ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Кваліфікаційна робота

Метод нечіткої кластеризації коротких текстів

Виконав:
студент групи СПм-22-5
Стеблянко Б.О.

Керівник:
проф. Кучук Г.А.

Харків 2024

Актуальність теми

Кластерний аналіз є одним з найважливіших розділів системного аналізу даних і застосовується в різних проблемних областях – технічних, природничих, соціальних. Кластеризація є прикладом завдання навчання без вчителя і зводиться до розбиття вихідної множини об'єктів на підмножини класів таким чином, щоб елементи одного класу були максимально схожі між собою, а елементи різних класів - відрізнялися.

Традиційні методи кластерного аналізу працюють із об'єктами, заданими у вигляді векторів ознак. При роботі з текстами першим кроком алгоритму кластеризації є визначення простору ознак і побудова в ньому векторів наявних текстів. Як правило, одержувані вектори мають велику розмірність і при роботі з ними традиційні методи кластерного аналізу не забезпечують достатню ефективність. У разі роботи з короткими текстами розмірність векторів не зменшується, а лише додається властивість розрідженості до векторів ознак, що створює додаткові труднощі при їх обробці методами кластерного аналізу.

Метою роботи є розробка методу у системі підтримки прийняття рішень для кластеризації коротких текстів українською мовою з урахуванням експертної інформації .

Для досягнення поставленої мети необхідно вирішити такі **завдання**:

1) провести огляд сучасних методів чіткої та нечіткої кластеризації;

2) розробити метод кластеризації коротких текстів;

3) провести реалізацію та верифікацію методу нечіткої кластеризації коротких текстів.

Об'єктом дослідження є кластеризація наборів даних, що складаються з коротких текстів українською мовою та експертна інформація, що надходить під час інтерактивної обробки текстів.

Предметом дослідження є методи нечіткої кластеризації коротких текстів та обробки експертної інформації.

3

Основні завдання актуальні у сфері обробки коротких текстів

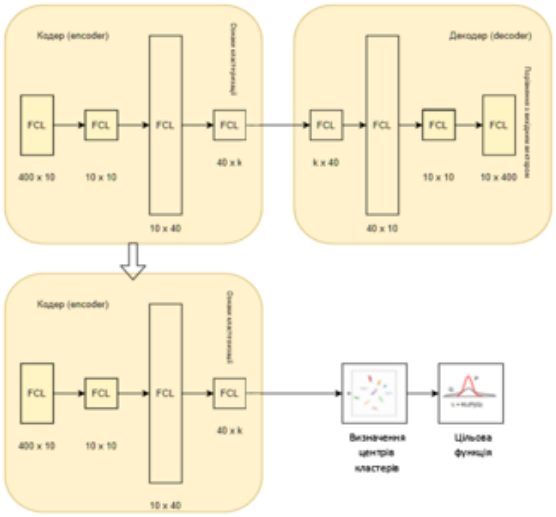
Завдання	Опис	Методи вирішення
Розпізнавання іменованих сутностей	Виділення з тексту дат, прізвищ, найменувань географічних об'єктів тощо.	Морфемний аналіз, Мовні моделі
Визначення тональності тексту	Поділ набору даних на тексти з позитивною та негативною тональністю, або за ширшим спектром емоцій	Визначення за ключовими словами
Класифікатори на базі мовних моделей	Класифікація текстів по заздалегідь визначеним тематкам	LDA, Мовні моделі

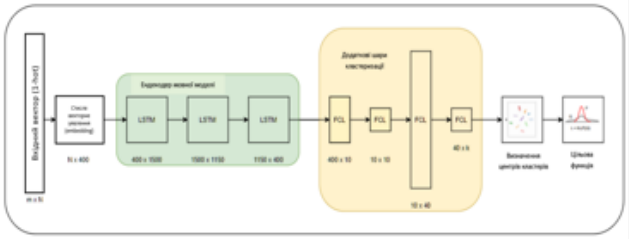
4

Метод нечіткої кластеризації

5

Схема роботи нейронної мережі для обробки текстів





Архітектура нейронної мережі для обробки текстів

N – кількість елементів у наборі.
k – задане число кластерів.

d-розмірність вхідного набору даних

N – кількість елементів у наборі.

