not and how similar they are in terms of the closeness of their material.

With category and keywords extracted from a text, we can group the text with similar categories. After that, we can combine the keywords from texts in the same category into a set of topics. It is possible to create a graph of topics from this set. The possible structure of such a graph can be the following: the nodes will be keywords of the text, and the connection will be present between keywords from the same text.

We will not discuss the possible structure of the graph in this work and will only concentrate on text analysis.

3.3 Expected results

After analyzing the text, we expect to get two results: the category of the text and the set of discussed topics in the form of keywords.

Chapter 4

Dataset description

4.1 Resources

While considering the sources for e-learning content, the base properties to consider are:

- size: if the source is small, it can't be used to train models;
- content diversity;
- ease of access;
- structure of the content: it is preferable that content have a fixed structure;
- additional information about content: it may be helpful if the content has additional tags or other metadata.

With all these properties taken into account, in this work, we will consider two sources of e-learning materials:

- Wikipedia articles;
- YouTube videos;

4.1.1 Wikipedia

Wikipedia is a multilingual free online encyclopedia whose combined editions comprise more than 58 million articles [*Wikipedia* 2001]. It is a great source of strict and structured material. The articles are easily accessible and well divided into sections, which may be helpful as additional information about the text. It is easy to find content about any topic on Wikipedia. The disadvantages of Wikipedia as of e-learning content source is the lack of content for a particular topic. It is usually represented as a single article, most often too short to describe the topic. Also, the size of the article may vary: one article can contain several hundred words while the other several thousand.

4.1.2 YouTube

YouTube is an excellent source of free educational content. There are lots of popular science channels and creators talking about numerous topics, lots of how-to guides, and even some universities post their lectures as online courses there. Such a variety of content makes YouTube an almost perfect place for e-learning; it helps create a diverse dataset for NLP purposes. A great bonus is an automatically created transcript for videos. The main disadvantage of this platform is the lack of content verification. Any person may post a video there, which may lead to inaccurate and low-quality content. In November 2021, YouTube made the dislike counter private, making the situation only worse. Previously, we could count on the likes/dislikes relation to predict the quality of the video; right now, only the number of views and likes are available.

4.2 Categorization dataset

4.2.1 References

As the first step of text analysis, we will create global categories with subcategories of possible themes to differentiate the separate topics and narrow down the overlapping of essential knowledge topics. As the reference of categories, we can take the US Department of Education Classification of Instructional Programs codes [*CIP codes* 2020]. It was initially introduced in 1996 [Morgan, 1996] to collect, analyze, and disseminate statistics and other data related to education in the United States and in other countries. It has several editions, and the latest contains 47 main categories and more than 2000 subcategories that may describe almost any university program in the world. It has a nested structure with 3 levels, and each program has its own title and definition. The drawback of this system is its complexity, similarity of programs, and unintuitive list of program categories. There are many unpopular categories that will not help a lot in open e-learning materials categorization, but it is a decent reference point.

The other possible source of categories is universities' program courses divisions. They represent pretty much the same structure but often are more straightforward and intuitive. The disadvantage of such categories is that they do not cover the whole range of course topics, and usually, the university targets a particular sphere of knowledge.

Also, as a reference, we can take basic branches of academic disciplines [*Outline of academic disciplines* 2022]

4.2.2 Categories

As discussed previously, the nested categories are great for flexibility and the ability to change or upgrade the model afterward. At the first, global level of the categories should separate the whole information into some general categories that would cover most of the spheres. We will use 6 categories: Society, Culture and Arts, Technology, Social Science, Management, and Natural Science.

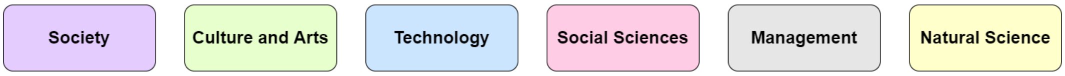

FIGURE 4.1: First level of categories, global categories

These topics would be too general and contain more information than we need during the standard exploration. For a second level, we will divide global categories into subcategories that would be more specific. There are 30 subcategories on the second level.

