Consequently, modern technologies are expanding the boundaries of coiled tubing in the oil and gas industry.

References:

1. Dyachenko Yu.G. Vorkshop – intensyvna forma profesijnogo navchannya / Dyachenko Yu.G., Zotova O.M. // Pedagogichnyj poshuk: ideyi, dosvid, praktyka: Zbirnyk metodychnyx pracz. Vypusk 3. – Poltava: PKNG PoltNTU, 2019.
2. Bileczkyj V.S. Osnovy naftogazovoyi spravy / V.S. Bileczkyj, V.M. Orlovskyj, V.I. Dmytrenko, A.M. Poxylko. – Lviv: Novyj Svit -2000, 2018.
3. Osnovy naftogazovoyi inzheneriyi: Pidruchnyk. / Bileczkyj V.S., Orlovskyj V.M.,Vitryk V.G. – Lviv: Novyj Svit -2000, 2019.
4. Koltyubingovi ustanovky ta yix zastosuvannya v naftogazovij promyslovosti / Zhuravchak V.Yu., Shhepansky`j M.I. SWorld – November 2017 Intellectual potential of the XXI century 2017. http://www.sworld.education
5. Mojshevy`ch L.R. Zastosuvannya kolony gnuchkyx trub u naftogazovij galuzi ta lovylni roboty z nymy // Nafta i gaz. Nauka – Osvita – Vyrobnycztvo: shlyaxy integraciyi ta innovacijnogo rozvytku: Materialy vseukrayinskoyi naukovo-texnichnoyi konferenciyi. – Drogobych: TzOV «Trek-LTD», 2016.

INFORMATION TECHNOLOGY FOR CONTENT SIMILARITY IDENTIFICATION IN NEWS TEXT CORPORA

Ph.D., Associate professor Petrasova Svitlana, Galkina Yana

National Technical University «Kharkiv Polytechnic Institute»
Ukraine

The aim of the research is to determine common information spaces, which represent matters of topical interest in news data streams, by modelling intelligence functions of understanding and classification of sense.

Nowadays, in conditions of constant growth of news data, the task of natural language processing is becoming more and more in demand. The difficulty of semantic analysis of natural language content, contained in news reports, is particularly caused by ambiguity and synonymy, proper to all the language levels. Therefore, it affects the identification of semantic similarity of linguistic units.

Generally, applied systems for semantic analysis require studying large volumes of natural-language texts. The results of the researches [1] of text corpora show that not certain words but regularly reproducible constructions as collocations are used in the process of communication.

In the broad sense, collocations are defined as the combination of two words or the usual use of certain tokens together. In the narrow sense, a collocation means not random but frequent co-occurrence of words in the text [2].

In modern computational linguistics there are several approaches for collocation extraction: semantic-syntactic, context-oriented, and corpus-oriented. In this research we apply both semantic-syntactic and corpus-oriented approaches

based on Fillmore's construction grammar and the notion of a collocation as a lexical combinatory of syntactically linked and frequently occurred elements.

Accordingly, the paper proposes the information technology for content similarity identification in news text corpora (fig. 1). The technology comprises the developed syntactic rules and MI-measure to detect collocations and the use of WordNet and the semantic similarity model to identify synonymous collocation words and text fragments in the corpus.

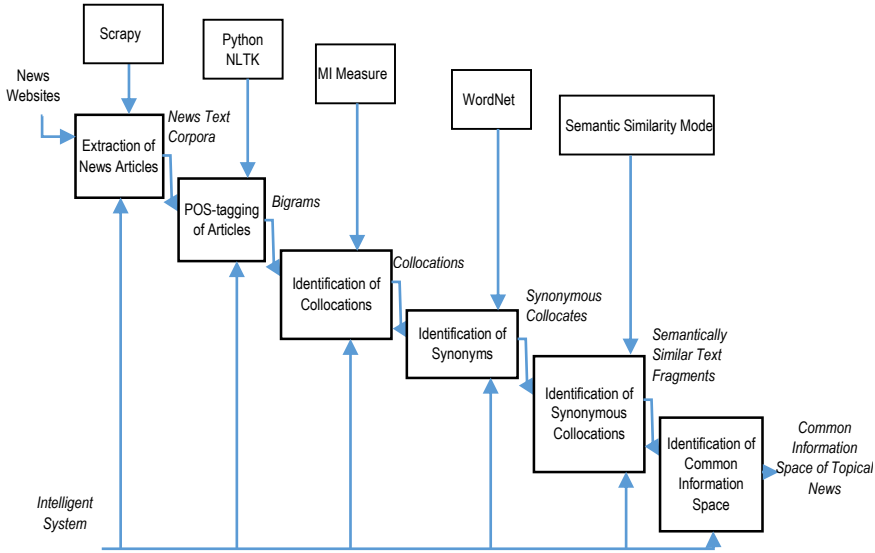

Fig. 1. **Stages of the Information Technology**

In order to determine semantically close content, the corpora of news texts have been developed. The corpora include the articles from BBC and CNN news websites [3, 4], extracted by Scrapy. Each corpus of texts consists of 100 000 words.

