Text Mining and Social Media Mining

THS. LÊ NHẬT TÙNG

Regular expressions

THS. LÊ NHẬT TÙNG

Biểu thức chính quy (Regular Expression)

Định nghĩa:

Ngôn ngữ cho phép khớp chuỗi ký tự theo mẫu

Sử dụng các toán tử đặc biệt (metacharacters)

Hỗ trợ tìm kiếm linh hoạt trong văn bản

Biểu thức chính quy (Regular Expression)

Phạm vi sử dụng: Tài liệu văn bản, Mã nguồn, File log, Bảng tính

Loại ký tự hỗ trợ:

- ASCII cơ bản:
  - Chữ cái (a-z, A-Z), Số (0-9), Dấu câu, Ký hiệu đặc biệt (%#$@!)
- Unicode:
  - Văn bản quốc tế, Ký hiệu đặc biệt, Hỗ trợ đa ngôn ngữ

Biểu thức chính quy (Regular Expression)

Chữ cái (Letters)

- Quy tắc cơ bản: một ký tự là biểu thức chính quy đơn giản nhất
- Cách hoạt động: khi tìm A, a, abc, xyz - nó sẽ tìm chính xác các ký tự đó

Văn bản: abcdefg abcde abc

Nếu tìm "abc" -> sẽ tìm được 3 kết quả chứa "abc"

Biểu thức chính quy (Regular Expression)

Chữ số (Digits)

- Quy tắc cơ bản: số cũng được xử lý như ký tự
- Cách hoạt động: khi tìm 1, 123, 9, 0 - nó sẽ tìm chính xác các số đó

Văn bản: abc123xyz 123dollars 123

Nếu tìm "123" -> sẽ tìm được 3 kết quả chứa "123"

Biểu thức chính quy (Regular Expression)

\d (digit - chữ số)

- Khớp với bất kỳ chữ số nào
- Có thể khớp với bất kỳ số nào từ 0 đến 9

. (dot - dấu chấm)

- Khớp với bất kỳ ký tự nào
- Có thể khớp với bất kỳ ký tự đơn lẻ nào (chữ cái, số, khoảng trắng)
- Nhiều dấu chấm có nghĩa là nhiều ký tự

\. (ký tự dấu chấm)

- Để khớp chính xác với dấu chấm, sử dụng dấu gạch chéo ngược

Ví dụ:

•Văn bản: cat. 896. ?=+.

•Mẫu khớp (tìm) tất cả các từ từ văn bản trên: ...\.

Biểu thức chính quy (Regular Expression)

[] (dấu ngoặc vuông)

- •Có thể khớp với bất kỳ ký tự đơn nào trong ngoặc
- •Bao gồm (Inclusion): [abc] - chỉ a, b, hoặc c
- •Loại trừ (Exclusion): [^abc] - không phải a, b, hoặc c

Ví dụ 1:

- •Văn bản: can man fan
- •Mẫu khớp (tìm) tất cả các từ từ văn bản trên: [cmf]an *(mẫu này sẽ tìm các từ có chữ c, m, hoặc f ở đầu, theo sau bởi "an")*

Ví dụ 2:

- •Văn bản: dan ran pan
- •Mẫu loại trừ (bỏ qua) tất cả các từ từ văn bản trên: [^drp]an *(mẫu này sẽ tìm các từ KHÔNG có chữ d, r, hoặc p ở đầu)*

Ví dụ 3:

- •Văn bản: hog dog bog
- •Mẫu khớp chỉ với động vật sống (hog, dog, nhưng không phải bog): [^b]og *(mẫu này sẽ tìm các từ kết thúc bằng "og" nhưng không bắt đầu bằng "b")*

Biểu thức chính quy (Regular Expression)

Ký tự đặc biệt trong Biểu thức chính quy

^ (dấu mũ)

- Bắt đầu với
- Khớp với tất cả các từ bắt đầu bằng một mẫu cho trước

$ (dấu đô la)

- Kết thúc với
- Khớp với tất cả các từ kết thúc bằng một mẫu cho trước

^...$ (khớp chính xác)

- Bắt đầu và kết thúc với
- Ví dụ: với biểu thức ^measure$, chỉ từ "measure" sẽ khớp

Biểu thức chính quy (Regular Expression)

^cat - sẽ khớp:
cat, cats, catch
không khớp: scat, location

ing$ - sẽ khớp:
running, walking, singing
không khớp: rings, bring

^cat$ - chỉ khớp:
chính xác từ "cat"
không khớp: cats, catch, scat

Biểu thức chính quy (Regular Expression)

*** (Kleene Star - Dấu sao Kleene)**

•Khớp với 0 hoặc nhiều lần lặp lại của ký tự đứng trước nó

•Ví dụ:

- .* - khớp với 0 hoặc nhiều ký tự bất kỳ
- ^m(.*)r$ - khớp: mr, monitor, mentor, mirror, v.v.

