Regular Expression

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

실무 팁: RegEx 적용 방법

1. Text Editor 활용

- 파일을 열어 적용 과정을 보면서 정제
- 바로 결과를 확인할 수 있음
- 적용 과정이 log로 남지 않음
  - 재활용 불가

2. 전용 모듈 작성 및 활용

- Python등을 활용하여 모듈을 만들고 regex 리스트를 파일로 받아서 처리
- 한번에 모든 regex를 적용
  - 중간 결과 확인 불가
- regex 재활용 가능

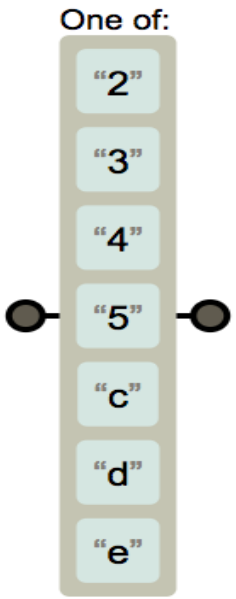
Text Editors with RegEx

- Sublime Text 3, VSCode
  - 무료

- EmEditor
  - 유료
  - 다양한 인코딩 지원
  - 대용량 corpus (GB 단위) 로딩 가능

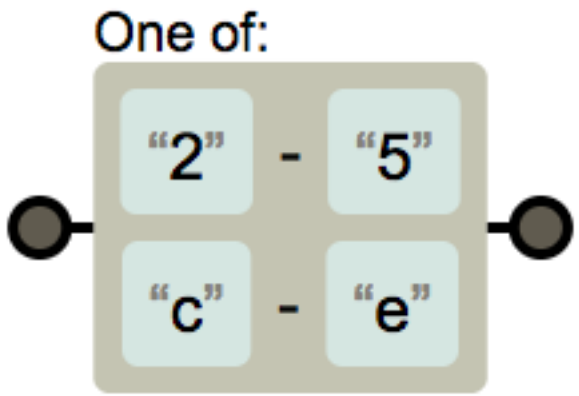
[ ]

- 2, 3, 4, 5, c, d, e 중의 character
- [2345cde]
- (2|3|4|5|c|d|e)



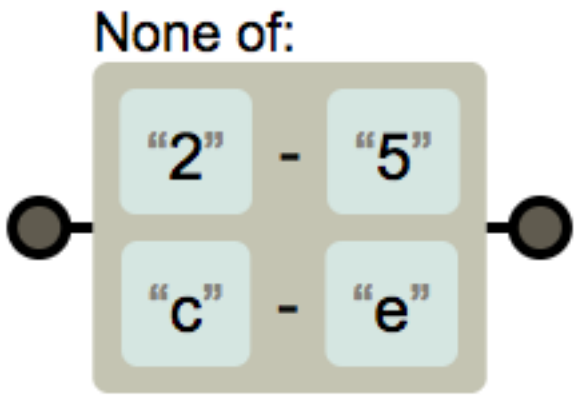
[-]

- 2, 3, 4, 5와 c, d, e 중의 character
- [2-5c-e]



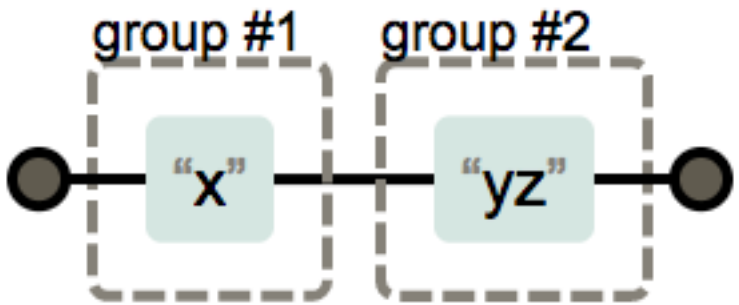
[^]

- 2, 3, 4, 5와 c, d, e를 제외한 모든 character
- [^2-5c-e]



()

- x를 ₩1에 지정, yz를 ₩2에 지정
- (x)(yz)



RegEx의 꿀기능

- 양 끝에 알파벳(소문자)으로 둘러싸인 'bc'를 제거하기
  - abcd
  - 0bc1

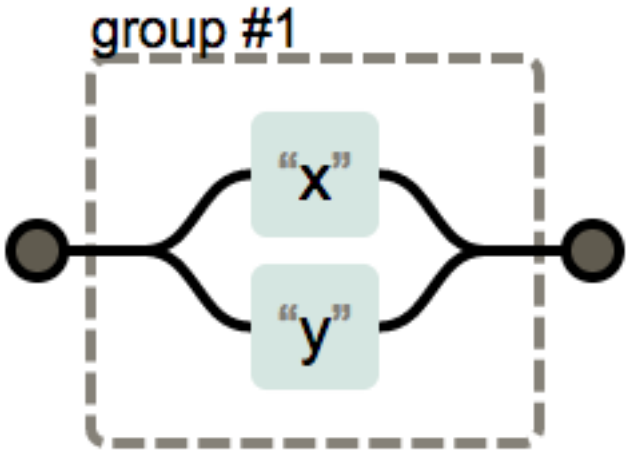
- 적용 예제

([a-z])bc([a-z]) → ₩1₩2

  - abcd → ad
  - 0bc1 → 0bc1

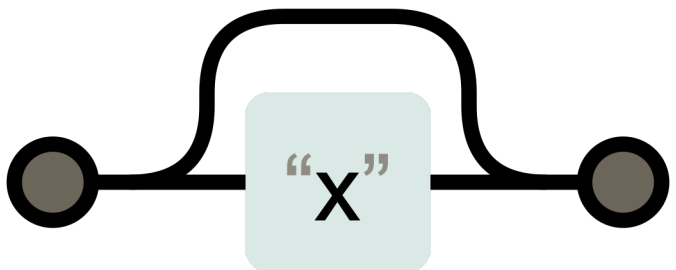
|

- x 또는 y가 나타남. 그리고 ₩1에 지정
- (x|y)