Using Python NLTK, we identify nouns ('NN', 'NNS', 'NNP', 'NNPS'), adjectives ('JJ', 'JJR', 'JJS'), verbs ('MD', 'VB', 'VBD', 'VBG', 'VBN', 'VBP', 'VBZ'). In this way, we extract a sequence of words that forms a syntactic construction (bigram): Noun + Noun, Adjective + Noun, Verb + Noun.

The result of the designed implementation (fig. 2) is extraction of bigrams from the corpus as well as the frequency of their occurrence in the texts.

So, the second stage of the information technology, i.e. extraction of potential collocations, brings establishment of the syntactic stability and the presence of word dependency.

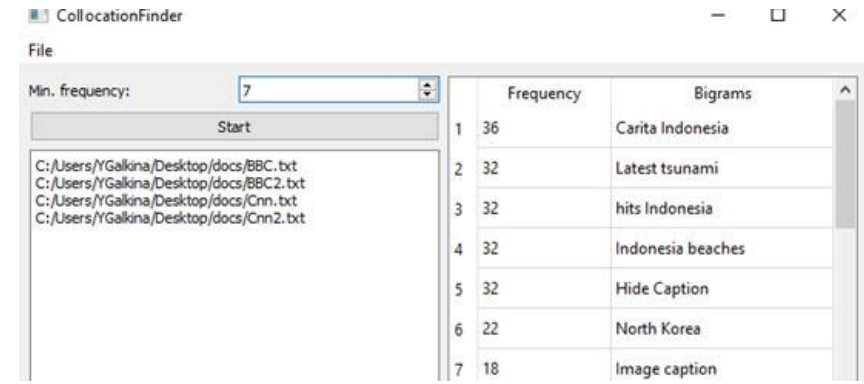

Fig. 2. **Implementation of Bigrams Extraction**

According to the corpus-oriented approach, we obtain substantive, attributive, and verbal collocations that represent the most frequently occurred meaningful text fragments. Conseque1ntly, the next stage of our technology is to compute mutual information (MI).

The coefficient of MI (1) compares the dependent context-related frequencies with independent ones (when words randomly appear in the context) [5]:

$$MI(n,c) = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}, \quad (1)$$

where:
n is a keyword;
c is a collocate;
f (n, c) is the frequency of the keyword n in pairs with the collocate c;
f(n), f(c) are absolute (independent) frequencies of the keyword n and the collocate c in the corpus;
N is the total number of words in the corpus.

Then, for extracting semantically close collocations, we apply WordNet 3.1.0 to identify synonymous collocates (collocation words) and the logical-linguistic model to identify semantically connected fragments. The model [6] is based on the use of algebraic-predicate operations and the predicate of semantic equivalence γ (2):

$$\gamma(x_1, y_1, x_2, y_2) = x_1^{VTr} y_1^{NObj} x_2^{Vtr} y_2^{NObj} \vee \vee (x_1^{NSubjOf} \vee x_1^{NSubj}) y_1^{NObj} (x_2^{NSubjOf} \vee x_2^{NSubj}) y_2^{NObj} \vee \vee x_1^{NSubj} (y_1^{AAtt} \vee y_1^{A\Pr}) x_2^{NSubj} (y_2^{AAtt} \vee y_2^{A\Pr}), \quad (2)$$

where:
VTr is a verb, transitive (a verb that can have a direct object);
NSubj is a noun, subject (the main word x in the substantive or attributive collocations);
NSubjOf is a noun, subject with the preposition "of";

NObj is a noun, object (the dependent word y in the substantive or verbal collocations);
AAtt is an adjective, attributive (an adjective used as an attribute before a noun);
APr is an adjective, predicative (an adjective used as a nominal part of the predicate).

Thus, text fragments are considered to be semantically similar in case the grammatical characteristics of collocates (synonymous collocation words) satisfy the predicate of semantic equivalence.

The example of the employment of the predicate of semantic equivalence to extract similar text fragments is shown in table 1.

