Мирош Юрій Михайлович, бакалавр

Національний університет «Львівська політехніка», Львів

0000-0001-5829-9322

**Інтелектуальна система освоєння іноземної мови з використанням статистичного аналізу**

**Вступ.** Область контент аналізу містить багато досліджень, про одне з найбільш цитованих було здійснене в області політології, , Laver, Benoit, and Garry (2003)[1], воно порівнює ефективність наявних у той час, традиційних методів з запропонованим ними методом частоти слів. З однієї сторони – метод ручного відбору слів, який вимагає велику кількість зусиль і значну затрату часу. З іншої ж сторони – автоматизований, машинний відбір контенту, який є цілком надійним і відтворюваним, але варто зважати, що складні алгоритми розпізнавання фраз можуть коштувати дорого й мають необхідність у регулярному коригуванні. Варто також зазначити, що алгоритми фраз можуть бути не такими доступними для мов відмінних від англійської.

**Мета роботи.** Метою створення даної системи можна вважати зменшення витрат часу на прочитання матеріалу й вивчення іноземних. Оскільки користувач заздалегідь знайомий з словами й фразами, що йому будуть траплятися найчастіше то він може скоротити час на прочитання на близько 20-30% у залежності від обсягу тексту.

**Основна частина роботи.** Опис аналогу. У цій частині роботі буде розглядатися альтернативна система wordfreq її робота і застосування. Загальний принцип роботи полягає у тому, що wordfreq завантажує веб-сторінку або локальний файл і готує розподіл частот різних слова [1].

Синтаксис використання виглядає наступним чином: wordfreq using filename [, min length(integer) nonumbers nogrammar nowww nocommon clear append], де filename – це ім’я файлу, що буде оброблятися, min length(integer) визначає мінімальну кількість символів, необхідну в слові, щоб зберегти його в статистичному розподілі, за замовченням значення 0, nonumbers вказує на видалення слів, що містять числа, nogrammar визначає відкидання слів, які є частиною загальної граматики (наприклад, is або are), nowww вказує на видалення слів, пов'язаних із http або html, clear очищає дані в пам'яті та append вказує на додавання нового розподілу частоти слів до існуючого розподілу частоти слів. За замовчуванням вони зберігаються.. Один з варіантів використання може бути адреса у мережі Інтернет, що обов’язково має починатися з http або https або це може бути локальний файл з довільним розширенням, для усіх файлів буде оброблене ASCII джерело файлу.

Робота програми wordfreq починається з обробки веб-сторінки або локального файлу. На першому етапі створюється статистичний розподіл усіх унікальних слів, що містяться у тексті. Вміст файлу обробляється як один рядок, у якому всі символи, що не входять до кодової таблиці ASCII (A–Z, a–z, 0–9, а також деякі неанглійські літери), замінюються пробілами. Потім очищений текст розбивається на слова за пробілами як розділювачами.

Для обробки веб-сайтів програма стикається з додатковими труднощами, такими як наявність реклами чи службових даних у JavaScript-тегах. Через це у списку слів можуть з’являтися "недослова" — довгі рядки, що представляють назви змінних, функцій або класів. Щоб уникнути таких проблем, користувач отримує чотири окремі списки, які дозволяють виключити небажані слова, наприклад, ті, що містять цифри, посилання (http, html) або інші технічні терміни. Це забезпечує більш точний аналіз тексту та зручність у роботі.

**Опис роботи власної системи.** Розроблена інтелектуальна система освоєння іноземної мови з використанням статистичного аналізу має вигляд веб-сайту, що значно спрощує використання її у порівнянні з аналогом. Також відмінністю можна вважати чіткий напрямок системи у застосуванні, який, однак можна й віднести до звуження теоретичних можливостей. Щодо файлів з якими працює система, то це також обмеження, що запобігає появі «недослів», адже на вхід йдуть лише такі формати pdf, docx, та txt. До параметрів системи також варто віднести PERCENT\_OF\_FREQ, що базується на статті Harald Baayen (1993) [2] про статистичні моделі розподілу слів, їх лігвістичну оцінку та бере пропоновані дані для частот слів. COUNT\_OF\_WORDS, ще один параметр, що вказує скільки саме слів має відібрати система для здійснення навчання користувача. WORD\_LENGTH – останній, але не менш важливий параметр, значення якого базуються на відомостях зі статті Reginald Smith(2012) [3] про роль різних довжин слів, розподіл та ентропія символів. За допомогою цих даних система здійснює вибір щодо кількості й якось бажаних слів та здійснює генерування тестувань з застосуванням побічних систем, що надають синоніми, антоніми, рими й просто випадкові слова, що допомагає створити максимально варіативні результати. Для статистичного аналізу вибираються слова лише з наперед заданих алфавітів, щоб уникнути потрапляння зайвих лексем. Оскільки речення з найуживанішими слова становлять велику частину тексту то створення тестів на їх основі значно просуне людину, що буде їх виконувати в загальному розумінні книги й безсумнівно добавить знань щодо книги.

**Висновки.** У ході даної роботи було порівняно існуючу систему зі власною як результат було виокремлено переваги й недоліки кожної, потенційні області застосування кожної з них. Що до конкретно нової системи, то можна зазначити, що дослідні цифри підтверджують бажані параметри ефективності й можна вважати, що й на практиці її робота буде цілком справна.

**Література**

1. Лейвер, М., Бенуа, К., Гаррі, Дж. Витяг політичних позицій із політичних текстів за допомогою слів як даних [Електронний ресурс] // The American Political Science Review. 2003. Т. 97, № 2. С. 311–331. Режим доступу: https://www.jstor.org/stable/3118211?origin=JSTOR-pdf (дата звернення: 08.05.2025).

2. Баайен, Г. Статистичні моделі для розподілу частот слів: лінгвістична оцінка [Електронний ресурс] // The Stata Journal. 1992. Т. 26, № 5/6. С. 347–363. Режим доступу: https://www.jstor.org/stable/30204630 (дата звернення: 08.05.2025).

3. Сміт, Р. Частоти довжин окремих слів: розподіли та ентропії символів [Електронний ресурс] // Glottometrics. 2012. С. 7–22. Режим доступу: https://arxiv.org/ftp/arxiv/papers/1207/1207.2334.pdf (дата звернення: 08.05.2025).