КІБЕРНЕТИКА та КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

УДК 004.9 DOI:10.34229/2707-451X.21.2.7

Ю.В. КРАК, А.І. КУЛЯС, В.М. ПЕТРОВИЧ, В.О. КУЗНЄЦОВ

ПРО МЕТОДИ КЛАСИФІКАЦІЇ ПРИХОВАНИХ КОНЦЕПТІВ МОВИ У СПЕЦІАЛІЗОВАНИХ ТЕКСТАХ ІЗ ЗАЛУЧЕННЯМ ПСЕВДООБЕРНЕННЯ, КЛАСТЕРИЗАЦІЇ І ГРУПУВАННЯ ОЗНАК

В роботі розглянуто задачу порівняння концептів мови у наукових текстах. Для обробки текстів сформовано корпус текстів, які перетворювалися за мірою TF-IDF у поєднанні з перетворенням Карунена – Лоева та Т-стохастичним групуванням найближчих сусідів (T-SNE). Отримана структура класифікатора прихованих концептів у вибірці наукових текстів із використанням дерев рішень, досягнуто точність розпізнавання (97–99 %) на зразках текстових даних. Досліджено стійкість до збурення вихідних даних варіаційним автокодувальником.

Ключові слова: *аналіз тексту, концепти мови, псевдообернення, кластеризація, групування ознак.*

Вступ. Однією із задач при науковому пошуку в спеціалізованих текстах є порівняння текстів за відповідністю певному критерію, а саме входження певної фрази або набору фраз у заданий науковий текст, щоб отримати результати, що містять жаданий набір фраз. Такий метод наразі успішно застосовується при пошуку інформації за ключовими словами, проте йому властиві недоліки – по-перше, необхідність перебору всього тексту і пошуку кожного ключового слова за заданим критерієм, що викликає появу результатів пошуку, які є не релевантними відносно шуканого тексту.

Дослідженню спеціалізованих та наукових текстів присвячена чимала кількість публікацій, серед яких варто відмітити наступні: в [1] досліджено амплітудні і фазові характеристики термінів для оцінки зосередженості категорій термінів у тексті, в [2] розроблено експериментальну інформаційну технологію аналізу частотних характеристик термінів та сполучень слів у тексті та побудовано модель Sub-Verb-Sub. Не зважаючи на наявність цих досліджень з даної тематики, не вирішеним є питання, що стосується, зокрема, аналізу частот появи окремих речень у тексті, зв'язку сенсу реферованого тексту (анотацій, рефератів) із відповідним текстом, статистичної близькості характеристик текстів, написаних різними авторами на одну тему та вивченню способів покращення результатів класифікації текстів у контексті задачі аналізу прихованих концептів мови, зокрема, із застосуванням методів групування і зменшення розмірності вектора характеристичних ознак. Саме тому дане дослідження – важливе і актуальне для розв'язування задач аналізу наукових текстів.

Постановка задачі. Виходячи із проведеного аналізу предметної області та виділення проблематики досліджень сформульовано такі задачі:

- сформувати вибірку зразків наукових текстів за різною тематикою;

• отримати представлення векторів ознак окремих речень у тексті;

• проаналізувати спорідненість текстів з анотацій текстів із вихідними текстами;

• дослідити представлення векторів ознак із залученням різних методів зменшення розмірності даних, кластеризації та групування ознак;

• оцінити інформаційну зосередженість змісту різних вибірок тексту за представленнями векторів ознак окремих речень тексту;

• провести класифікацію окремих речень текстів за тематикою;

• дослідити виділення прихованих концептів методами кластеризації;

• оцінити стійкість алгоритмів класифікації до перетворень розмірності ознак.

Отримання даних. У рамках дослідження сформовано три вибірки зразків наукових текстів з моделювання і розпізнавання комунікативної інформації за трьома тематиками: мімічні прояви, Українська жестова мова та тексти українською мовою (близько 1500 речень).