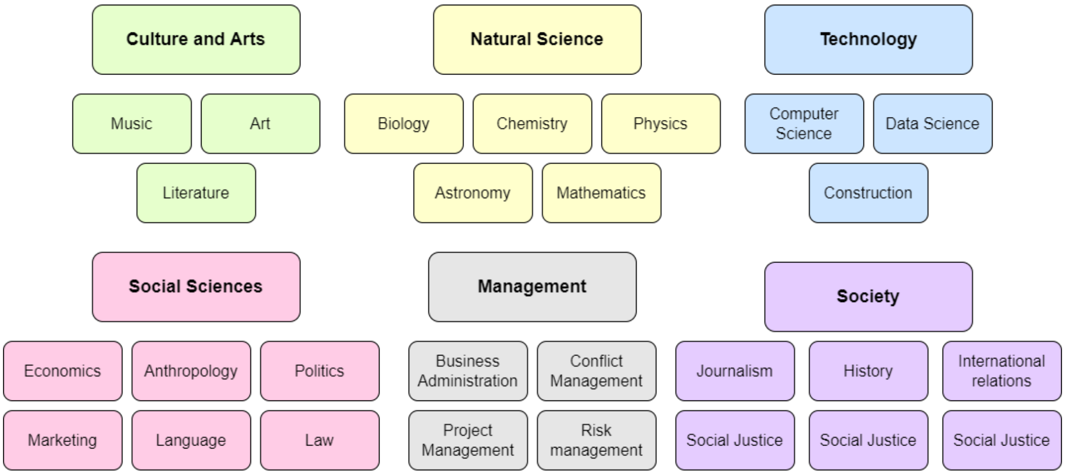

FIGURE 4.2: Two-level categories used for categorization

It is possible to divide each second-level category into more subcategories if needed, but at the same time, each subsequent level must require a more significant number of articles for analysis and a much more complex training process. Due to this, for the categorization task we will stick to two-level categories, where the first one is a global category and the second is a more specific subcategory.

Nevertheless, we will proceed with further, more detailed categorization, but for another purpose. The third level of categories will be used to create diverse sources of text for training.

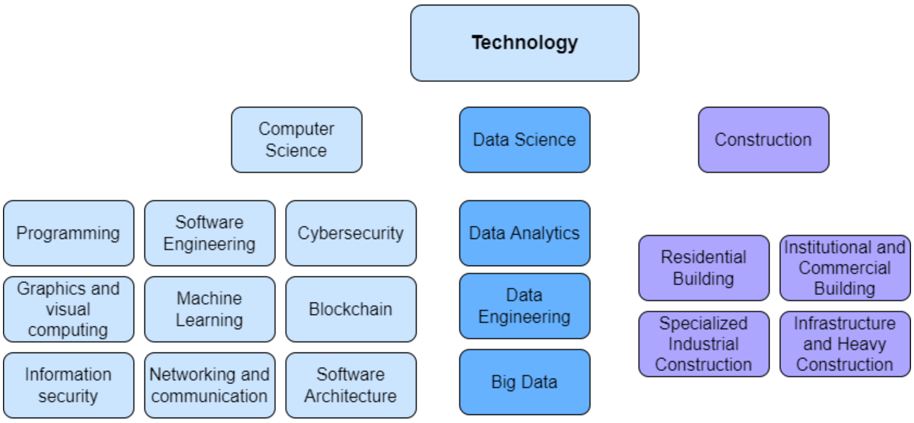

FIGURE 4.3: Third-level categories used for the creation of the train dataset

4.2.3 Categorization dataset

To train the categorization model, we need an appropriate amount of diverse train data representing different categories. To achieve this, we will create a third level of categories. The full categories tree can be found in Appendix 1. We will find one or several corresponding Wikipedia articles for each third-level category and at least three short YouTube videos. Such a dataset will have 115 low-level categories.

Firstly, we will parse each Wikipedia article and YouTube video. We will divide each article into paragraphs and, if needed, prepare it to a size smaller than 512 words as it is the maximum input text length for the BERT model. We will download the transcript of the YouTube videos and divide it into parts so that each one contains a complete sentence and is close to 512 words.

After such preparation, we will get around 10000 labeled text fragments. If any category is much smaller than others, we need to find additional sources of content that represent this category. It is not required as it may be corrected during training process, but it is beneficial.

Chapter 5

Experiments

5.1 Categorization techniques

For categorization purpose, we will experiment with the classic BERT model [Devlin et al., 2018]. It is convenient as it is pretrained and can be easily modified to serve as a classifier, as discussed in [Sun et al., 2019] and [Devlin et al., 2018].

