

ЛАТЕНТНО-СЕМАНТИЧНОГО АНАЛІЗУ ТЕКСТІВ

Лісовик Д. П.

Науковий керівник – доц., к. т. н. Арсенюк І. Р.

Латентно-семантичний аналіз (ЛСА) – це метод обробки інформації природною мовою, що дозволяє проаналізувати взаємозв'язок між колекцією документів і термінами, які в них зустрічаються. Існує 3 види задач, що вирішуються ЛСА: порівняння двох термінів між собою, порівняння двох документів між собою, порівняння терміну і документа. При використанні методу вважається, що текст є набором слів, а семантичне значення тексту визначається його словами. Для спрощення вважається, що кожне слово має одне значення.

Латентно-семантичний аналіз передбачає виконання таких кроків:

- виключення з тексту стоп-символів (розділових знаків, сполучників, слів, що не мають смислового навантаження);
- виконання стеммінгу, тобто виділення основи слова (розглядається алгоритм Портера);
- виключення слів, що зустрічаються один раз (не обов'язковий пункт);
- складення частотної матриці (зліва – слова, зверху – тексти);
- сингулярне розкладання матриці та спрощення;
- побудова графіка, на основі якого робляться висновки.

Основна мета роботи полягає у створенні програми для автоматизованої генерації тестів на основі вихідного тексту. Автоматизованість обумовлена тим, що програма може допускати граматичні помилки (рід, число, відмінок, час), помилки смислового характеру, а також генерувати надто прості або надто складні тести. А отже втручання людини обов'язкове.

Для реалізації обрано мову програмування C#, адже в ній є технології для роботи з текстом і базами даних.

Створюється база даних з термінами відповідної предметної області, які розділені за категоріями. На основі ЛСА здійснюється порівняння тексту з базою даних. На основі бази даних знаходяться терміни до яких ставляться запитання, а неправильні варіанти відповідей вибираються з відповідної категорії, для того, щоб не було очевидним, які саме варіанти неправильні.

Інтерфейс програми передбачатиме можливість редагування питання, варіантів відповідей, вилучення питання і варіантів відповідей, збереження результатів.

До основних переваг програми можна віднести:

- можливість застосування викладачами в навчальному процесі;
- можливість редагування результатів роботи;
- самовдосконалення програми (користувач може додавати терміни в базу даних термінів і програма враховує їх при подальшій роботі).