Таким чином, використовуючи близькість тематик, можна визначити як змінюється представлення цих текстів у просторі характеристичних ознак, що відповідно дозволить визначити спільні та відмінні концепти мови в цих текстах. Для розв'язування поставленої задачі наукові тексти подано у вигляді матриці, в якій кожному рядку відповідає окреме речення, включаючи і назву, анотацію, підписи до рисунків, висновки та інші текстові елементи, що містить даний текст.

На етапі попередньої обробки тексти аналізувалися синтаксичним аналізатором (стемером), який відсіював слова, які не мали суттєвого сенсу (стоп-слова) та відтинав афікси (суфікси та закінчення) слів.

Параметри середовища аналізу текстових даних. Для вивчення цих даних було створено модуль інтелектуального аналізу наукових текстів на мові Python із залученням бібліотеки інтелектуальної обробки текстових даних scikit-learn у середовищі Jupyter. Отриманий модуль тестувався на апаратній платформі із ОС Windows та наступними характеристиками: процесор Intel Core i5-6600k, ОЗП 8 ГБ DDR4.

Представлення вектора ознак. Кожен рядок матриці був поданий у представленні TF-IDF (text frequency inverse document frequency) [3], де кожному з елементів відповідав окремий термін та частота його появи. Отримані частотні характеристики порівнювалися, і з вектора ознак відсікалися усі терміни та відповідно ознаки, які не входили хоча б в один із текстів. Це дозволяло зберігати розмірність даних і враховувати лише співвідношення кількості наукових термінів, спільних для текстів із різною тематикою. Крім того, при використанні деяких методів зменшення розмірності даних (зокрема, Карунена – Лоєва) це дозволило зменшити кількість нульових елементів і відповідно ступінь розрідженості матриці зразків.

Статистична спорідненість анотацій у спеціалізованих текстах. Для оцінки статистичної спорідненості в першій групі експериментів запропоновано використати наступні показники: декартову відстань, критерій Пірсона та середньоквадратичне відхилення. Дані значення опосередковано вказують на різнорідність речень і тому на попередньому етапі процесу дозволяють опосередковано перевірити валідність вибірки зразків. У результаті випробувань при порівнянні окремих речень із окремо взятого наукового тексту та анотації із цього тексту, відмічено, що речення, які мали неоднаковий сенс відрізнялися за запропонованими параметрами.

На прикладі 1-го речення з анотації та основного тексту показано подібність і відмінність речень за трьома параметрами (табл. 1).

ТАБЛИЦЯ 1. Показники тексту для речень

№ з/п	Анотації			Основний текст		
	Відстань	Пірсон	Ст. відхилення	Відстань	Пірсон	Ст. відхилення
1	0	1	0.490698	11	0.425414	0.347540
2	10	0.618363	0.529891	4	0.854699	0.325396
3	11	0.409801	0.300327	12	0.501629	0.490698
4	17	0.002661	0.300327	14	0.334416	0.418213
5	12	0.541444	0.529891	11	0.425414	0.347540

Синтез лінійних систем для зменшення розмірності даних і розпізнавання структурованих текстів. Умова побудови лінійної системи [4] – незалежність окремих ознак та елементів даних, їм задовольняє такий набір даних, в якому ранг матриці дорівнює кількості зразків даних, що вектор ознак містить кількість ознак більшу, ніж кількість зразків даних. Варіюючи кількість ознак, здійснювався відбір найбільш значущих, щоб максимізувати ранг із урахуванням зв'язної структури тексту. Після виконання процедури відбору елементів даних та ознак виконується нормування матриці зразків за величиною математичного сподівання. Обчислення коваріаційної матриці дає змогу оцінити міру надлишкової інформації досліджуваного тексту, а її наявність викликана переважно внутрішньою власною структурою вибірки речень за тематикою.