Table 1

Extraction of Semantically Similar Text Fragments

BCC World News	CNN World News
A flooding emergency [$y_1^{NObj}x_1^{NSubj}$] in the Washington DC area left commuters [$x_1^{VTr}y_1^{NObj}$] in hazardous conditions. Torrential downpours led to road closures and left drivers [$x_1^{VTr}y_1^{NObj}$] stranded as well as dangerous flooding on the underground rail-lines.	A flash flood emergency [$y_2^{NObj}x_2^{NSubj}$] in Washington left roads submerged and cars [$x_2^{VTr}y_2^{NObj}$] stranded as heavy rains poured over the region.

As a result, the information technology for content similarity identification has been improved on the basis of the developed logical-linguistic model with the use of algebraic-predicate operations for the formalization of the grammatical characteristics of collocations. Establishing implicit semantic relations between news text fragments, the proposed technology allows identifying common information spaces of topical news (fig. 3).

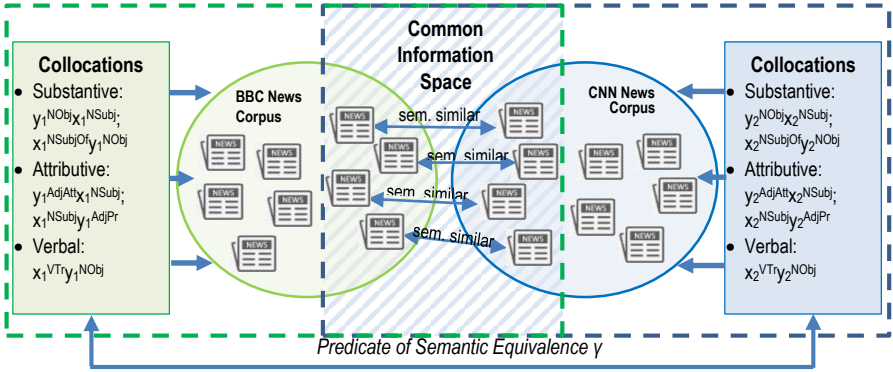

Fig. 3. **Identification of Common Information Spaces**

One of the future works might be implementation of the proposed technology in information retrieval, expert, or information-analytical systems.

References:

1. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, 2012. 294 p.

2. Sinclair J. Corpus, Concordance, Collocation. Oxford: Oxford University Press, 1991. 200 p.
3. BBC News. Available at: https://www.bbc.com/news.
4. CNN News. Available at: https://edition.cnn.com.
5. Evert S., Krenn B. Methods for the qualitative evaluation of lexical association measures. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France, 2001. p. 188–195.
6. Petrasova S., Khairova N., Lewoniewski W., Mamyrbayev O., Mukhsina K. Similar Text Fragments Extraction for Identifying Common Wikipedia Communities. Data (Special Issue: Stream Mining and Processing). MDPI AG, Basel, Switzerland, 2018. 3(4), 66. Available at: www.mdpi.com/2306-5729/3/4/66.

АДАПТИВНИЙ АЛГОРИТМ БІНАРИЗАЦІЇ ДЛЯ БІОМЕТРИЧНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ ЛЮДИНИ ЗА ВІДБИТКОМ ПАЛЬЦЯ

Гайдук Дар'я Володимирівна, Клочко Вікторія Олександрівна
Науковий керівник: старший викладач Трифонова К.О.
Одеський Національний Політехнічний Університет
Україна

Задача розпізнавання людини за відбитком є актуальною на сьогоднішній день, оскільки майже третина ринку біометричних технологій становлять саме системи за відбитками пальців. Сфери застосування таких систем різноманітні – правоохоронні органи, прикордонні служби, відділи кадрів, медицина, освіта, кібербезпека, фізичний доступ до об'єктів та інші.

Згідно алгоритму порівняння відбитків пальців за особливими точками, процедура вирішення задачі складається з наступних етапів: попередня обробка зображення, вилучення ознак – особливих точок, порівняння ознак.

Першою найважливішою задачею, від якої залежить якість наступних кроків алгоритму - це бінаризація зображення. Процес бінаризації представляє значний інтерес у задачах аналізу зображень, оскільки обробка бінарних зображень потребує менших ресурсів обчислювальної техніки та часових затрат, дозволяє виділити об’єкти, що містять необхідну інформацію.

За способом визначення порогу існуючі методи бінаризації поділяють на дві групи: глобальні (порогові) та локальні (адаптивні) [1]. У глобальних методах порогова величина обчислюється в процесі обробки всього зображення та залишається незмінною протягом процесу бінаризації, у локальних – зображення поділяють на області, в кожній з яких обчислюється локальна порогова величина.

Для дослідження якості виконання бінаризації зображення відбитка пальця обрано локальний метод Бредлі-Рота. Метод Бредлі-Рота заснований на використанні інтегрального зображення. Інтегральне зображення (також відоме як таблиця просумованих областей) – це матриця, яку