

ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І.І.МЕЧНИКОВА

Факультет математики, фізики та інформаційних технологій

Кафедра математичного забезпечення комп'ютерних систем

Дипломна работа

на здобуття освітньо-кваліфікаційного рівня «бакалавр»

на тему тему Обробка потоків неструктурованих даних в інформаційно пошукових системах

Unstructured data flow processing data in information-retrieval systems

Виконав: студент денної форми навчання

напряму підготовки 123 – Комп'ютерна інженерія

(шифр і назва напрямку підготовки, спеціальності)

Жужа Георгій Юрійович

(прізвище, ім'я, по-батькові)

Керівник _____ ст. викладач Максимов О.С.

(науковий ступінь, вчене звання, прізвище та ініціали, підпис)

Рецензент _____ доц. Пенко В.Г.

(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:

Протокол засідання кафедри

№ Від « » 2020 р.

Завідувач кафедри

Захищено на засіданні ЕК №

протокол № від « » 2020 р.

Оцінка _____ / _____ / _____
(за національною шкалою, шкалою ECTS, бали)

Голова ЕК

Є.В. Малахов

(підпис)

(прізвище, ініціали)

Н.Ф. Казакова

(підпис)

(прізвище, ініціали)

АНОТАЦІЯ

Метою дипломної роботи є модернізація ядра системи «ScanStream» з додаванням нових можливостей аналізу неструктурованого тексту методами штучного інтелекту.

Основне призначення АС «ScanStream», розробленої на платформі «СвітОсвіти» є систематизація різномірних неструктурованих інформаційних потоків для попередньої їх обробки, групування і подальше їх зберігання в частково-структурованого вигляді.

Модернізованого ядро використовуючи способи регуляризації наївного Байєса зі згладжуванням в загальному вигляді Лідстоун дозволяє застосовувати методи штучного інтелекту для більш ефективного аналізу неструктурованих текстів. Результатом досліджень є розроблений програмний комплекс обробки неструктурованих даних з використанням штучного інтелекту. Модернізації ядра системи «ScanStream» дозволила підвищити коефіцієнт достовірності визначення категорії тексту до 95%.

ABSTRACT

The purpose of the thesis is to modernize the core of the system "ScanStream" with new features for the analysis of unstructured text by artificial intelligence.

The main purpose of the automated system "ScanStream", developed on the platform "СвітОсвіти" allows you to systematize a variety of unstructured information flows for pre-processing, grouping and subsequent storage in a partially structured form.

The upgraded kernel using naive Bayes regularization methods with anti-aliasing in general Lidstone allows the use of artificial intelligence methods for more efficient analysis of unstructured texts. The result of the research is a developed software package for processing unstructured data using artificial intelligence. Upgrading the kernel of the "ScanStream" system allowed to increase the reliability of determining the category of text to 95%.

АННОТАЦИЯ

Целью дипломной работы является модернизация ядра системы «ScanStream» с добавлением новых возможностей анализа неструктурированного текста методами искусственного интеллекта.

Основное назначение АС «ScanStream», разработанной на платформе «СвітОсвіти» есть систематизация разнородных неструктурированных информационных потоков для предварительной их обработки, группировки и дальнейшее их хранения в частично-структурированном виде.

Модернизированное ядро, использующие средства регуляризации наивного Байеса со сглаживанием в общем виде Лидстоуна, позволяет применять методы искусственного интеллекта для более эффективного анализа неструктурированных текстов. Результатом проведенных исследований является разработанный программный комплекс обработки неструктурированных данных с использованием искусственного интеллекта. Модернизация ядра системы «ScanStream» позволила повысить коэффициент достоверности определения категории текста до 95%.

ЗМІСТ

ВСТУП	6
1 АКТУАЛЬНІ ПРОБЛЕМИ ІНФОРМАЦІЙНИХ ПОТОКІВ ПІДПРИЄМСТВА В СУЧАСНИХ УМОВАХ ЦИФРОВІЗАЦІЇ	8
1.1 Проблеми обробки різнорідних інформаційних потоків управлінських даних	8
1.2 Структурована і неструктурована інформація	12
1.3 Дослідження проблем неструктурованих даних підприємства	15
1.4 Обробка, зберігання та доступ до неструктурованої інформації	19
2 МЕТОДИ ОБРОБКИ НЕСТРУКТУРОВАНОЇ ІНФОРМАЦІЇ	22
2.1 Методи обробки неструктурованої інформації	22
2.2.1 Математичний інструмент обробки неструктурованої інформації	29
2.2.2 Поліноміальний метод Байєса	33
2.3 Аналіз існуючих методів зберігання і доступу до неструктурованої інформації	34
2.3.1 СУБД: реляційні бази даних	34
2.3.2 Об'єктне зберігання даних	36
2.3.3 Нативне XML сховище	39
2.4 XML сховище даних	40
2.4.1 Модель даних для пошуку в XML	41
3 ОБРОБКА ПОТОКІВ НЕСТРУКТУРОВАНИХ ДАНИХ В ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМАХ НА ПРИКЛАДІ АС «SCANSTREAM»	46
3.1 Опис функціональних можливостей АС «ScanStream»	46
3.2 Існуючі методи і технології обробки неструктурованих потоків даних в «ScanStream»	49
3.3 Модернізація ядра системи із застосуванням штучного інтелекту.	52
3.4 Приклад роботи «ScanStream»	59
ВИСНОВОК	61
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	62
ДОДАТОК А Код програмного комплексу	63

ВСТУП

На сучасному етапі розвитку інформаційних технологій та інформаційно-пошукових систем, важливе значення має застосування різних видів інформації в потоці управлінських даних.