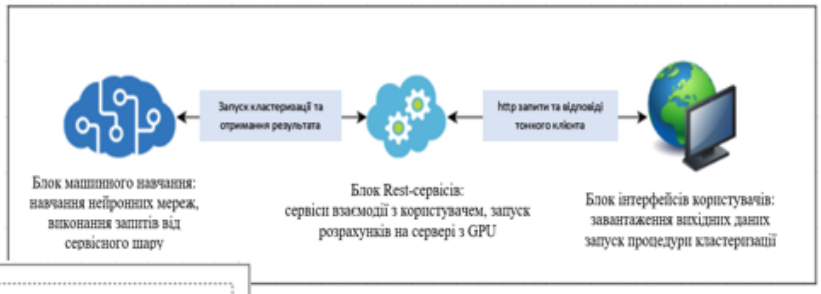
6

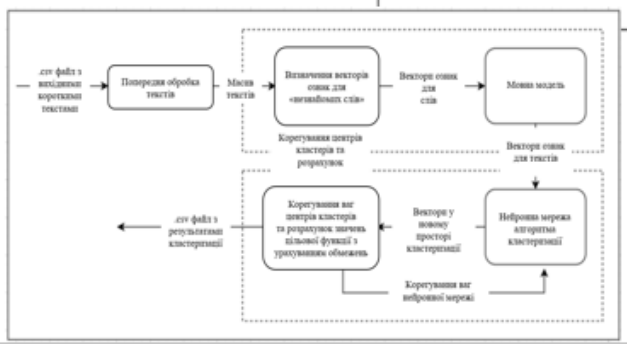
Метод інтерактивної кластеризації коротких текстів

1) Передобробка текстів: виправлення орфографії, токенізація.
2) Розширення словника мовної моделі.
3) Точне налаштування мовної моделі.
4) Ініціалізація шарів нейронної мережі блоку кластеризації.
5) Ініціалізація центрів кластерів (з використанням k- means).
6) Первинна кластеризація (синхронне навчання нейронної мережі кластеризації коротких текстів та визначення центрів кластерів).
7) Кроки інтерактивної кластеризації (цикл до досягнення прийнятної для експерта якості кластеризації).

7

Триланкова архітектура програмного комплексу для інтерактивної кластеризації коротких текстів

