



# Gcon - 신규기능 중간보고서

작성일: 2021년 1월 19일

작성자: 오지혜

프로젝트 완성도 : **50%**/100%


## 목차

1. 데이터 추출 항목
2. 추출 방법
3. 결과
4. 개선 방향

---

### 1. 데이터 추출 항목

 파일명: **input.xlsx (예시)**

 추출 해야하는 데이터

I am Seonho.
I am Sunho,
I am Seohoo.
I need to drink something
I need to drunk something
I need to drank something
I like this person.
I like that person.
I like those person.
How did you learn of this position at Snowman Press?
How did you learn of this position at SnowPooh Press?
How did you learn of this position at SnowWhite Press?
Why don't you come to the beach with us?
Why don't you know to the beach with us?
Why don't you require to the beach with us?

- 위 그림과 같이, 문장 내에서 **다른게 표시된 단어에 태그 표시를 해서 추출**해야한다.

## 2. 추출 방법

### a. 3개씩 끊어 가져와서 리스트에 담기

`[[case_a(3문장)], [case_b] ..... ]`

### b. case 별로 각 문장을 형태소 분석을 한다.

`"list > tuple > str"` 형식으로 형태소 분석

### 😊 Good Case

- Input Data :

**I am Sunho.**

**I am Suuunwo,**

#### 1) mecab try

모든 것이 si 혹은 . , 으로 분류되어  
테스트만 사용했다.

#### 2) nltk try

영 단어의 품사별로 나눠줬다.

- for : '동사 + 전치사' 함께 태그 표시
- because : 의미도, 목적도 함께 챙겨서 태그 표시를 하는 것이 사용자가 편리할 방법이라고 생각
- example : learn of = know

I am Sunhoooo.

- Output Data :

I am <b>Sunho</b><b>.</b>  
I am <b>Suuunwo</b><b>,</b>  
</b>  
I am <b>Sunhoooo</b><b>.</b>  
</b>

### 3. 결과

No	Input	Output
1	I am Seonho.	I am <b>Seonho</b><b>.</b>
	I am Sunho,	I am <b>Sunho</b><b>,</b>
	I am Seohoo.	I am <b>Seohoo</b><b>.</b>
2	I need to drink something	I need to <b>drink</b> something
	I need to drunk something	I need to <b>drunk</b> something
	I need to drank something	I need to <b>drank</b> something
3	I like this person.	I like <b>this</b> person.
	I like that person.	I like <b>that</b> person.
	I like those person.	I like <b>those</b> person.
4	How did you learn of this position at Snow	How did you learn of this position at <b>Snowman</b> Pre
	How did you learn of this position at Snow	How did you learn of this position at <b>SnowPooh</b> Pr
	How did you learn of this position at Snow	How did you learn of this position at <b>SnowWhite</b> P

- case set 수 : 29쌍 (87줄) < - 부족하므로 데이터를 가져올 예정입니다.
- 소요시간: 1초
- 첨부파일

[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/64037aab-887d-4f0f-9f35-4c5f02a23f15/input\\_result\\_01191727.xlsx](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/64037aab-887d-4f0f-9f35-4c5f02a23f15/input_result_01191727.xlsx)

## 4. 개선방향

### a. 주의했던 작은 사항들

▲ 1) **띄어쓰기 반영** : 단순히 형태소를 나눠서 붙여준 것이 아니라, 띄어쓰기를 반영하기 위해 참고만 했다.

2) 다른 부분이 1개가 아니더라도 다르게 표시된 것이 어느것 인지 태그 표시를 할 수 있도록 했다.

### b. 개선 사항

✳ 1) 2-b-2)에 언급한 것 처럼, '**동사 + 전치사**'를 묶어서 하나의 태그로 다른 단어로 쓰여졌음을 표시하고자 한다.

→ 현재까지 해결하지 못한 이유:

1. 전치사도 하나의 단어이므로, 형태소 분석을 하면 3 문장의 단어들의 총 길이가 때문에 index 에러가 발생하고 비교대상이 원활하게 일치되지 않았다.
2. 나눠지는 품사가 분명 문장에서는 동사일지라도, 명사로 인식되어 형태소 분석이 되는 경우가 있었다. 따라서, 특정 품사로 지칭하는 데에 문제가 있었다.
3. 따라서, nltk 뿐만 아니라, Stanford도 사용할 예정이다.

2) **Mr. vs Mr.** 를 비교하면 . 와 , 가 균등하게 분리되지 않는다. 따라서, 이러한 에러도 해결해 나가고자한다.