Таким чином, алгоритм побудови лінійних систем передбачає ітераційну процедуру впорядкування зв'язків між елементами даних з урахуванням їх внутрішньої структури. Для зменшення розмірності вихідного вектора характеристичних ознак окремих речень, застосовано перетворення Карунена – Лоєва [5, 6]. Перевага даного підходу – суттєве зменшення розмірності вектора ознак, можливість відкидання не інформативних ознак і зменшення часу навчання методів класифікації, що розглядаються у роботі.

В результаті даного експерименту отримано набори ознак для вибірки зразків, що аналізувалися надалі. Цей розклад тестової вибірки зразків для першого корпусу текстів (міміка), показав, що 95 % енергії власних векторів містилося в перших трьох векторах, що дозволило візуалізувати просторове розташування векторів ознак елементів даних у вигляді тривимірної діаграми, що показано на рис. 1.

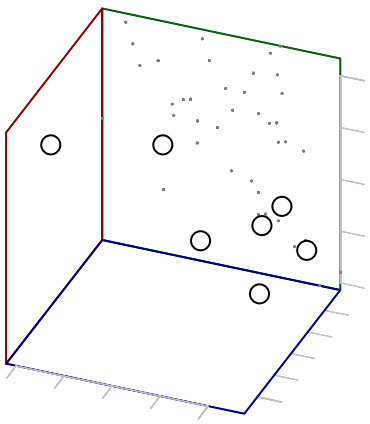


РИС. 1. Візуалізація перших трьох власних чисел перетворення Карунена – Лоева

Як видно з діаграми, по-перше, перші три власні числа формують окремі кластери точок, що дозволяє стверджувати про відмінність концептів у даних реченнях. По-друге, відстань між елементами із перевірочної вибірки (обведені кругом) та навчальної вибірки (інші точки) дозволяє візуально оцінити близькість зразків даних між собою, а, отже, застосовуватися для задачі порівняння прихованих параметрів тексту.

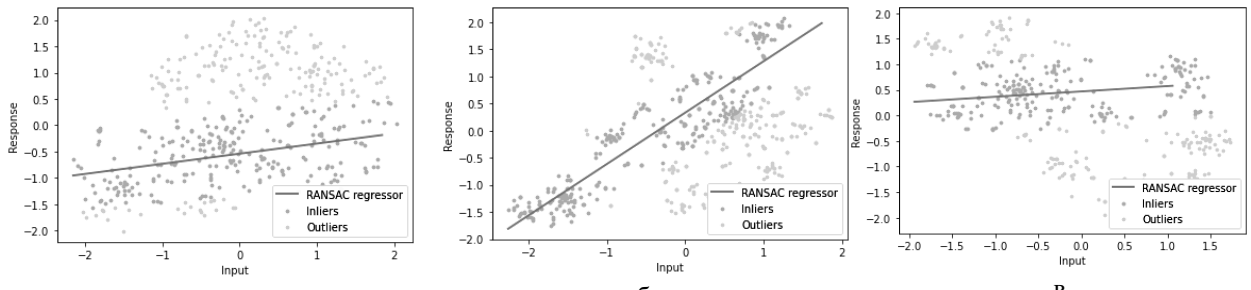
Групування і кластеризація ознак. Оскільки тривимірне подання не є зручним, було запропоновано виконати групування ознак для можливості візуалізації у вигляді плаского зображення, для чого додатково застосовувалося Т-стохастичне групування найближчих сусідів [5], щоб зменшити розмірність вектора ознак до двох. Для наочності, далі на рис. 2 показано представлення вектора ознак для другого корпусу текстів шляхом Т-стохастичного групування найближчих сусідів із контрастуванням кластерів точок методом K-середніх [4, 7].

РИС. 2. Представлення вектора ознак для другого корпусу текстів шляхом Т-стохастичного групування найближчих сусідів і кластеризації точок даним методом K-середніх

Інформаційна зосередженість змісту окремих речень. При порівняно великій кількості зразків у корпусі текстів і відповідно навчальній вибірці, на перший план мають виступати саме речення і їх розташування у просторі ознак [8]. Відповідно, тексти, що суттєво відрізняються за змістом, будуть мати велику кількість невходжень елементів-речень у кластери, що вказує на відмінність тематики тексту та прихованих концептів.