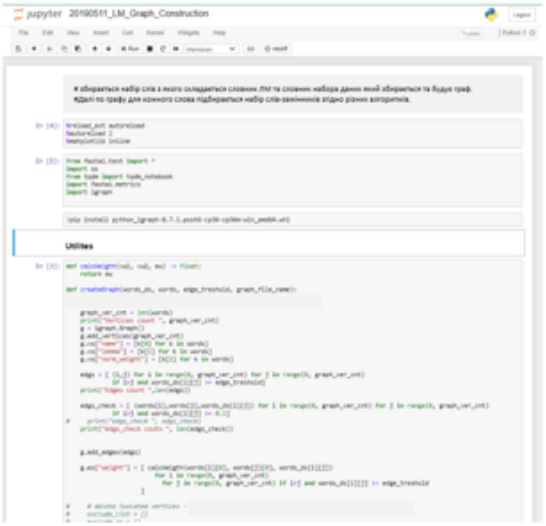




Функціональна структура блоку машинного навчання

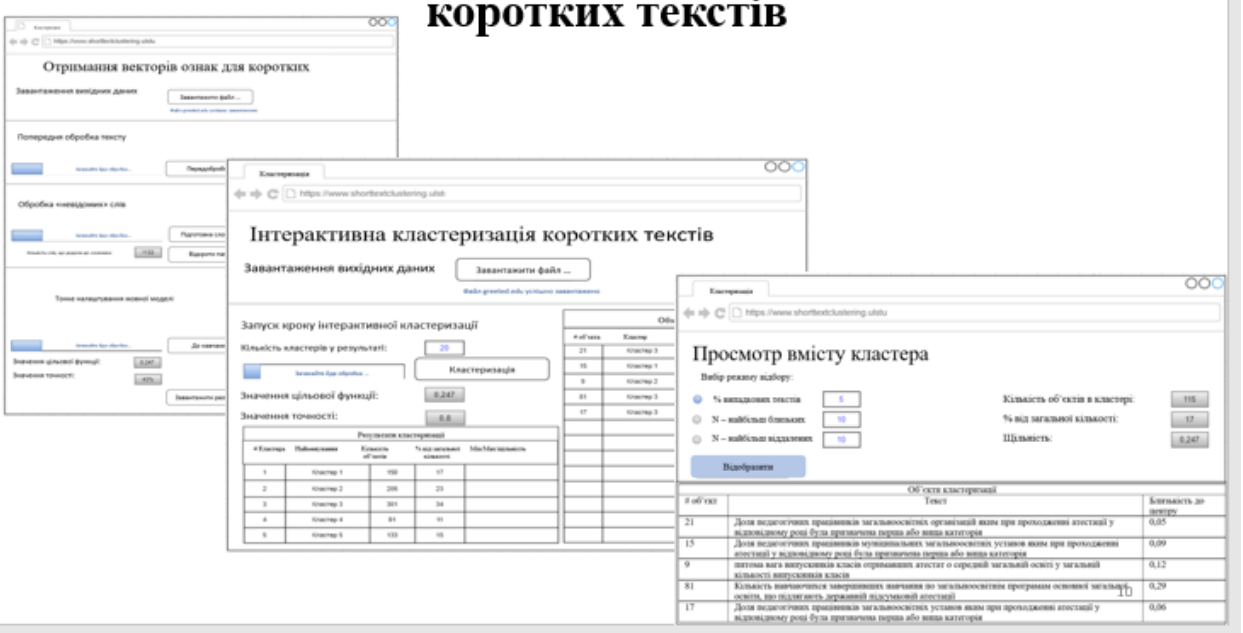
8

Модуль побудови векторів ознак для “невідомих” слів



9

Екранна форма отримання векторів ознак для коротких текстів

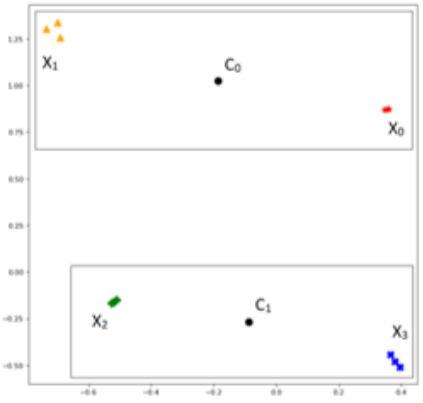


10

Верифікація методу

Список перших 4 векторів набору даних

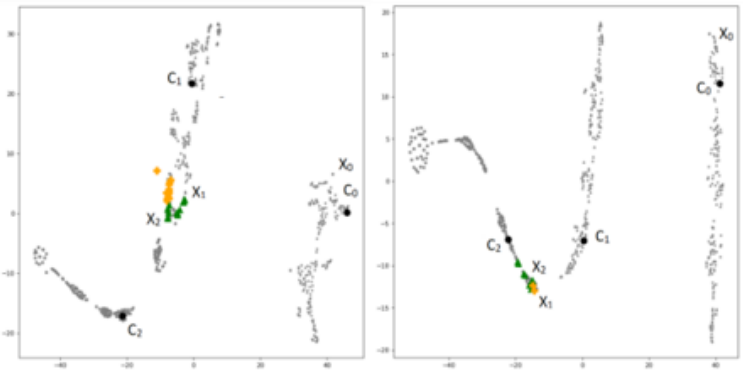
№	Координати				Кластер
0	1.0191519	0.06221088	0.04377278	0.07853585	C_0
1	0.07799758	1.0272592	0.02764643	0.08018722	C_0
2	0.09581394	0.08759326	1.0357817	0.05009951	C_1
3	0.06834629	0.07127021	0.03702508	1.0561196	C_1



