pynori 동의어 사전

[1]:
```
from pynori.korean_analyzer import KoreanAnalyzer
nori = KoreanAnalyzer(
            decompound_mode='DISCARD', # DISCARD or MIXED or NONE
            infl_decompound_mode='DISCARD', # DISCARD or MIXED or NONE
            discard_punctuation=True,
            output_unknown_unigrams=False,
            pos_filter=False, stop_tags=['JKS', 'JKB', 'VV', 'EF'],
            synonym_filter=False, mode_synonym='NORM', # NORM or EXTENSION
        )
```

[7]:
```
nori.set_option_tokenizer(decompound_mode='DISCARD', infl_decompound_mode='DISCARD')
nori.set_option_filter(mode_synonym='EXTENSION')
print(nori.do_analysis("레드 가방")['termAtt'])
```

```
['레드', '가방']
```

[8]:
```
nori.set_option_tokenizer(decompound_mode='DISCARD', infl_decompound_mode='DISCARD')
nori.set_option_filter(mode_synonym='EXTENSION')
print(nori.do_analysis("red 가방")['termAtt'])
```

```
['red', '가방']
```

[13]:
```
nori.set_option_tokenizer(decompound_mode='DISCARD', infl_decompound_mode='DISCARD')
nori.set_option_filter(mode_synonym='EXTENSION')
print(nori.do_analysis("elasticsearch 개발자")['termAtt'])
```

```
['elasticsearch', '개발', '자', 'developer']
```

왜 개발자는 developer로 변환이 되는데 레드는 red로 변환이 되지 않을까?

https://github.com/gritmind/python-nori 를 보고 공부한 결과 내가 동의어사전을 바꾸지 않았음을 알게 되었다.

Resources

- 시스템 사전은 `~/pynori/resources/mecab-ko-dic-2.1.1-20180720` 에서 수정
  - 사전 변경사항은 다음 두 항목을 실시하면 곧바로 적용 가능
    - 기존 csv 파일 수정/삭제 or 새로운 csv 파일 추가 (주의. mecab 단어 작성 규칙)
    - 기존 `~/pynori/resources/pkl_mecab_csv/mecab_csv.pkl` 삭제
    - (참고. `mecab_csv.pkl` 파일이 없으면 KoreanAnalyzer 초기화 시에 최신 csv 파일을 기반으로 재생성)
    - (참고. `~/pynori/resources/pkl_mecab_matrix/matrix_def.pkl` 파일은 수정/삭제하지 말 것)
    - (참고. 다른 버전의 mecab-ko-dic 적용을 위해서는 코드 내의 path 수정 필요)
- 사용자 사전은 `~/pynori/resources/userdict_ko.txt` 에서 수정 (곧바로 적용 가능)
- 동의어 사전은 `~/pynori/resources/synonyms.txt.txt` 에서 수정 (곧바로 적용 가능)