We will use the classic BERT tokenizer and encoder and add a single softmax classifier and the end of the model. To get the relevant results, we need to fine-tune the softmax classifier. For this purpose, we will use the previously described categorization dataset.

We need to divide the dataset to train, validate and test parts with a ratio of 80:10:10. After nine epochs and several hours of training, we can evaluate the result on the test dataset.

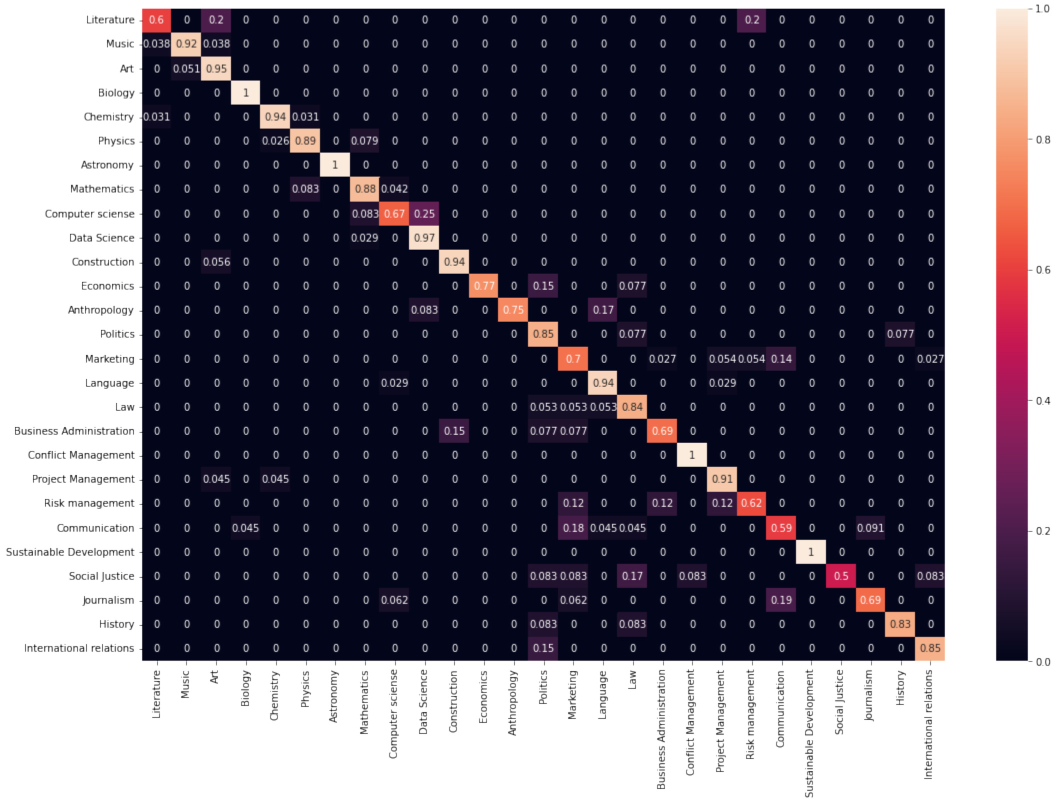

FIGURE 5.1: Confusion matrix for classification of second-level categories

The average weighted accuracy is around 85%, but it depends on the category. The worst result is observed with Social Justice: it is confused with Law and Conflict Management, and they are somehow similar topics. The other topics that have some part of inaccuracies are Communication, which is confused with Marketing. And Computer Science, which is confused with Data Science,.

All in all, those results are expected and the errors are not critical. The better results may be obtained with further diversification of the dataset and its amplification.

5.2 Keyword extraction techniques

For keyword extraction, we can use the BERT model as well. The possibility of using it for keyword extraction was researched in [Qian, Jia, and Liu, 2021], [Sahrawat et al., 2019], [Sharma and Li, 2019]. In this work, we will use the KeyBERT - a minimal and easy-to-use keyword extraction technique that leverages BERT embeddings [*KeyBERT* 2020]. It is a simple solution that helps evaluate the method and the pipeline of the analysis.

With KeyBERT, we can receive the n-grams of a specific length or in particular ranges of lengths. The model is such that, in general, a phrase is more likely to be chosen as a keyword than a single word. So, while extracting keywords from the text, we will run the model two times. At first, extract the phrases up to four words, and then extract single-word keywords.

