Regular Expression

정규표현식

지난 시간 요약

자연어처리

자연어처리란 전처리 -> 임베딩 -> 분석의 단계로 이루어진다

전처리 단계

정규표현식

정규표현식은 파이썬에서 굉장히 강력한 find, replace

import re

전처리 단계

정규표현식

아래 문장의 어순을 바꿀려면?....

김준태 - 학점 2.0

학점 2.0 - 김준태

전처리 단계

정규표현식

index 찾거나 split 하거나 할 때보다 굉장히 짧음!

```
a = '김준태 - 학점 2.0'
p = re.compile(r'(?P<name>₩w+)₩s.₩s+(?P<grade>.+₩s.+)')
print(p.sub("₩g<grade> - ₩g<name>", a))
```

전처리 단계

정규표현식

우선은 기본 메타 문자부터.....

[] : 문자 (안에 - 쓰면 범위, ^은 not)
. : 모든 문자 (줄바꿈 제외, 공백포함)
*** : 0부터 반복**
+ : 1부터 반복
{m,n} : 횟수 지정 (m, n 중 하나만 있어도 됨)
? : 있어도 되고 없어도 되고~
| : or
^ : 시작
$: 끝

정규표현식

re 모듈?

match() : 문자열 처음부터 매치
search() : 문자열 전체 매치
findall() : 매치된거 문자열 리스트로 반환
finditer() : 매치된거 각각 match로 반환

정규표현식

실습1.

[abbbbba, cddb, aerb, aeeeb, almnj]
(모두 문자열)

일 때

1, 3번째만 True를 리턴하는 정규표현식을 작성하시오

정규표현식

match 는 그냥 Object, 그 안에는...

group() : 매치된 문자열 전부 / 일부
start() : 매치된 문자열 시작 인덱스
end() : 매치된 문자열 끝 인덱스
span() : 매치된 문자열 (시작, 끝) 튜플

정규표현식

compile 할때! 쓸 수 있는 옵션

DOTALL(S) : . 에 줄바꿈 포함
IGNORECASE(I) : 대소문자 무시
MULTILINE(M) : 여러줄 매치 (^, $ 시)
VERBOSE(X) : 주석 작성가능

정규표현식

문자 클래스

\d : 숫자
\D : not 숫자
\s : 공백 (탭, 스페이스, 줄바꿈 포함)
\S : not 공백
\w : 문자+숫자
\W : not 문자+숫자

\b : 문자와 공백 사이의 문자
\B : not 문자와 공백 사이의 문자
\A : 문자열 처음 (multiline 무시)
\Z : 문자열 끝 (multiline 무시)
\\ 을 쓸거면 r을 앞에다가 써야한다!

정규표현식

실습2.

['abcd efghopqr-stuv', 'abcd-efgh', 'abc defghijk', 'ab cdhijk']
일 때

1, 3, 4번째만 match되도록 정규표현식을 작성하시오.

정규표현식

일치하는 것을 참조하기 위한 그루핑!

() 하면 그룹이 되고
이후에 .group(n) 으로 뽑을 수 있음!

내부에서 \1 \2 로 재참조도 가능

() 안에 ?P<name> 을 넣어서 이름을 붙일 수도!

전처리 단계

정규표현식

(?) 구문..!

() 안에 ?는 백슬래시 같은 느낌

(?P<이름>) : 이름 지정하겠다

(?=특정문자) : 특정문자 바로 앞까지만 매치하겠다

(?!특정문자) : 특정문자가 아닌 경우에만 통과

전처리 단계

정규표현식

실습3.

그룹을 사용해서
['코끼리는 코끼리', '강아지는 강아지', '고양이는 고양이']

모두에게 True를 리턴하는 정규표현식을 작성하고,
주어들을 순서대로 출력하세요.

전처리 단계

정규표현식

매치되는 것들을 바꾸고 싶을 때는?

.sub('바뀐 뒤의 문자','바꿀 대상')
.subn 도 비슷하지만 얘는 바뀐 횟수까지 출력!

바뀐 뒤의 문자에는 \g 를 통해서 그룹을 넣어줄 수도 있고,
함수를 넣어줄 수도 있다!

정규표현식

실습4.

['김준태 - 학점 2.0', '이재화 - 학점 3.0', '최주원 - 학점 4.0']
에서

학점을 모두 4.5로 바꿔주세요!

전처리 단계

정규표현식

전체 다 말고 하나만 바꾸고 싶을때!

s = '(안녕)하세(요)'

에서 () 로 매치하면 오류! (그대로 출력해버린다)

?로 써서 (안녕)까지만 되도록!

? -> 최소한의 반복

정규표현식

실습5.

** 뉴스스탠드 바로가기 주제별캐스트 바로가기 타임스퀘어 바로가기 쇼핑캐스트 바로가기 로그인 바로가기 **

에서 사용된 모든 태그(<>)를 출력하세요.

정규표현식

실습6.

'1100+1200+1300+1800=5400'
이라는 문장에서 등호표시 전의 숫자만
8진법으로 변경하세요.

정규표현식

실습7.

**['https://naver.com', 'https://korea.ac.kr',
'https://www.google.com', 'google.com',
'https://facebook.com', 'naver.com']
에서**

**https로 시작하고, www.이 없고, 마지막이 .com으로 끝나는 주소
만 True를 리턴하는 정규표현식을 작성하세요.**

정규표현식

실습8.

a = '''

mike 010-1234-0907 mikeman@naver.com hi! my name is mike!

jay 010-1112-3456 jayman@naver.com hi! my name is jay!

yohan 010-2223-5874 yohanman@naver.com hi! my name is yohan!

'''

위 문장에서 순서 배치를 전화번호, 이메일, 자기소개, 이름 순서로 변경하시오.

THANK YOU

kim jun tae