Результат кластеризації для перших 12 векторів

11

Верифікація методу на класичному завданні класифікації та кластеризації «Іриси Фішера»



№	Координати				Кластер
0	5.119	3.562	1.443	0.278	C_0 (setosa)
1	7.077	3.227	4.727	1.480	C_1 (versicolor)
2	6.395	3.387	6.032	2.550	C_2 (virginica)

12

Висновки

Проведено дослідження моделей та методів машинного навчання для обробки текстів, та розроблено архітектуру штучної нейронної мережі, що реалізує кластеризацію на базі простору ознак мовної моделі української мови.

Розроблено метод обробки текстів для розширення словника мовної моделі на базі нейронної мережі з використанням нечіткого ієрархічного класифікатора, який дозволяє підвищити точність кластеризації в середньому на 10%.

Програмний модуль, що виконан у триланковій архітектурі, дозволив ефективно організувати спільну роботу експертів у ході рішення задачі нечіткої кластеризації набору даних коротких текстів

13


ISSN 2073-7394

Національний університет
"Полтавська політехніка імені Юрія Кондратюка"

National University
"Yuri Kondratyuk Poltava Polytechnic"

Системи управління, навігації та зв'язку

Control, navigation and communication systems

Випуск 2 (76)

Issue 2 (76)


FUZZY INTERACTIVE CLUSTERING METHOD

Дякую за увагу!

14

ДОДАТОК Б

Публікація

ISSN 2073-7394

Національний університет
"Полтавська політехніка імені Юрія Кондратюка"

National University
"Yuri Kondratyuk Poltava Polytechnic"

Системи управління, навігації та зв'язку

Control, navigation and communication systems

Випуск 2 (76)

Issue 2 (76)

Щоквартальне видання

Засноване у 2007 році

У журналі відображені результати наукових досліджень з розробки та удосконалення систем управління, навігації та зв'язку у різних проблемних галузях.

Засновник і видавець:
Національний університет
"Полтавська політехніка імені Юрія Кондратюка"

Телефон:
+38 (050) 302-20-71

E-mail редколегії:
kuchuk_nina@ukr.net

Інформаційний сайт:
http://journals.nupp.edu.ua/sunz

Quarterly

Founded in 2007

Journal represent the research results on the development and improvement of control, navigation and communication systems in various areas

Founder and publisher:
National University
"Yuri Kondratyuk Poltava Polytechnic"

Phone:
+38 (050) 302-20-71

E-mail of the editorial board:
kuchuk_nina@ukr.net

Information site:
http://journals.nupp.edu.ua/sunz

За достовірність викладених фактів, цитат та інших відомостей відповідальність несе автор

Журнал індексується міжнародними наукометричними базами: Index Copernicus (ICV = 82.05), General Impact Factor, Google Scholar, Academic Resource Index, Scientific Indexed Service

Затверджений до друку Вченою Радою Національного університету "Полтавська політехніка імені Юрія Кондратюка" (протокол від 30 квітня 2024 року № 5).

Свідоцтво про державну реєстрацію КВ № 24464-14404 ПР від 27.03.2020 р.