After running the model, we can get the set of n-grams extracted from the text, but sometimes, they may be just do not have any meaning. As an example of raw keywords for a may look like this:

- 'map of physics', 0.65
- 'physics and mathematics', 0.6045
- 'of classical physics', 0.596
- 'physics describes how', 0.5775
- 'classical physics and', 0.5703
- 'of physics', 0.5671
- 'mathematics but physics', 0.5637
- 'physics and', 0.5636
- 'world of physics', 0.5564

The phrase like 'of classical physics', 'classical physics and', 'of physics', 'physics and' are not useful as they make no sense. So, after getting the keywords, we will keep only those that start and finish with nouns, verbs, or adjectives.

Also, when working with single-word keywords, we need to convert them to singular and present tense.

5.3 Real text analysis

After all, let's try running the complete analysis pipeline on some general internet content from different spheres like Art, Technologies, and Politics to see what are the actual results of the model.

5.3.1 YouTube videos

As an example of YouTube content, we will analyze the video from the channel "The School of Life" "What Rembrandt Can Teach Us About Love".

It is correctly categorized as Art with good confidence:

- ’Art’, 0.90;
- ’Literature’, 0.26;
- ’Law’, 0.17;
- ’Physics’, 0.15;
- ’Politics’, 0.15;
- ’Conflict Management’, 0.10;
- ’Music’, 0.10;
- other less then 0.1;

Extracted keywords:

- ’rembrandt painted self’, 0.69;
- ’rembrandt contemporaries’, 0.68;
- ’extravagant life rembrandt’, 0.63;
- ’rembrandt self’, 0.61;
- ’rembrandt van’, 0.57;
- ’rembrandt’, 0.56;
- ’life rembrandt van’, 0.55;
- ’rembrandt van rijn’, 0.54;
- ’painter’, 0.38;
- ’painting’, 0.35;
- ’portrait’, 0.32;
- ’extravagant’, 0.22;
- ’contemporaries’, 0.22;
- ’wastrel’, 0.22;
- ’famous’, 0.20;
- ’artists’, 0.20;
- ’self’, 0.20;
- ’van’, 0.19;
- ’gallery’, 0.19;

The model was able to recognize that the text was about Rembrandt and his famous paintings, including self portrait, but it didn’t recognize the other motive of the video about the expressions and feelings of the painter.

5.3.2 NASA news article

The next example is NASA news article: "Small Spacecraft Electric Propulsion Opens New Deep Space Opportunities".

- ’Astronomy’, 0.79;
- ’Physics’, 0.32;
- ’Project Management’, 0.24;
- ’Computer sciense’, 0.23;
- ’Chemistry’, 0.21;
- ’Construction’, 0.18;
- ’History’, 0.15;
- other less then 0.15;

Extracted keywords:

- ’spacecraft electric propulsion’, 0.58;
- ’electric propulsion technologies’, 0.57;
- ’electric propulsion ep’, 0.57;
- ’spacecraft electric’, 0.56;
- ’advanced electric propulsion’, 0.55;
- ’small spacecraft electric’, 0.55;
- ’electric propulsion technology’, 0.55;
- ’small electric propulsion’, 0.55;
- ’electric propulsion thrusters’, 0.55;
- ’power electric propulsion’, 0.54;
- ’propulsion’, 0.43;
- ’spacecraft’, 0.41;
- ’thruster’, 0.35;
- ’nasa’, 0.35;

- ’electric’, 0.35;
- ’thrusters’, 0.34;
- ’thrust’, 0.34;
- ’kilowatt’, 0.32;
- ’efficiency’, 0.32;
- ’moon’, 0.28;

The categorization was good; the model detects that it is about Astronomy and a little bit of physics. The keyword extraction part went really well. The main topic - new electric propulsion system for small spacecrafts was distinguished well. The other keywords amplified the topic.

5.3.3 Institute of the Study of War: Russia-Ukraine warning update

The last considered topic was a political article from February 22 about Russia-Ukraine relationships just before the war: "RUSSIA-UKRAINE WARNING UPDATE: RUSSIA LIKELY TO PURSUE PHASED INVASION OF UNOCCUPIED UKRAINIAN TERRITORY".