+ (Kleene Plus - Dấu cộng Kleene)

•Khớp với 1 hoặc nhiều lần lặp lại của ký tự đứng trước nó

•Ví dụ:

- a+ - khớp với một hoặc nhiều chữ 'a': a, aa, aaa
- ^me+ - khớp: me, mee, meee
- [abc]+ - khớp với một hoặc nhiều ký tự a, b, hoặc c

Biểu thức chính quy (Regular Expression)

So sánh * và +:

1.Dùng *:

1. Có thể không có ký tự nào
2. "ca*t" khớp: ct, cat, caat, caaat

2.Dùng +:

1. Phải có ít nhất 1 ký tự
2. "ca+t" khớp: cat, caat, caaat
3. KHÔNG khớp: ct

Biểu thức chính quy (Regular Expression)

? (Dấu hỏi)

•Không xuất hiện hoặc lặp lại ký tự đứng trước một lần (tính chất tùy chọn)

Ví dụ: Chúng ta có văn bản 4 dòng:

1 file found?
2 files found?
24 files found?
No files found.

Giải thích cách hoạt động:

•\d = khớp với chữ số
•\d+ = khớp với một hoặc nhiều chữ số
•files? = từ "file" có thể có hoặc không có chữ 's'
•found? = từ "found" kết thúc bằng dấu hỏi

Yêu cầu: Khớp chỉ những dòng có một hoặc nhiều file được tìm thấy

Giải pháp:

•Sử dụng ký tự đặc biệt \d để khớp với số lượng file
•Sử dụng biểu thức \d+ files? found? để khớp tất cả các dòng có file được tìm thấy

\s (Ký tự khoảng trắng)

•Khớp với bất kỳ khoảng trắng nào

•Cực kỳ hữu ích khi xử lý văn bản thô

Ví dụ: Chúng ta có văn bản 4 dòng:

1.abc

2. abc

3. abc

4. abc

Yêu cầu: Khớp chỉ những dòng có khoảng trắng giữa số thứ tự và 'abc'

Giải pháp:

•Sử dụng biểu thức \d.\s+abc để khớp:

- \d = một chữ số
- . = dấu chấm (phải có dấu gạch chéo ngược)
- \s+ = một hoặc nhiều khoảng trắng
- abc = văn bản cần

| (Ký tự pipe - dấu gạch đứng)

• Các tập ký tự khác nhau có thể có (lựa chọn thay thế)

Ví dụ 1:

• Buy more (milk|bread|juice)

• Chỉ khớp với: Buy more milk, Buy more bread, Buy more juice

Ví dụ 2:

• ^measure$|^sulfur$

• Chỉ khớp với một trong hai từ: measure hoặc sulfur

Ví dụ 3: Chúng ta có văn bản 4 dòng:
I love cats
I love dogs
I love logs
I love cogs
Yêu cầu: Khớp chỉ những dòng có động vật (cats và dogs)
Giải pháp:
•Sử dụng biểu thức: I love (cats|dogs)
•Kết quả:

- Khớp: "I love cats", "I love dogs"
- Không khớp: "I love logs", "I love cogs"

Biểu thức chính quy (Regular Expression)

Các ký tự đặc biệt nâng cao khác

- [a-z] : Các ký tự từ a đến z
- [0-9] : Các số từ 0 đến 9
- \w : Bất kỳ ký tự chữ và số
- \W : Bất kỳ ký tự không phải chữ và số
- {m} : Lặp lại m lần
- {m,n} : Lặp lại từ m đến n lần
- (...) : Nhóm khớp
- (a(bc)) : Nhóm khớp con
- (.*) : Khớp tất cả

Ví dụ:

1.[a-z]{2,4}

    1. Khớp: ab, abc, abcd
    2. Không khớp: a, abcde

2.\w+@\w+.\w{2,3}

    1. Khớp: user@gmail.com
    2. Không khớp: @gmail.com

3.(abc){2}

    1. Khớp: abcabc
    2. Không khớp: abc

4.\W+

    1. Khớp: @#$%
    2. Không khớp: abc123