

Текст

Лекција 5

Стеван Гостојић

Факултет техничких наука, Нови Сад

31. март 2022.

Преглед садржаја

- 1 Увод
- 2 Текст
- 3 Претрага и индексирање
- 4 Apache Lucene Core
- 5 Закључак

Преглед садржаја

- 1 Увод
- 2 Текст
- 3 Претрага и индексирање
- 4 Apache Lucene Core
- 5 Закључак

- Текст је, у општем случају, низ знакова који преносе неку поруку

Текст

- Велика количина информација складишти се и преноси у облику текста (тј. текстуалних докумената као јединичних носилаца информација)
- Технике и алати за складиштење и претраживање текста разликују се од техника и алата за складиштење других врста информација
- Познавање ових техника и алата доприноси писању ефикаснијих и ефектнијих рачунарских програма који обрађују текст

Преглед садржаја

- 1 Увод
- 2 Текст
- 3 Претрага и индексирање
- 4 Apache Lucene Core
- 5 Закључак

Репрезентација текста

- Посоји више метода за рачунарску репрезентацију текста
- Неке од тих метода су представљање текста као низа знакова, као вектора и као фреквенције n -торки

Низ знакова

- Текст можемо да представимо као низ знакова (који су кодирани одређеним кодом)
- Временом су настали различити кодови за кодирање знакова

Вектор

- Други начин за репрезентацију текста је вектор речи (тј. токена)
- У том случају су димензије вектора речи које се појављују у тексту, а интензитет појединачних компоненти вектора број појављивања одговарајуће речи у документу

Фреквенција n-торки

- Текст можемо да представимо и као фреквенцију појављивања различитих n-торки речи у тексту
- Можемо да посматрамо уређене парове, уређене тројке, уређене четворке итд.

Кодирање знакова

- С обзиром да рачунари податке складиште, преносе и обрађују као низ битова, знаке неког језика је потребно кодирати у неком коду
- Код је пресликавање скупа знакова на скуп бинарних речи
- Постоје многи стандарди за кодирање знакова од којих је данас у најширој употреби Unicode

ASCII

- American Standard Code for Information Interchange (ASCII)
је је стандард за кодирање текста
- Садржи 128 знакова (од којих су неки контролни знаци)
- Основа је модерних стандарда за кодирање текста

ASCII

| Знак | Код (децимално) | Код (бинарно) |
|------|-----------------|---------------|
| A | 65 | 100 0001 |
| B | 66 | 100 0010 |
| C | 67 | 100 0011 |
| a | 97 | 110 0001 |
| b | 98 | 110 0010 |
| c | 99 | 110 0011 |

Table 1: ASCII

ISO/IEC 8859

- ISO/IEC 8859 је фамилија стандарда за кодирање знакова која је заједнички прописана од ISO и IEC
- То је проширење ASCII стандарда
- ISO/IEC 8859 је подељен у неколико делова који стандардизују кодирање знакова различитих језика (нпр. ISO/IEC 8859-2 садржи српске латиничке знакове, а ISO/IEC 8859-5 српске ћириличне знакове)

Unicode

- Unicode је стандард за конзистентно кодирање, репрезентацију и руковање текстом
- Такође је проширење ASCII стандарда
- Садржи више 144,697 знакова који покривају 159 језика (верзија 14.0 од септембра 2021)
- Знакови се могу кодирати на више начина (UTF-8, UTF-16, UTF-32 итд.)
- Прописује га Јуникод конзорцијум (енг. Unicode Consortium)

Преглед садржаја

- 1 Увод
- 2 Текст
- 3 Претрага и индексирање**
- 4 Apache Lucene Core
- 5 Закључак

Проналажење

- Проналажење (енг. retrieval) или претрага (енг. search) текста је процес одабира докумената који задовољавају одређени упит
- На пример, претраживање интернета коришћењем Google претраживача

Прегледање

- Прегледање (енг. browsing) текста је процес одабира докумената на основу веза са другим документима
- На пример, праћење хиперлика од једног HTML документа до другог HTML документа

Индексирање текста

- Индексирање текста је процес прављења индекса у циљу убрзавања претраге текста
- Појам индекса потиче из библиотечке делатности
- У зависности од техника које се користе за (рачунарску) претрагу текста, индекси се имплементирају као различите структуре података

Претраживачи

- Претраживачи (енг. search engine) су рачунарски програми чија је основна функција индексирање и претрага текстуалних докумената
- Претраживачи могу да буду понуђени као сервиси или као библиотеке које се уграђују у рачунарске програме

Преглед садржаја

- 1 Увод
- 2 Текст
- 3 Претрага и индексирање
- 4 Apache Lucene Core**
- 5 Закључак