Для перевірки даної гіпотези запропоновано використати регресійну функцію – метод консенсусу випадкових зразків (RANSAC) [9]. На рис. 3 показано регресії на трьох вибірках наукових текстів у представленні ознак T-SNE.

Темно-сірим кольором на даному рисунку зазначено, які точки приймаються методом найбільш інформативними і використовуються для побудови регресії. Відповідно, порівнюючи графіки, можна вказати, що тексти мають спільну тематику, оскільки околи розташування точок, перетинаються. Крім того, темно-сірі точки на цьому рисунку також вказують на області з найбільшою інформаційною зосередженістю змісту кожного із корпусів.



а б в

РИС. 3. Представлення речень із трьох наборів текстів і їх інформаційна зосередженість

Класифікація речень із різних корпусів текстів. Проведено серію випробувань на T-SNE представленнях текстів із залученням методів інтелектуальної обробки даних [10]. Так, досліджувалися наступні методи: випадковий ліс, дерево рішень, метод найближчих сусідів, метод опорних векторів, наївний класифікатор Байєса, одношарова нейронна мережа. В рамках експерименту на отриманому наборі даних найкраще себе показав класифікатор Байєса (табл. 2).

ТАБЛИЦЯ 2. Ефективність класифікатора Байєса для трьох наборів даних

Dataset	Precision	recall	f1-score	support
Emotion	0.94	0.87	0.90	138
Gesture	0.82	0.94	0.87	163
NLP	0.95	0.87	0.91	159
accuracy	0.90	0.89	0.89	460
macro avg	0.90	0.89	0.89	460
weighted avg	0.90	0.89	0.89	460

На основі проведеного експерименту показано, що отриманий набір ознак TF-IDF дозволяє класифікувати тексти із високою достовірністю (87 %). Наявність помилок розпізнавання 1-го і 2-го роду (див. табл. 2) пояснюється великою спорідненістю текстів і близькості авторських стилів у наукових текстах, що передбачає застосування спільних термінів, а це видно з подібності представлень на рис. 3.

Класифікація текстів із використанням методів кластеризації. З метою вивчення можливості покращення результатів класифікації запропоновано наступний підхід: оскільки досліджувані тексти включають приховані концепти (підмножини) з категорії речень, які близькі за змістом, то виділення таких прихованих концептів дозволяє коректно поставити задачу класифікації текстів, а саме: класифікувати речення по їх мірі подібності до того чи іншого концепту тексту.

Для цього запропоновано використати одну із реалізацій методу K-середніх (mini-batch K-means), що найкраще підійшло для експериментальних даних. Отримані кластери даних обрано як позначки даних, що, таким чином дозволило перейти від класифікації за тематикою текстів до класифікації методом кластеризації із урахуванням внутрішньої будови даних, а саме прихованих концептів мови.