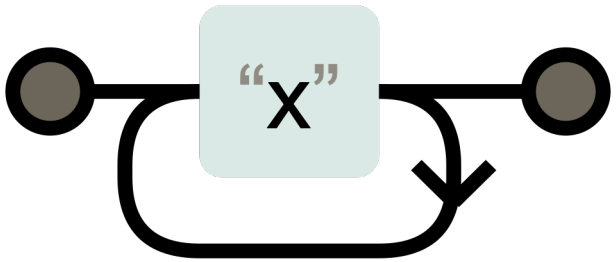
?

- x가 0번 또는 1번 나타남
- x?



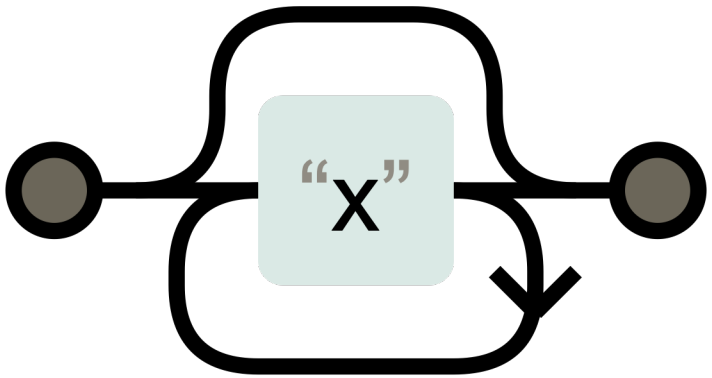
+

- x가 한 번 이상 나타남
- x+



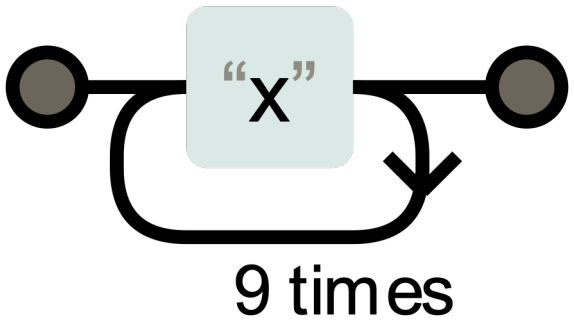
*

- x가 나타나지 않을 수도, 반복될 수도 있음
- 강력한 표현. 유의해서 사용해야 함
- x*



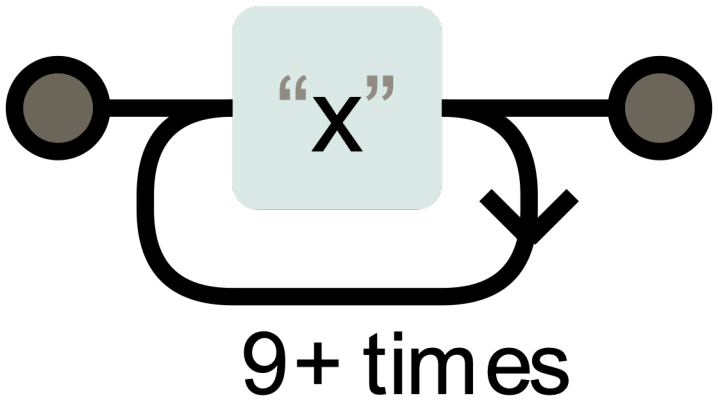
{n}, {n,}, {n,m}

- n번 반복
- x{n}



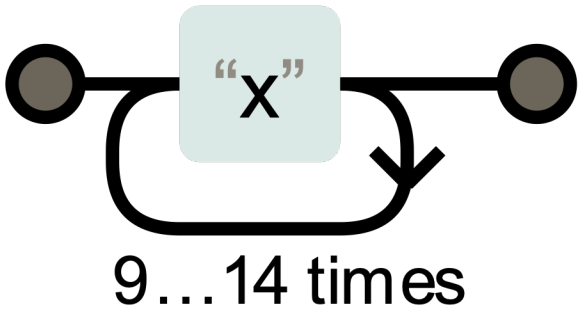
{n}, {n,}, {n,m}

- n번 이상 반복
- x{n,}



{n}, {n,}, {n,m}

- n번부터 m번까지 반복
- x{n,m}

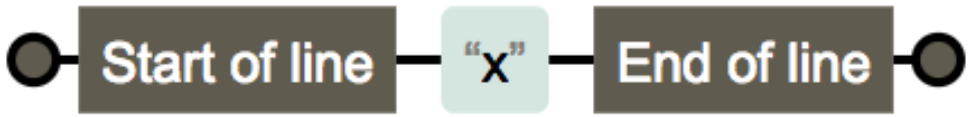


.

- any character
- 매우 강력한 표현. 유의해서 사용해야 함

^ $

- 문장의 시작과 끝을 표시
- ^x$



그 밖의 지정 문자

Meta Characters	Description
₩s	공백 문자white space
₩S	공백 문자를 제외한 모든 문자
₩w	alphanumeric(알파벳 + 숫자) + '_' ([A-Za-z0-9_]와 같음)
₩W	non-alphanumeric 문자 및 '_' 제외 ([^A-Za-z0-9_]와 같음)
₩d	숫자 ([0-9]와 같음)
₩D	숫자를 제외한 모든 문자 ([^0-9]와 같음)