# Лабораторна робота № 3

# ЗАСТОСУВАННЯ ЗАКОНІВ ЗІПФА ДЛЯ ОБРОБЛЕННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ

# **Ціль работи**

Отримати практичний досвід використання програм для автоматичного пошуку ключових слів в тексті.

# **Завдання**

1. Проаналізувати фрагмент навчального посібника Основы программной инженерии [Текст]: учеб. пособие /И. Б. Туркин, Е. В. Соколова. – Х.: Нац. аэрокосм. ун-т им. Н. Е. Жуковского «Харьк. авиац. ин-т», 2013. – 232 с. (папка std.), обраному відповідно до варіанту.
2. Вибрати 10-15 слів, що на ваш погляд, найчастіше зустрічаються в тексті. Записати їх.
3. Провести частотний аналіз фрагмента тексту з використанням інструменту <http://simple-seo-tools.com>.
4. Порівняти список ключових слів, складений на основі частотного аналізу тексту, зі списком, отриманим експертним шляхом в п.2.
5. Побудувати графік залежності ***частоти входження слова*** від ***рангу*** (закон Зіпфа), використовуючи табличний процесор EXCEL.
6. Зробити висновок про доцільність використання програм для автоматичного пошуку ключових слів тексту.

# **Зміст звіту**

1. Постановка завдання.
2. Номер розділу фрагмента навчального посібника.
3. Список ключових слів, отриманих експертним шляхом.
4. Список ключових слів, отриманих на основі частотного аналізу.
5. Графік залежності частоти входження слова від рангу.
6. Висновки.

# **Варіанти.**

|  |  |
| --- | --- |
| № варіанту | Розділ підручника |
| 1 | 1.1 |
| 2 | 1.2 |
| 3 | 1.3.1 |
| 4 | 1.3.2 |
| 5 | 1.4 |
| 6 | 1.4.1 |
| 7 | 1.4.2 |
| 8 | 1.4.3 |
| 9 | 1.4.4 |
| 10 | 2.1 |
| 11 | 2.2 |
| 12 | 3.1 |
| 13 | 3.1.1 |
| 14 | 3.1.2 |
| 15 | 3.2 |
| 16 | 3.2.1 |
| 17 | 3.2.2 |
| 18 | 3.2.3 |
| 19 | 3.2.4 |
| 20 | 3.3 |
| 21 | 3.3.1 |
| 22 | 3.3.2 |
| 23 | 3.3.3 |
| 24 | 3.3.4 |
| 25 | 3.3.5 |
| 26 | 3.4.1 |
| 27 | 3.4.2 |
| 28 | 3.4.3 |

**Теоретичні відомості**

**Ключовими словами** називають повнозначні слова, стійкі складні і складноскорочені слова та термінологічні словосполучення, що несуть істотне смислове навантаження в текстах документів відповідних галузей знання.

**Закони Зіпфа. Перший закон Зіпфа**

Всі створені людиною тексти побудовані за єдиними правилами. Нікому не вдається обійти їх. Яка б мова не використовувалася, хто б не писав — класик чи графоман, — внутрішня структура тексту залишиться незмінною. Вона описується законами Зіпфа (G.K. Zipf).

Виберемо в тексті будь-яке слово і порахуємо, скільки разів воно зустрічається в тексті. Ця величина називається частотою входження слова. Виміряємо частоту входження кожного слова тексту. Об'єднаймо слова з однаковою частотою входження в одну групу. Розташуємо частоти в міру їх зменшення та пронумеруємо. Порядковий номер частоти назвемо рангом частоти. Так, слова, що найбільш часто зустрічаються, матимуть ранг 1, наступні за ними - 2 і т. Д. Визначимо ймовірність входження конкретного слова в текст. Ймовірність буде дорівнює відношенню частоти входження цього слова до загальної кількості всіх слів в тексті.

*Ймовірність = Частота входження слова / кількість слів у тексті*

Джорж Зіпф (GK Zipf) в 1949 році виявив цікаву закономірність. Виявляється, якщо помножити ймовірність входження слова в текст на ранг частоти, то отримана величина (С) приблизно постійна:

*С = (Частота входження слова х Ранг частоти) / Кількість слів в тексті = const*

Перетворивши цю формулу до вигляду:

*Частота входження слова = (Кількість слів в тексті х С) / Ранг частоти,*

і враховуючи, що не тільки C, але і кількість слів у конкретному тексті - величини постійні, легко помітити, що функція Частота входження слова (Ранг частоти) є функція виду y = k / x (де k = const, x - аргумент функції y ) і її графік - гіпербола, представлена ​​на рис 1.

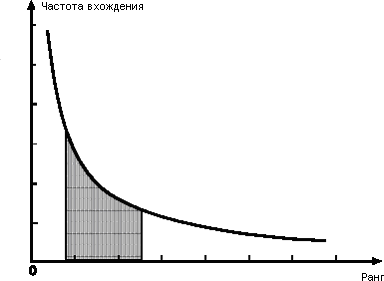


Рис 1 - Залежність Частоти входження слова від Рангу частоти

Отже, за першим законом Зіпфа, якщо найпоширеніше слово зустрічається в тексті, наприклад, 100 раз (відповідає значенням y = 100/1), то наступне за частотою слово навряд чи зустрінеться 99 раз. Частота входження другого за популярністю слова, з високою часткою ймовірності, виявиться близько 50 (y = 100/2), третього - близько 33 (y = 100/3), четвертого - близько 25 (y = 100/4) і т. д. Значення константи в різних мовах відрізняється, але всередині однієї мовної групи залишається незмінним, який би текст ми не взяли. Так наприклад, для англійських текстів константа Зіпфа дорівнює приблизно 0,1, для української мови - 0,06-0,07.

**Другий закон Зіпфа**

Розглядаючи перший закон, не бралося до уваги, що різні слова входять в текст з однаковою частотою. Зіпф встановив, що частота і кількість слів, що входять в текст з цією частотою, теж пов'язані між собою. Якщо побудувати графік, відклавши по осі абсцис частоту входження слова, а по осі ординат - кількість слів у групі з даної частотою, то отримана крива буде зберігати свої параметри в межах однієї мови для всіх створених людиною текстів. Однак і міжмовні відмінності невеликі. На якій би мові текст не був написаний, форма кривої Зіпфа залишиться незмінною. Можуть трохи відрізнятися лише коефіцієнти, що визначають параметри кривої. Сказане ілюструється кривими, представленими на рис. 2, на якому показані криві для англійської (нижня), французької та російської мов (верхня).

Рис 2 - Залежність Кількості слів в тексті від Частоти входження слова