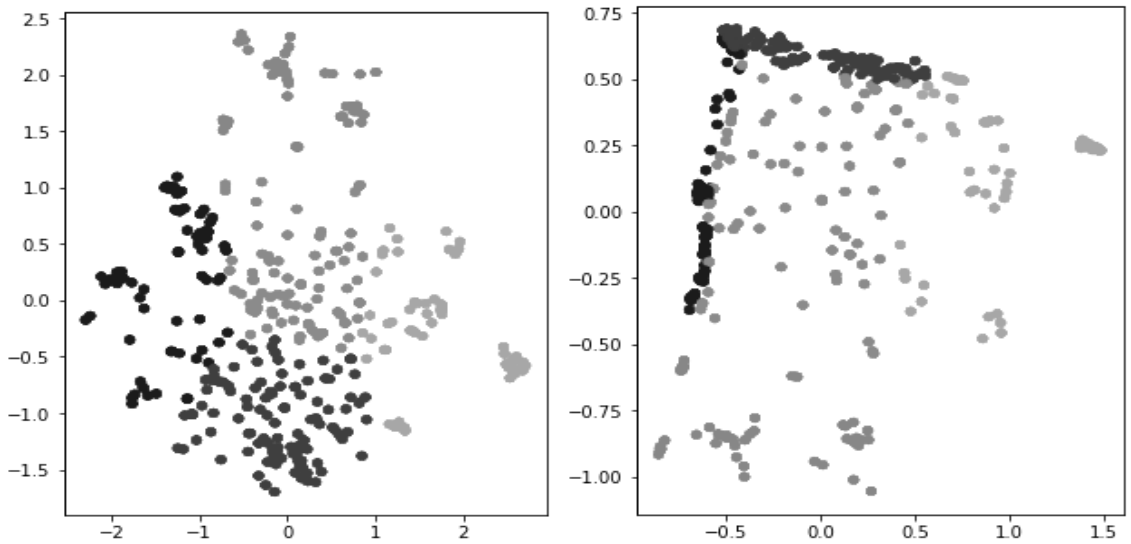
Як методи класифікації обрано такі: метод опорних векторів із лінійною і нелінійною гіпотезою, одношарову нейромережу, класифікатор Байєса, дерева рішень та споріднені методи, зокрема адаптивну машину підсилення і екстремальне підсилення градієнта. В результаті випробування показано, що на отриманому наборі даних із досліджуваних методів найбільшої точності розпізнавання прихованих концептів досягають метод опорних векторів із лінійною гіпотезою (97,4 %), одношарова нейромережа (97,4 %), випадковий ліс (99,1 %), дерево рішень із екстремальним підсиленням градієнта (99,1 %).

Стійкість алгоритмів класифікації до перетворень розмірності ознак. Для верифікації алгоритму проведено додатковий експеримент з дослідження стійкості визначення класів даних при внесенні збурення у вектор характеристичних ознак. Збурення виконувалося шляхом перетворення вихідного представлення T-SNE (двовимірне) у латентний простір ознак (двохвимірний), в якому елементам даних із вихідного простору відповідало розташування в латентному просторі ознак. Для цього використано варіаційний автокодувальник (variational autoencoder або VAE) [8].

Даний метод мінімізує величину середньоквадратичної похибки між вхідним (в представленні T-SNE) і вихідним набором даних (у латентному просторі ознак), і генерує додаткові елементи даних, що мають такий же розподіл, як вибірка навчальних зразків.

Для того, щоб досягти високої точності представлення даних, прихований шар автокодувальника має розмірність набагато вищу, ніж у звичайному застосуванні (прихований шар меншої розмірності).

Використовуючи перетворення розмірності, отримане автокодувальником, відмічено, що представлення ознак у латентному просторі (рис. 4) дозволяє побудувати лінійну гіпотезу при класифікації один напроти одного (один клас напроти іншого класу). Далі на рис. 5 показано гіпотези і позначки даних для представлення у латентному просторі ознак.



а б

РИС. 4. Ілюстрація елементів тестової вибірки у просторі ознак: а – T-SNE, б – VAE

Розглядаючи рис. 4 можна заначити, що відстань між елементами даних зменшилась. Відповідно це призвело до зменшення точності розпізнавання методами класифікації до 85 % для методу опорних векторів, а до 97,1 % – для дерева рішень із екстремальним підсиленням градієнта. Не зважаючи на зменшення точності розпізнавання, даний експеримент цікавий вивченням впливу збурення на точність розпізнавання різними методами класифікації.

Такий результат викликаний тим, що збурення вхідних даних варіаційним автокодувальником впливає на збіжність алгоритму навчання класифікатора. Варіаційний автокодувальник мінімізує середньоквадратичну похибку (дисперсію) між елементами даних. Дисперсія є вираженням величини збурень у даних і напряму залежить від характеру спадання енергії векторів ознак перетвореної матриці в латентному просторі ознак.

