Семилітко Микола, група 3

Огляд на розроблену систему інформаційного пошуку

Розроблена система отримала наступні оцінки:

Оцінка не ранжованих результатів			
	Релевантні	Не релевантні	
Знайдені	6	0	
Не знайдені	2	6	
Точність(P)	1		
Повнота(R)	0,75		
β=	1		
F-міра	0,857142857		
Правильність	0,857142857		
Оцінка ранжованих результатів			
	Так	Ні	Усього
Так	5	2	7
Ні	8	6	14
Усього	13	8	21
P(A)	0,523809524		
P (нерелевантний)	0,523809524		
P (релевантний)	0,476190476		
P(E)	0,501133787		
Каппа-статистика	0,045454545		

В розробленій системі дозволяється використовувати пошук з джокерами та з відстанню Левенштейна.

Пошук відбувається досить довго, оскільки пошук виконується в 1 потоці, також на швидкість впливає сама мова Python. Ця мова була обрана, тому що розробка відбувається трошки швидше ніж в мові C++

(наскільки мені відомо, саме цю мову використовують для побудови інвертованих індексів та інших алгоритмів в Google та Yandex). Дані для пошуку зберігались в словнику(dict) для швидкого пошуку за ключем(шуканого словом). Алгоритмом для створення індексу великої колекції був BSBI, з використанням memory map для часткового читання даних з файлів. Алгоритмами стискакання були стрічка та байтове кодування. Документи ранжуються за схемою зважування tf-$idf_{t,d}$. Кластеризація відбувається за принципом скорочення словника. Пошук в xml документах не використовує теги для ранжування. Також була протестована бібліотека Okapi BM25. Щодо неї, вона ініціалізується набагато довше, ніж індекс написаний мною, результати ранжованого пошуку моїх алгоритмів та алгоритмів бібліотеки зходяться, найбільш релевантні файли однакові.

До проблемних алгоритмів хотів би віднести джокері, в лекції дуже гарно описані всі випадки окрім “*word”, що створило деякі проблеми в розробці.

Відповідь моєї системи з ранжованим пошуком:

```
D:\projects\SearchEngine\Lab-8>python Positional_Indexes.py
creating index...
for request ' ['to', 'be'] '
was analyzed requests 1
8 samples/The-Letters-of-a-Por-Marianna-Alcofo-[ebooksread.com].txt score : 842.7309100547261
7 samples/Sir-Edwin-Landseer-Frederick-G--St-[ebooksread.com].txt score : 841.4238785288403
```

Відповідь бібліотеки BM25 (імена файлів за зменшенням їх релевантності)

```
D:\projects\SearchEngine\Lab-12>python Positional_Indexes.py
for request ' to be '
0 0.35665000580705924
1 0.5390127659181464
2 0.0
3 0.0
4 0.0
5 0.0
6 0.0
7 1.3386077566671277
8 0.0
9 3.2397125098742454
10 3.2509964620602387
11 0.0
was analyzed requests 1
samples/The-Letters-of-a-Por-Marianna-Alcofo-[ebooksread.com].txt
samples/Sir-Edwin-Landseer-Frederick-G--St-[ebooksread.com].txt
samples/Navarro_Mir-matematiki_31_Taynaya-zhizn-chisel_RuLit_Me.txt
samples/Arbones_Mir-matematiki_12_Tom-12-Chisla-osnova-garmonii-Muzyka-i-matematika_RuLit_Me.txt
samples/Alsina_Mir-matematiki_11_Tom-11-Karty-metro-i-neyronnye-seti-Teoriya-grafov_RuLit_Me.txt
samples/Kasalderrey_Mir-matematiki_16_Obman-chuvstv_RuLit_Me.txt
samples/Levshin_Karlikaniya_2_Puteshestvie-po-Karlikanii-i-Al-Dzhebre_RuLit_Me.txt
samples/Levshin_V-labirinte-chisel_RuLit_Net.txt
samples/Loyd_Samyie_znamenityie_golovolomki_mira_RuLit_Net.txt
samples/matematicheskie_chudesa_i_tajjny.u.txt
samples/Smallian_Priklyucheniya_Alisyi_v_Strane_Golovolomok_RuLit_Net.txt
samples/test.fb2
```

Додаток. Файл з розрахунками оцінки системи ІС.