Різнорідна інформація – це файли документів, електронні повідомлення, аудіофайли, цифрові зображення, відеокліпи та інше.

Актуальність обраної теми обумовлена важливістю управлінською інформацією її все зростаючою значимістю для прийняття управлінських рішень. Недостовірні інформація призводить до непередбачуваних наслідків.

В ході дослідження визначено, що за підрахунками вчених близько 80% інформації в компаніях є неструктурованою. Працівники фірм використовують тільки 20% всіх інформаційних ресурсів компанії, а це впливає на продуктивність працівників, всієї компанії і на якість прийняття рішення.

Щоб позбутися від «інформаційного хаосу» як неструктурованою інформації, потрібно вміти нею грамотно розпорядитися.

Метою дипломної роботи є модернізація ядра системи «ScanStream» з додаванням нових можливостей аналізу неструктурованого тексту методами штучного інтелекту.

Відповідно до поставленої мети необхідно вирішити такі завдання:

- 1) розглянути методи перетворення та зберігання неструктурованої інформації;
- 2) спроектувати і розробити підсистему обробки неструктурованих даних із застосуванням штучного інтелекту;
- 3) здійснити вбудовування розробленої підсистеми в існуючу інтегровану інформаційну систему «ScanStream»;

4) виконати тестування модернізованої системи на прикладі обробки мінімум 20 текстів і підготовленого за допомогою експертних оцінок сховища.

Об'єктом дослідження є АС обробки неструктурованих потоків даних «ScanStream», її структура, вбудовані засоби інтеграції, а також інформаційні потоки, якими маніпулює система.

Предметом дослідження є «інформаційний хаос», представлений у вигляді неструктурованих різних даних.

Практична значимість дослідження полягає в модернізації ядра системи «ScanStream» для розширення можливостей аналізу неструктурованих потоків даних методами штучного інтелекту.

Робота складається з трьох розділів, висновків, списку використаної джерел, додатків.

У першому розділі описуються проблеми інформаційних потоків, проблема неструктурованих даних, теоретичні основи неструктурованих даних, і проблема зберігання неструктурованих даних.

У другому розділі описуються методики, що застосовуються для обробки неструктурованої інформації, застосування математичного методу теореми Байєса, опис способів зберігання неструктурованих даних і теоретичні основи зберігання даних в нативному XML сховищі.

У третьому розділі описується практичне застосування обробки неструктурованих даних, опис програми та опис зберігання отриманих результатів і неструктурованих даних.

ВИСНОВОК

У результаті виконання дипломної роботи модернізовано ядро системи «ScanStream». За допомогою методів взаємодії з ядром впроваджено новий метод обробки неструктурованих потоків даних.

Підготовлено вхідні тексти для тестування в розмірі 20 текстів із заздалегідь відомими категорій.

При проведенні тестування виявлені наступні результати: при обробці 20 неструктурованих текстів із заздалегідь відомими категоріями в режимі до модернізації, коефіцієнт достовірності визначення категорії тексту становив 80%. При застосуванні модернізованих методів коефіцієнт достовірності визначення категорії тексту склав 95%.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Актуальные проблемы архивного хранения электронных документов. [Электронный ресурс] / Режим доступа: works.doklad.ru/view/bZQB4wx3SoA.html
2. XML as Data Format. [Электронный ресурс] / режим доступа: docstore.mik.ua/orelly/xml/xmlnut/ch15_01.htm
3. Data mining. [Электронный ресурс] / Режим доступа: britannica.com/technology/data-mining/Pattern-mining
4. Comparison some of ontologoy editors. [Электронный ресурс] / Режим доступа: ef.uns.ac.rs/mis/archive-pdf/2013%20-%20No2/MIS2013-2-.pdf
5. Unstructured Data Management with Oracle Database. [Электронный ресурс] / Режим доступа: oracle.com/technetwork/database/information-management/unstructured-data-management-wp-12c-1896121.pdf
6. Why Object Storage? [Электронный ресурс] / Режим доступа: blog.westerndigital.com/why-object-storage/
7. Object storage. [Электронный ресурс] / Режим доступа: ibm.com/cloud/learn/object-storage
8. Object Storage use cases. [Электронный ресурс] / Режим доступа: cloudacademy.com/blog/object-storage-block-storage/
9. What is NAS? [Электронный ресурс] / Режим доступа: seagate.com/gb/en/tech-insights/what-is-nas-master-ti/
10. SAX. [Электронный ресурс] / Режим доступа: accu.org/index.php/journals/515
11. XML [Электронный ресурс] / Режим доступа: xul.fr/en-xml.php.
12. XML Schema [Электронный ресурс] / Режим доступа: w3.org/XML/Schema
13. XPath [Электронный ресурс] / Режим доступа: w3.org/TR/xpath