Знаючи величину зміни дисперсії при застосуванні автокодувальника можна оцінити кількість інформативних елементів досліджуваних даних. Особливістю латентного представлення є зменшення відстані між класами даних і полоси розділення (рис. 5), що в кінцевому результаті впливає на кількість ітерацій оптимізаційної процедури та на збіжність алгоритму класифікації.

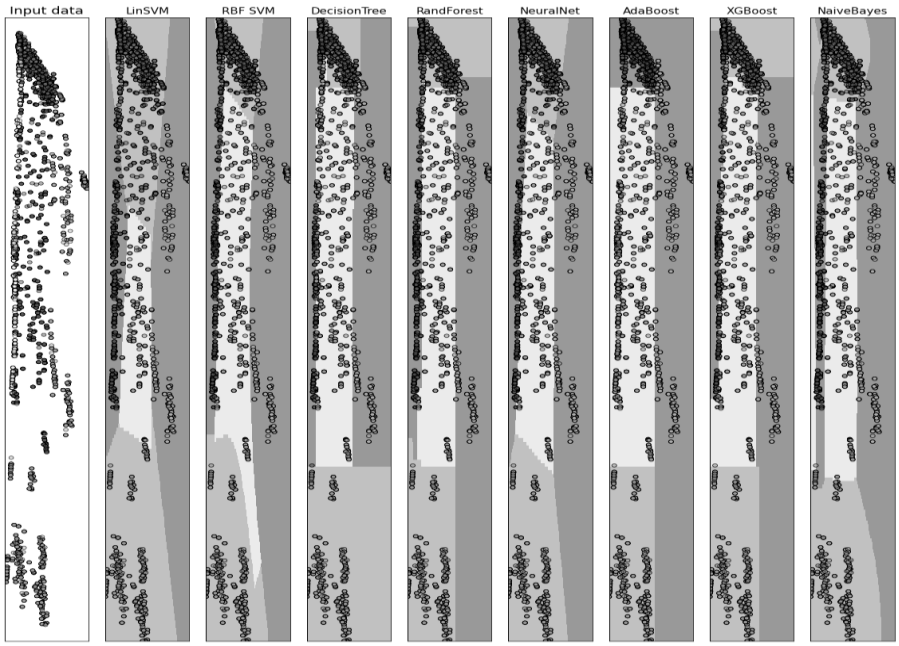


РИС. 5. Візуалізація гіпотез різних методів класифікації в латентному просторі ознак

Висновки. Застосування методів зменшення розмірності та групування даних вказало на наявність зосереджень точок даних, що може бути використано для оцінки авторського стилю за вживанням типових конструктів речень. Окремо варто вказати на отримані результати: при класифікації прихованих концептів із внесенням збурень у вихідні дані. Вони дають змогу визначити подальший напрямок досліджень, а саме: вивчення вибірки наукових текстів із іншими тематиками і авторськими стилями, застосування інших методів класифікації структурованих текстів.

Список літератури

1. Джурабаєв О.В., Бармак О.В., Манзюк Е.А., Скрипник Т.К. Інформаційна зосередженість змістовності в тексті. Вісник Хмельницького національного університетру. Сер. *«Технічні науки».* 2019. **4** (275). С. 80 – 83.
2. Бармак О.В., Мазурець О.В., Живілік А.В. Інформаційна технологія автоматизованого анотування та реферування цифрових текстів. Вісник Хмельницького національного університету. Сер. *«Технічні науки».* 2017. **4** (251). С. 147 – 158.
3. Робертсон С. Розуміння зворотної частоти документа: про теоретичні аргументи для IDF. *Journal of Documentation*. 2004. **60** (5). С. 503–520.
4. Крак Ю.В., Бармак А.В., Манзюк В.С. Информационная технология синтеза разделяющих гиперплоскостей для линейных классификаторов. *Проблемы управления и информатики.* 2019. № 1. С. 245–254.
5. Візуалізація даних за допомогою t-SNE. *Journal of Machine Learning Research*. 2017. **9**. С. 2595.
6. Кривонос Ю.Г., Кириченко М.Ф., Крак Ю.В., Донченко В.С., Куляс А.І. Аналіз та синтез ситуацій в системах прийняття рішень. Київ: Наукова думка, 2009. 336 с.
7. Хаст А., Нисьо Й., Марчетті А. Оптимальний RANSAC – до повторюваного алгоритму пошуку оптимального набору. *WSCG* . 2013. **21** (1). С. 21–30.
8. Гінтон Дж., Салакхутдинов Р. Зниження розмірності даних за допомогою нейронних мереж. *Science.* 2006. **313**. С. 504–507.
9. Крак Ю., Кручинін К., Бармак О., Манзюк Е. Візуальна аналітика в системах машинного навчання для ефективного прийняття рішень. Springer, 2020. С. 327–338.
10. Крак Ю.В., Кудин Г.И., Куляс А.И. Многомерное шкалирование средствами псевдообратных операций. *Кибернетика и системный анализ*. 2019. **55** (1). С. 47–57.