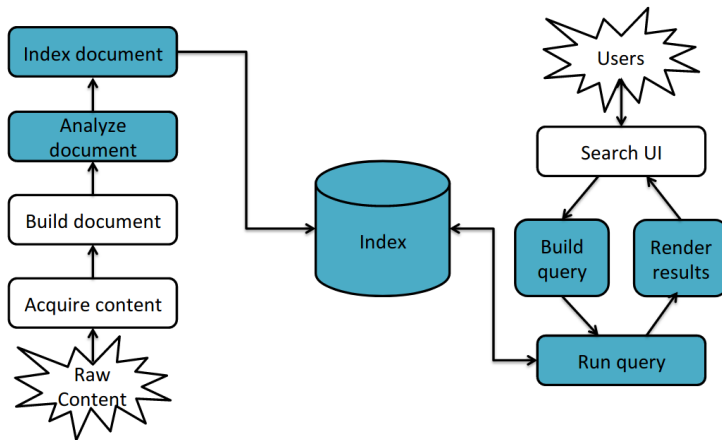
Apache Lucene Core

- Apache Lucene Core је Java API за (индексирање) и претрагу текста
- Постоје конектори и за друге програмске језике и платформе (нпр. Python)
- Apache Lucene Core имплементира напредне методе претраге текста, има високе перформансе и захтева релативно мало ресурса

Apache Lucene Core

- Основна јединица индексирања и претраге су (текстуални) документи
- Документи се састоје од поља (различити типови докумената састоје се од различитих поља)
- Сирови текстуални документи се преводе у једно или више поља (која могу а не морају да се индексирају и анализирају)

Apache Lucene Core



Lucene анализатори

- WhitespaceAnalyzer (дели текст на токене по белим знацима)
- SimpleAnalyzer (дели текст на токене по знацима који нису слова, па их претвара у мала слова)
- StopAnalyzer (исто као и SimpleAnalyzer, али и уклања тзв. "стоп речи")
- StandardAnalyzer (најчешће коришћен анализатор који, за разлику од StopAnalyzer анализатора, нуди и неке "напредне" функције као што је препознавање URL адреса, адреса електронске поште итд.)
- итд.

WhitespaceAnalyzer

```
1 ' 'The quick brown fox jumped over the lazy dog' '  
2 [The] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dog]  
3
```

WhitespaceAnalyzer

```
1 ' 'XY&Z Corporation - xyz@example.com' '  
2 [XY&Z] [Corporation] [-] [xyz@example.com] ' '  
3
```

SimpleAnalyzer

```
1 ' 'The quick brown fox jumped over the lazy dog' '  
2 [the] [quick] [brown] [fox] [jumped] [over] [the] [lazy] [dog]  
3
```

SimpleAnalyzer

```
1 ' 'XY&Z Corporation — xyz@example.com' '  
2 [xy] [z] [corporation] [xyz] [example] [com]  
3
```

StopAnalyzer

```
1  ' 'The quick brown fox jumped over the lazy dog' '  
2  [quick] [brown] [fox] [jumped] [over] [lazy] [dog]  
3
```

StopAnalyzer

```
1 ' 'XY&Z Corporation — xyz@example.com ' '  
2 [xy] [z] [corporation] [xyz] [example] [com]  
3
```

StandardAnalyzer

```
1 ' 'The quick brown fox jumped over the lazy dog' '  
2 [quick] [brown] [fox] [jumped] [over] [lazy] [dog]  
3
```


StandardAnalyzer

```
1 ' 'XY&Z Corporation — xyz@example.com ' '  
2 [xy&z] [corporation] [xyz@example.com]  
3
```

Apache Lucene Core

- Колекције текстуалних докумената могу да се претражују на два начина
- Први је Apache Lucene Core API
- Други је Apache Lucene Core Query Parser упитног језика

Query Parser упитни језик

- Синтакса (и семантика) Query Parser упитног језика није стандардизована и може да (незнатно) варира од верзије до верзије Lucene-а
- У примерима је приказан упитни језик за Apache Lucene Core 9.1

Термини

- Термини су појединачни термини или фразе
- Појединачни термини су речи
- Фразе су низ речи под двоструким наводницима

Термини

```
1 Hello
2 "Hello World"
3
```

Поља

- Поља су уређени парови кључ:вредност
- Вредности поља су термини
- Једно поље може да буде подразумевано (тада термини не морају да се квалификују кључем)

Поља

```
1 title:"Hello World"  
2 title:Hello World  
3 text:go  
4 go  
5
```

Упити

- Упити се састоје од поља и логичких оператора
- Логички оператори се користе да би се више поља комбиновало у (сложени) упит

Lucene упити

```
1 title:"Hello World" AND text:go
2 title:"Hello World" AND go
3 title:Hello World
4
```

Модификатори

- Модификатори су речи који се користе за модификовање термина (и упита)

Модификатори

- Претраге "џокер" знацима (енг. wildcard searches)
- Претраге регуларним изразима (енг. regular expression searches)
- "Фази" претраге (енг. fuzzy searches)
- Претраге по близини (енг. proximity searches)
- Претраге по интервалима (енг. range searches)
- "Појачавање" термина (енг. boosting a term)
- Логички оператори (енг. boolean operators)
- Груписање (енг. grouping)