Включений до "Переліку наукових фахових видань України, в яких можуть публікуватися результати дисертаційних робіт на здобуття наукових ступенів доктора наук, кандидата наук та ступеня доктора філософії" до категорії Б – наказами МОН України від 17.03.2020 № 409 та від 09.02.2021 № 157

Полтава • 2024

Control, Navigation and Communication Systems. 2024. No. 2 ISSN 2073-7394

UDC 004.9 **doi: 10.26906/SUNZ.2024.2.152**

B. Steblyanko, O. Ni, H. Kuchuk, D. Volk

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

FUZZY INTERACTIVE CLUSTERING METHOD

Abstract. The article examines an example of a system in which a large number of short texts are generated. In it, participants create strategic planning documents, within which key performance indicators are determined. The formulations of key performance indicators form a data set consisting of short texts. Within the framework of this system, there is an urgent task of forming and updating a classifier based on this set. A solution to this problem is presented using the fuzzy interactive clustering method. This method allows expert to perform clustering sets of short texts, issuing reverse communication based on the results of each step interactive clustering. Collection procedure reverse does not imply any connection availability of an expert special knowledge about work neural network and is assembled in human-readable form matrices reverse communications. Such an approach has advantages over clustering methods requiring adjustments metaparameters algorithm not related directly with the clustering results. Also important advantage the proposed method is opportunity realize clustering sets data related to various language domains that do not match the domain on which was produced education language models, due to proposed extension method dictionary language models This property allows use the proposed algorithm in a narrow way specialized domains, as well as in domains that do not allow you to obtain a full-fledged corpus of texts for yourself training language models.

Keywords: clustering, data, decision making, efficiency, neural net.

Introduction

Cluster analysis is one of the most important sections of system data analysis and is used in various problem areas - technical, natural science, social.

Clustering is an example of an unsupervised learning problem and comes down to dividing the original set of objects into subsets of classes in such a way that elements of one class are as similar as possible to each other, and elements of different classes are different.

Traditional cluster analysis methods work with objects specified as vectors signs [1–4]. When working with texts, the first step of the algorithm is clustering is definition space signs and construction in it vectors available texts [5, 6]. Typically received vectors have big dimensions and when working with them traditional cluster analysis methods do not provide sufficient efficiency[7–10].

When working with short texts dimension vectors does not decrease, but only is added property sparsity to feature vectors that creates additional difficulties with them processing by cluster analysis methods [11, 12]. Below the short texts in this research implied texts consisting from one or several sentences with a total number of words ranging from 5 to 100.

In addition, additional complicating factors solution tasks clustering for short texts are: synonymy, homonymy, more frequent, compared to ordinary texts, use abbreviations, slang expressions and neologisms and most the main thing is partial or complete absence context for short texts.

Swift height arrays information consisting from short text sets fragments, contributes intensification research in the field development methods processing texts using machine learning.

Problem annually dedicated to a significant number of studies. Big Part carried out research refers to texts in English language.

In the article an example of a system is considered in which is happening generation big number of short texts. In it the participants form documentation strategic planning, within the framework of which are determined key indicators efficiency.

Formulations key indicators efficiency form a data set consisting from short texts.

Within these systems acute the task is to form and update classifier based on this set.

This task can be solved with clustering.

Main Part

Modern methods clustering using neural networks are usually used neural network for preparation vectors signs and then used analytical method (based on formulas with hyperparameters) for clustering these signs.

As a result, the result of clustering due to quality received vectors signs, i.e. quality training neural network.

At the same time, in the last time appear methods allowing solve the clustering problem directly using neural network that allows combine process receiving vectors signs and actually clustering.

Sticking to ideas for the researcher most simple and accurate reverse communication there will be criticism received results clustering, in this work supposed reverse connection two types:

1 – "element X_i must belong to the cluster C_j";

2 – "element X_i should not be in cluster C_j".

Simultaneously may be received arbitrarily the number of such restrictions, in particular, is easily specified restriction "change" elements X_i and X_j in places" combination two restrictions first kind. In Table 1 shows an example of a formalized reverse connections from an expert in the form of a matrix reverse communications.

At the intersection of the line corresponding element (object) from the data set and cluster to which the an object was ranked with the highest degree confidence neural network (maximum value in the corresponding output component vector) is placed reverse connection two the above types in the form "Include" or "Exclude" respectively. in Table 1.