Одержано 24.02.2021

Крак Юрій Васильович,
член-кореспондент НАН України, доктор фізико-математичних наук, професор
Київського національного університету імені Тараса Шевченка,
провідний науковий співробітник, зав. відділом
Інституту кібернетики імені В.М. Глушкова НАН України, Київ,
https://orcid.org/0000-0002-8043-0785

Куляс Анатолій Іванович,
старший науковий співробітник, кандидат технічних наук, провідний науковий співробітник
Інституту кібернетики імені В.М. Глушкова НАН України, Київ,
https://orcid.org/0000-0003-3715-1454

Петрович Валентина Миколаївна,
кандидат технічних наук, старший науковий співробітник
Інституту кібернетики імені В.М. Глушкова НАН України, Київ,
Petrovych@nas.gov.ua

Кузнєцов Владислав Олександрович,
молодший науковий співробітник
Інституту кібернетики імені В.М. Глушкова НАН України, Київ.
kuznetsov.wlad@incyb.kiev.ua

UDC 004.9

Iurii Krak [1, 2 *], Anatoliy Kulias [1], Valentina Petrovych [1 *], Vladyslav Kuznetsov [1 *]

About Methods for Classifying Hidden Language Concepts in Specialized Texts Involving Pseudoinverse, Clustering and Data Grouping

[1] V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv
[2] Taras Shevchenko National University of Kyiv of Ukraine
[*] Correspondence: krak@univ.kiev.ua Petrovych@nas.gov.ua kuznetsov.wlad@incyb.kiev.ua

This paper discusses the problems of analysis of hidden language concepts in scientific texts in the Ukrainian language, using methods of text mining, dimensionality reduction, grouping of features and linear classifiers.

A corpus of scientific texts and dictionaries, as well as stop words and affixes, has been formed for processing specialized texts. The resulting texts were analyzed and converted into text frequency-inverse document frequency (TF-IDF) feature representation. In order to process the feature vector, we propose to use methods of dimensionality rteduction of the data, in particular, the algorithm for the synthesis of linear systems and Karunen – Loeve transform and grouping of features: T-stochastic grouping of nearest neighbors (T-SNE). A series of experiments were performed on test examples, in particular, for the determination of informational density in the text and classification by keywords in specialized texts using the method of random samples consensus (RANSAC). A method of classification of hidden language concepts was proposed, making use of clustering methods (K-means). As a result of the experiment, the structure of the classifier of hidden language concepts was obtained in structured texts was obtained, which gained a relatively high recognition accuracy (97 – 99 %) using such linear classification algorithms: decision trees and extreme gradient boost machine. The stability of the proposed method is investigated by using the perturbation of the original data by a variational autoencoder, test runs shown that sparse autocoder reduces the mean square error, but the separation band decreases, which affects the convergence of the classification algorithm.

In further research, we propose to apply other methods of analysis of structured texts and ways to improve the separability of specialized texts with similar authorial styles and different topic using a proposed set of parameters.

Keywords: text processing, language concepts, pseudoinverse, clusterization, methods of data groupings.