ВИЛУЧЕННЯ ІНФОРМАЦІЇ З ОСВІТНІХ ДЖЕРЕЛ НА ОСНОВІ NATURAL LANGUAGE PROCESSING

Худік Б.О.[1], аспірант, bohdankhudik@gmail.com;
Золотухіна О.А.[1], к.т.н., доцент, zolotukhina.oks.a@gmail.com
[1] Державний університет інформаційно-комунікаційних технологій, Київ, Україна

В процесі навчання студенти використовують різноманітні джерела як для забезпечення самостійної роботи в межах вивчення дисциплін навчального плану, так і для самоосвіти. Бібліотечні фонди закладів освіти зазвичай містять доволі велику кількість різноманітної літератури. Це може ускладнити процес пошуку та підбору необхідних джерел, зважаючи на те, що студенти не завжди мають час та відповідний досвід у визначенні релевантних до конкретних навчальних потреб джерел. В якості інструменту, що дозволяє співставити деякі освітні джерела до потреб студентів чи якихось навчальних завдань можуть виступати методи та засоби Natural Language Processing [1].

Метою дослідження є аналіз застосування методів Natural Language Processing для вилучення інформації з освітніх джерел.

Постановка задачі:

1) аналіз особливостей вхідних текстів, що використовуються в начальному процесі;
2) визначення етапів вилучення інформації з освітніх джерел на основі методів Natural Language Processing.

Вхідні тексти, які використовуються в освітньому процесі, можна поділити на декілька груп:

– методичні матеріали – містять чітко структуровану інформацію та стосуються певних видів роботи в процесі вивчення дисципліни, як правило, містять невеликі або середні обсяги тексту, можуть містити певну кількість рисунків, схем, формул чи інших елементів, які не відносяться до сфери обробки Natural Language Processing;

– посібники, підручники – орієнтовані на вивчення однієї дисципліни чи блоку дотичних дисциплін і мають значно більші обсяги у порівнянні з методичними матеріалами, характеризуються великими обсягами текстових пояснень у порівнянні з графічними чи іншими видами даних;

– книги – схожі за характером контенту з підручниками та посібниками, але в загальному випадку не спрямовані на вивчення конкретної дисципліни, а містять дотичну інформацію за дисципліною чи окремою частиною/темою дисципліни;

– наукові джерела (статті, тези доповідей конференцій) – визначаються невеликим обсягом та вузькою спрямованістю, орієнтовані на конкретну задачу, а не на область досліджень навчальної дисципліни в цілому.

Методи обробки текстів природною мовою дозволяють підготувати вхідні «сирі» текстові дані вирішення більш складних завдань обробки текстової інформації. Вони включають в себе як тривіальні операції, типу переведення всіх літер в тексті в нижній або верхній регістри, видалення знаків пунктуації та спеціальних символів, видалення символів пробілів (зазвичай, для розріджених фрагментів), токенізація, видалення цифр (чисел) або заміна на текстовий еквівалент тощо, так і операцій, які потребують застосування більш складних алгоритмів обробки, що можуть залежати від характеру даних (виправлення помилок в тексті, нормалізація слів, збагачення даних тощо) [2].

Визначений розподіл освітніх джерел та різноманіття характеристик їх контенту, обсягів та внутрішньої структури не дозволяє застосовувати єдиний підхід до вилучення ключової для подальшої обробки інформації. Загальний алгоритм вилучення корисної інформації може бути описаний наступним чином:

1. Визначення типу джерела.
2. Визначення набору методів попередньої обробки тексту в залежності від джерела.
3. Визначення переліку елементів тексту, цінних для подальшого аналізу, тобто релевантних до певного користувацького запиту.
4. Токенізація тексту з урахуванням задачі вилучення (в якості токенів для різних задач можуть виступати слова, словосполучення, речення, абзаци тощо).
5. Векторизація тексту з урахуванням його розміру та типу джерела.
6. Вилучення токенів, релевантних запиту.

Вилучення інформації з освітніх джерел різних типів передбачає використання різних методів Natural Language Processing, що обумовлено особливостями внутрішньої структури текстових джерел, властивостями контенту, розмірами текстів та іншими характеристиками. Диференційований підхід в залежності від типу джерела дозволить більш точно визначити відповідність джерела потребам користувача та вилучити максимально цінну для нього інформацію. Вилучена інформація може бути використана для побудови рекомендації того чи іншого джерела для використання в процесі самостійної роботи студента під час вивчення навчальної дисципліни, підготовки курсової роботи чи дипломного проєкту. Формування характеристик джерел на основі їх контенту дозволить визначити близькість джерел до запиту користувача, в тому числі з урахуванням їх семантичної складової, що, в свою чергу, може гарантувати підвищення якості рекомендації джерела, як потенційного об’єкту інтересу користувача.

Література

1. Balush, I., Vysotska, V., & Albota, S. (2021). Recommendation System Development Based on Intelligent Search, NLP and Machine Learning Methods. In MoMLeT+ DS (pp. 584-617).

2. Eisenstein, J. (2019). Introduction to natural language processing. MIT press.

Анотація

Розглянуто підхід до формування характеристик освітніх джерел та вилучення з них цінної для зацікавлених сторін інформації на основі методів Natural Language Processing. Визначено етапи алгоритму вилучення текстової інформації з урахуванням специфіки контенту.

Ключові слова: рекомендаційна система, Natural Language Processing, вилучення текстової інформації.

Abstract

An approach to the formation of characteristics of educational sources and the extraction of valuable information for stakeholders based on Natural Language Processing methods is considered. Defined stages of the textual information extraction algorithm taking into account the specifics of the content.

Keywords: recommendation system, Natural Language Processing, extraction of text information.