2021학년도 2학기 언어와 컴퓨터

제10강 파일 처리 (1)

박수지

서울대학교 인문대학 언어학과

2021년 10월 18일 월요일

오늘의 목표

1. 정규표현식의 고급 기능으로 lookahead, lookbehind, 특수 메타 문자를 사용할 수 있다.
2. `pandas` 모듈을 사용하여 스프레드시트 파일을 읽고 처리할 수 있다.
3. `re` 모듈을 사용하여 문자열에서 한글만 추출할 수 있다.
4. `unicodedata` 모듈을 사용하여 한글 음절 문자를 자모로 분해하고 초성만 뽑아낼 수 있다.

Lookahead 형식

`pattern(?=suffix)`

- pattern 뒤에 suffix가 나오는 조건을 만족시키는 것을 매치시키기

Lookahead 예시

`[A-Za-z]+(?=ing\b)`

- 영어 접미사 '-ing'이 붙은 어간만 매치시키기

Lookbehind 형식

`(?<=prefix)pattern`

- pattern 앞에 prefix가 나오는 조건을 만족시키는 것을 매치시키기

Lookbehind 예시

`(?<=\b[Nn]on-?)[A-Za-z]+`

- 영어 접두사 'non-'이 붙은 어간만 매치시키기

특수 메타 문자

\d digit
\D non-digit
\w alphanumeric
\W non-alphanumeric
\s whitespace (space, tab)
\S non-whitespace

단어 토큰화(Word tokenization)

```
1 string = ”It’s very cold today!”
2 word_pattern = re.compile(r’\w+’)
3 word_pattern.findall(string)
```

오늘의 모듈

pandas

- 파이썬에서 엑셀 파일을 읽고 쓸 수 있다.

오늘의 실습

https://colab.research.google.com/drive/1Wd5gP-zEMfACWaQ4fH8GcjMwRlVzC6Rq