Лабораторна робота 1. Реалізація моделей подання документів

Мета

Реалізувати теоретико-множинну (стандартну булеву) та алгебраїчну (векторно-просторову) модель подання документів, таким чином ознайомившись з найбільш поширеними моделями та набувши практичні навички з їх реалізації.

Вимоги до програмного забезпечення

1. Для реалізації програмного забезпечення в ході лабораторної роботи може використовуватись будь-який стек (мова програмування, фреймворк і так далі) технологій.
2. Програмне забезпечення може мати будь-який з перелічених інтерфейсів користувача: консольний, веб, мобільний, настільний.
3. Програмне забезпечення повинно мати два режими роботи: перший – з використанням теоретико-множинної моделі, другий – з використанням алгебраїчної моделі. Це може бути реалізовано будь-яким способом на вибір студента, наприклад параметром запуску для консольного застосунку, перемикачем у графічному інтерфейсі тощо. Вимоги для кожного із режимів див. далі.

Режим з використанням стандартної булевої моделі подання

Робота з розробленим програмним забезпеченням у даному режимі повинна відбуватися наступним чином:

Етап 1. Введення множини індексних термів

1. Користувач після запуску програмного забезпечення у теоретико-множинному режимі повинен відразу ввести множину індексних термів.
2. Користувач повинен мати можливість ввести будь-яку невід’ємну кількість індексних термів на свій вибір (хоча б один індексний терм є обов’язковим).
3. Допускається (на вибір студента) реалізація введення множини індексних термів за допомогою її зчитування з файлу будь-якого формату. У такому випадку, на даному етапі користувач повинен вводити шлях до цього файлу.

4. Після закінчення введення множини індексних термів, користувач повинен автоматично перейти на наступний етап.
5. Повернення до цього етапу не допускається.

Етап 2. Введення колекції документів

1. Користувач повинен мати можливість ввести будь-яку невід'ємну кількість документів (хоча б один документ у колекції є обов'язковим).
2. Допускається (на вибір студента) реалізація введення колекції документів за допомогою її зчитування з окремих текстових файлів, розміщених у певній папці (один текстовий файл = один документ). У такому випадку, на даному етапі користувач вводить шлях до папки з файлами.
3. Після закінчення введення, користувач автоматично переходить на наступний етап.
4. Допускається (на вибір студента) реалізація можливості переходу користувача між другим та третім етапом і назад.

Етап 3. Виконання пошукових запитів

1. На даному етапі користувач вводить пошуковий запит і переглядає його результати.
2. На вибір студента, пошуковий запит потрібно вводити у нормальній кон'юнктивній або диз'юнктивній формі.
3. Після виконання пошукового запиту, користувач повинен залишатися на цьому ж етапі, з можливості ввести новий пошуковий запит.

Режим з використанням векторно-просторової моделі

Даний режим роботи розробленого програмного забезпечення передбачає два етапи – введення колекції документів (або зчитування з папки) та виконання необмеженої кількості пошукових запитів, аналогічно до попереднього режиму.

- В якості міри подібності документу до пошукового запиту використати косинусну міру.
- Граничне значення подібності, за якої документ повинен відображатися у результатах запиту, обирається студентом самостійно.

- При показі результатів запиту потрібно обов’язково вивести обчислене значення подібності для кожного документа.
- Для реалізації векторно-просторової моделі необхідно **обов’язково** використати міру TF-IDF **відповідно** до свого варіанту (див. далі).

Таблиця варіантів для режиму з векторно-просторовою моделлю

Варіант визначається за порядковим номером студента у списку його групи. Якщо у групі більше 25-ти студентів, то для студентів з порядковим номером більше 25 потрібно взяти залишок від ділення номеру на 25.

№ варіанту	Формула *tf*	Формула *idf*
1	$f_{t,d}$	1
2	$f_{t,d}$	$\log\frac{N}{n_t}$
3	$f_{t,d}$	$\log\left(\frac{N}{1+n_t}\right)+1$
4	$f_{t,d}$	$\log\left(\frac{\max_{\{t'\in d\}} n_{t'}}{1+n_t}\right)$
5	$f_{t,d}$	$\log\frac{N-n_t}{n_t}$
6	$f_{t,d}\Big/\sum_{t'\in d} f_{t',d}$	1
7	$f_{t,d}\Big/\sum_{t'\in d} f_{t',d}$	$\log\frac{N}{n_t}$
8	$f_{t,d}\Big/\sum_{t'\in d} f_{t',d}$	$\log\left(\frac{N}{1+n_t}\right)+1$
9	$f_{t,d}\Big/\sum_{t'\in d} f_{t',d}$	$\log\left(\frac{\max_{\{t'\in d\}} n_{t'}}{1+n_t}\right)$

10	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$	$\log \frac{N - n_t}{n_t}$
11	$\log(1 + f_{t,d})$	1
12	$\log(1 + f_{t,d})$	$\log \frac{N}{n_t}$
13	$\log(1 + f_{t,d})$	$\log\left(\frac{N}{1 + n_t}\right) + 1$
14	$\log(1 + f_{t,d})$	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1 + n_t}\right)$
15	$\log(1 + f_{t,d})$	$\log \frac{N - n_t}{n_t}$
16	$0.5 + 0.5 \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$	1
17	$0.5 + 0.5 \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$	$\log \frac{N}{n_t}$
18	$0.5 + 0.5 \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$	$\log\left(\frac{N}{1 + n_t}\right) + 1$
19	$0.5 + 0.5 \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1 + n_t}\right)$
20	$0.5 + 0.5 \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$	$\log \frac{N - n_t}{n_t}$
21	$K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$	1
22	$K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$	$\log \frac{N}{n_t}$
23	$K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$	$\log\left(\frac{N}{1 + n_t}\right) + 1$

24	$K + (1-K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
25	$K + (1-K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$	$\log\frac{N-n_t}{n_t}$

Для варіантів 21-25, значення параметру K обрати довільно так, щоб $K \neq 0.5$.

Вимоги до звіту з лабораторної роботи

Звіт з лабораторної роботи повинен містити:

1. Титульну сторінку
2. Постановку завдання – варіант для реалізації векторно-просторової моделі
3. Код розробленого програмного забезпечення
4. Знімки екрану з прикладами результатів роботи розробленої програми в обох режимах (вхідні дані обираються на розсуд студента)
5. Висновки