- ’International relations’, 0.69;
- ’History’, 0.41;
- ’Sustainable Development’, 0.32;
- ’Astronomy’, 0.26,
- ’Physics’, 0.19;
- ’Conflict Management’, 0.19;
- ’Law’, 0.15;
- other less then 0.15;

Extracted keywords:

- ’russian military operations’, 0.63;
- ’ukraine putin demands’, 0.62;
- ’russian military operation’, 0.62;
- ’military activity putin’, 0.61;
- ’operations against ukraine’, 0.60;
- ’eastern ukraine putin’, 0.58;
- ’deploy russian forces’, 0.58;
- ’authorization for putin’, 0.57;
- ’operation against ukraine’, 0.57;
- ’oblasts’, 0.45;
- ’putin’, 0.44;
- ’crimea’, 0.42;
- ’donetsk’, 0.42;
- ’ukraine’, 0.41;
- ’russian’, 0.38;
- ’russia’, 0.38;
- ’vladimir’, 0.37;
- ’ukrainian’, 0.33;
- ’kremlin’, 0.29;

It is correctly categorized as International Relationship, but with less confidence. The keywords help detect the narrative of the text and the topics discussed in it.

Chapter 6

Conclusions

In this paper, we have investigated the methods of e-learning content analysis, have suggested the possible solution and create a working model of it. The model helps to categorize the content and get the understanding about discussed topics. After extracting the keyword it is possible to integrate the content into recommendation system. The next steps may be the analysis of the big amount of e-learning content and building the knowledge map.

Appendix A

Categories tree

GlobalCategory	Subcategory	Topic
Culture and art	Literature	Fiction
Culture and art	Literature	Classics
Culture and art	Literature	Technical Literature
Culture and art	Literature	Contemporary literarure
Culture and art	Literature	Other
Culture and art	Music	Music History
Culture and art	Music	Classical Music
Culture and art	Music	Modern Music
Culture and art	Music	Musical instrument
Culture and art	Music	Other
Culture and art	Art	Painting
Culture and art	Art	Architecture
Culture and art	Art	Sculpture
Culture and art	Art	Dance
Culture and art	Art	Drama
Culture and art	Art	Performing-arts
Culture and art	Art	Cinema
Culture and art	Art	Other
Technology	Computer sciense	Theoretical Computer Science
Technology	Computer sciense	Computer Systems and Computational Processes
Technology	Computer sciense	Applied Computer Science
Technology	Computer sciense	Other
Technology	Data Science	Data Engineering
Technology	Data Science	Machine Learning and Deep Learning
Technology	Data Science	Big Data
Technology	Data Science	Other
Technology	Construction	Residential Building
Technology	Construction	Institutional and Commercial Building
Technology	Construction	Infrastructure and Heavy Construction
Technology	Construction	Other

TABLE A.1: All categories p.1

Natural Science	Biology	Anatomy
Natural Science	Biology	Botany
Natural Science	Biology	Zoology
Natural Science	Biology	Microbiology
Natural Science	Biology	Other
Natural Science	Chemistry	Organic Chemistry
Natural Science	Chemistry	Non-Organic Chemistry
Natural Science	Chemistry	Periodic table
Natural Science	Chemistry	Chemical industry
Natural Science	Chemistry	Other
Natural Science	Physics	Classical Mechanics
Natural Science	Physics	Thermodynamics and statistical mechanics
Natural Science	Physics	Electromagnetism
Natural Science	Physics	Relativity
Natural Science	Physics	Quantum mechanics
Natural Science	Physics	Other
Natural Science	Astronomy	Astrophysics
Natural Science	Astronomy	Astrochemistry
Natural Science	Astronomy	Stellar astronomy
Natural Science	Astronomy	Other
Natural Science	Mathematics	Algebra
Natural Science	Mathematics	Geometry
Natural Science	Mathematics	Probability and statistics
Natural Science	Mathematics	Other
Social Science	Economics	Microeconomics
Social Science	Economics	Macroeconomics
Social Science	Economics	Economic Theories
Social Science	Economics	Other
Social Science	Anthropology	Physical Anthropology
Social Science	Anthropology	Linguistic Anthropology
Social Science	Anthropology	Ethnology
Social Science	Anthropology	Other
Social Science	Politics	Political Theory
Social Science	Politics	Public Law
Social Science	Politics	Public Administration
Social Science	Politics	Other
Social Science	Marketing	Market Research
Social Science	Marketing	Brand Management
Social Science	Marketing	Social Media
Social Science	Marketing	Product Marketing
Social Science	Marketing	Marketing tools
Social Science	Marketing	Other
Social Science	Language	Translation
Social Science	Language	Linguistics
Social Science	Language	Philology
Social Science	Language	Other
Social Science	Law	Fundamental Law
Social Science	Law	Natural$_{law}$
Social Science	Law	Special Law
Social Science	Law	Other