Претраге "џокер" знацима

- Lucene подржава претрагу "џокер" знацима "?" и "*" над појединачним терминима
- "џокер" знак "?" замењује било који знак
- "џокер" знак "*" замењује било који низ знакова
- "џокер" знаци "?" и "*" не могу да се налазе на почетку термина

Претраге "џокер" знацима

```
1 te?t
2 test*
3
```

Претраге регуларним изразима

- Lucene подржава претрагу регуларним изразима
- И синтакса (и семантика) регуларних израза може да (незнатно) варира од верзије до верзије Lucene-а
- Регуларни изрази наводе се између знакова "/"

Претраге регуларним изразима

1 `/[mb] oat /`

2

Регуларни изрази

- Регуларни изрази су низ знакова који специфицирају текстуални образац

Регуларни изрази

| Израз | Опис |
|-------|----------------------|
| \d | цифра |
| \D | све осим цифре |
| \s | бели знак |
| \S | све осим белог знака |
| \w | слово |
| \W | све осим слова |

Table 2: Регуларни изрази

Регуларни изрази

| Израз | Опис |
|----------|------------------------|
| . | било који знак |
| "текст" | текст |
| [abcde] | класа знакова |
| [^abcde] | негација класе знакова |
| | унија |

Table 3: Регуларни изрази

Регуларни изрази

| Израз | Опис |
|------------|----------------------------|
| $?$ | нула или једно појављивање |
| $*$ | нула или више појављивања |
| $+$ | једно или више појављивања |
| $\{n\}$ | n појављивања |
| $\{n, \}$ | n или више појављивања |
| $\{n, m\}$ | n до m појављивања |

Table 4: Регуларни изрази

Претраге регуларним изразима

```

1 [a-z0-9_-]{3,16}
2 #?([a-f0-9]{6}|[a-f0-9]{3})
3 ([a-z0-9_\.-]+)@([\da-z_\.-]+\.[a-z_\.-]{2,6})
4 (https?:\/\/\/?)([\da-z_\.-]+\.[a-z_\.-]{2,6})([\/\w_\.-]*)*\/?
5

```

"Фази" претраге

- Lucene подржава "фази" претраге (тј. претраге по сличним речима)
- Сличност је дефинисана мером Levenshtein Distance (минималним бројем знакова које је потребно променити да би се једна речи трансформисала у другу реч)

"Фази" претраге

- 1 roam~
- 2 roam~0.8
- 3

Претраге по близини

- Lucene подржава претрагу докумената у којима се одређене речи налазе на одређеном растојању

Претраге по близини

```
1 "jakarta apache"~10
```

```
2
```


Претраге по интервалима

- Lucene подржава претрагу докумената чија поља имају вредности у одређеном интервалу
- Интервали се специфицирају навођењем доње и горње границе и могу да буду отворени ("{" и "}") и затворени ("[" и "]")

Претраге по интервалима

```
1 mod_date:[20020101 TO 20030101]
2 title:{Aida TO Carmen}
3
```

"Појачавање" термина

- Lucene подржава појачавање релевантности једног термина у односи на друге термине (релевантност термина утиче на рангирање резултата претраге)
- Што је фактор појачавања већи, термин је јелевантнији (фактор појачавања мора да буде позитиван)

"Појачавање" термина

```
1 jakarta apache
2 jakarta^4 apache
3 "jakarta apache"^4 "Apache Lucene"
4
```

Логички оператори

- Lucene подржава комбиновање термина коришћењем логичких оператора
- То су оператори AND, "+", OR, NOT и "-"
- Логички оператор OR се подразумева

Логички оператори

```
1 "jakarta apache" jakarta
2 "jakarta apache" OR jakarta
3 +jakarta lucene
4 "jakarta apache" AND "Apache Lucene"
5 "jakarta apache" NOT "Apache Lucene"
6 "jakarta apache" -"Apache Lucene"
7
```

Груписање

- Lucene подржава груписање израза коришћењем заграда

Груписање

1 (jakarta OR apache) AND website

2

Преглед садржаја

- 1 Увод
- 2 Текст
- 3 Претрага и индексирање
- 4 Apache Lucene Core
- 5 **Закључак**

Закључак

- Методе репрезентације текста
- Низ знакова
- Вектор
- Фреквенција n -торки
- Кодови
- Unicode

Закључак

- Претрага
- Прегледање
- Индексирање
- Претраживачи

Закључак

- Apache Lucene Core
- Анализатори
- Query Parser упитни језик
- Термини
- Поља
- Упити

Закључак

- Претраге "џокер" знацима
- Претраге регуларним изразима
- "Фази" претраге
- Претраге по близини
- Претраге по интервалима
- "Појачавање" термина
- Логички оператори
- Груписање

Литература

- Apache Lucene Core, <https://lucene.apache.org/>

Хвала на пажњи!