TABLE A.2: All categories p.2

Management	Business Administration	Corporate management
Management	Business Administration	Startups
Management	Business Administration	Entrepeneurship
Management	Business Administration	Other
Management	Conflict Management	Conflict resolution
Management	Conflict Management	Other
Management	Project Management	Agile
Management	Project Management	Waterfall
Management	Project Management	Other
Management	Risk management	Internal control
Management	Risk management	Other
Society	Communication	Media
Society	Communication	TV
Society	Communication	News agencies
Society	Communication	Social media
Society	Communication	Other
Society	Sustainable Development	Economic Sustainability
Society	Sustainable Development	Social Sustainability
Society	Sustainable Development	Environmental Protection
Society	Sustainable Development	Other
Society	Social Justice	Discrimination
Society	Social Justice	Justice and equity
Society	Social Justice	Other
Society	Journalism	Political Journalism
Society	Journalism	Media communication and digital production
Society	Journalism	Internet Journalism and blogging
Society	Journalism	Other
Society	History	World History
Society	History	History of the USA
Society	History	Other
Society	International relations	Foreign policy and national security
Society	International relations	Diplomacy and international cooperation
Society	International relations	Security in international relations
Society	International relations	Other

TABLE A.3: All categories p.3

Bibliography

Allahyari, Mehdi et al. (2017). “Text summarization techniques: a brief survey”. In: *arXiv preprint arXiv:1707.02268*.

Altindag, Duha Tore, Elif S Filiz, and Erdal Tekin (2021). *Is Online Education Working?* Working Paper 29113. National Bureau of Economic Research. DOI: 10.3386/w29113. URL: http://www.nber.org/papers/w29113.

Chanaa, Abdessamad et al. (2021). “E-learning Text Sentiment Classification Using Hierarchical Attention Network (HAN)”. In: *International Journal of Emerging Technologies in Learning (iJET)* 16.13, pp. 157–167.

CIP codes (2020). https://nces.ed.gov/ipeds/cipcode/browse.aspx?y=56. Accessed: 2022-05-30.

Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

KeyBERT (2020). https://maartengr.github.io/KeyBERT/. Accessed: 2022-05-30.

Khanal, Shristi Shakya et al. (2020). “A systematic review: machine learning based recommendation systems for e-learning”. In: *Education and Information Technologies* 25.4, pp. 2635–2664.

Lan, Zhenzhong et al. (2019). “Albert: A lite bert for self-supervised learning of language representations”. In: *arXiv preprint arXiv:1909.11942*.

Liu, Yinhan et al. (2019). “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692*.

Morgan, Robert L (1996). *Classification of instructional programs*. US Department of Education, Office of Educational Research and Improvement . . .

Outline of academic disciplines (2022). https://en.wikipedia.org/wiki/Outline_of_academic_disciplines. Accessed: 2022-05-30.

Qian, Yili, Chaochao Jia, and Yimei Liu (2021). “Bert-Based Text Keyword Extraction”. In: *Journal of Physics: Conference Series*. Vol. 1992. 4. IOP Publishing, p. 042077.

Roser, Max and Esteban Ortiz-Ospina (2016). “Global Education”. In: *Our World in Data*. https://ourworldindata.org/global-education.

Sahrawat, Dhruva et al. (2019). “Keyphrase extraction from scholarly articles as sequence labeling using contextualized embeddings”. In: *arXiv preprint arXiv:1910.08840*.

Sharma, Prafull and Yingbo Li (2019). “Self-supervised contextual keyword and keyphrase retrieval with self-labelling”. In.

Sun, Chengyu et al. (2020). “A review of unsupervised keyphrase extraction methods using within-collection resources”. In: *Symmetry* 12.11, p. 1864.

Sun, Chi et al. (2019). “How to fine-tune bert for text classification?” In: *China national conference on Chinese computational linguistics*. Springer, pp. 194–206.

Thangaraj, M and M Sivakami (2018). “Text classification techniques: a literature review”. In: *Interdisciplinary Journal of Information, Knowledge, and Management* 13, p. 117.

Wikipedia (2001). https://en.wikipedia.org/wiki/Wikipedia. Accessed: 2022-05-30.