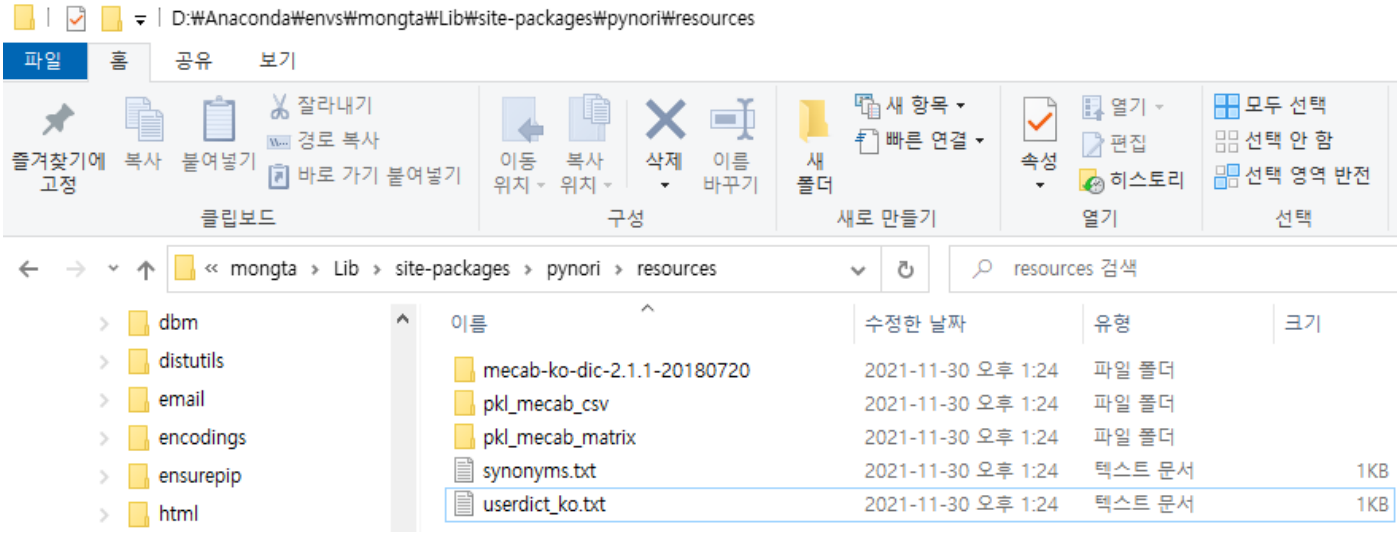
우선 pynori가 깔려진 곳을 찾아야 된다.

우선 cmd창에 "pip freeze"를 검색한다.

```
(mongta) C:₩Users₩MongTa>pip freeze
WARNING: Ignoring invalid distribution -umpy (d:₩anaconda₩envs₩mongta₩lib₩site-packages)
pynori==0.2.4
```

pynori는 버전정보만 나오는데 그러면 pynori 패키지는 노란색 글씨에 해당되는

"d:₩anaconda₩envs₩mongta₩lib₩site-packages" 에 있는 것이다.

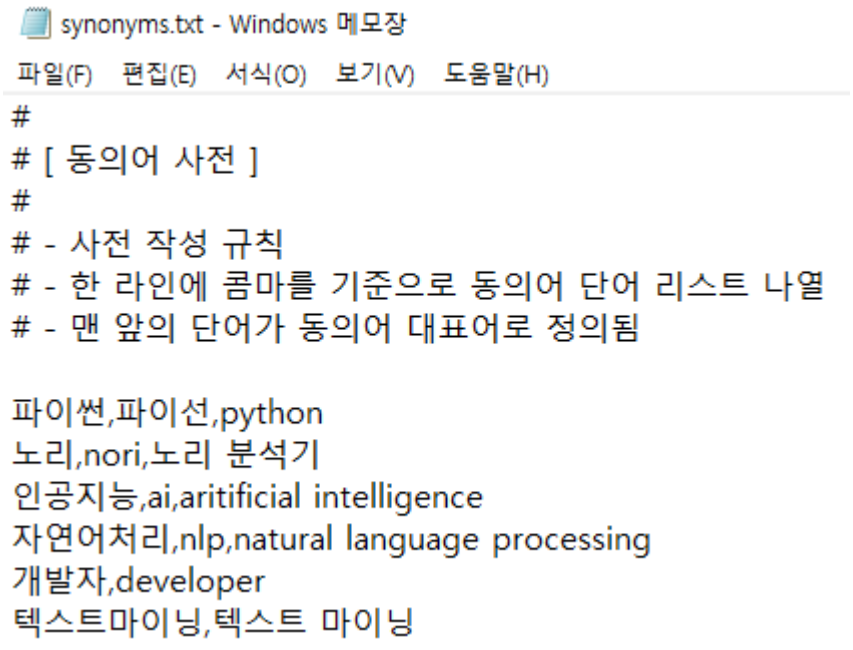


https://github.com/gritmind/python-nori 를 참고해

사용자 사전은 ~/pynori/resources/userdict_ko.txt 에서 수정 (곧바로 적용 가능)

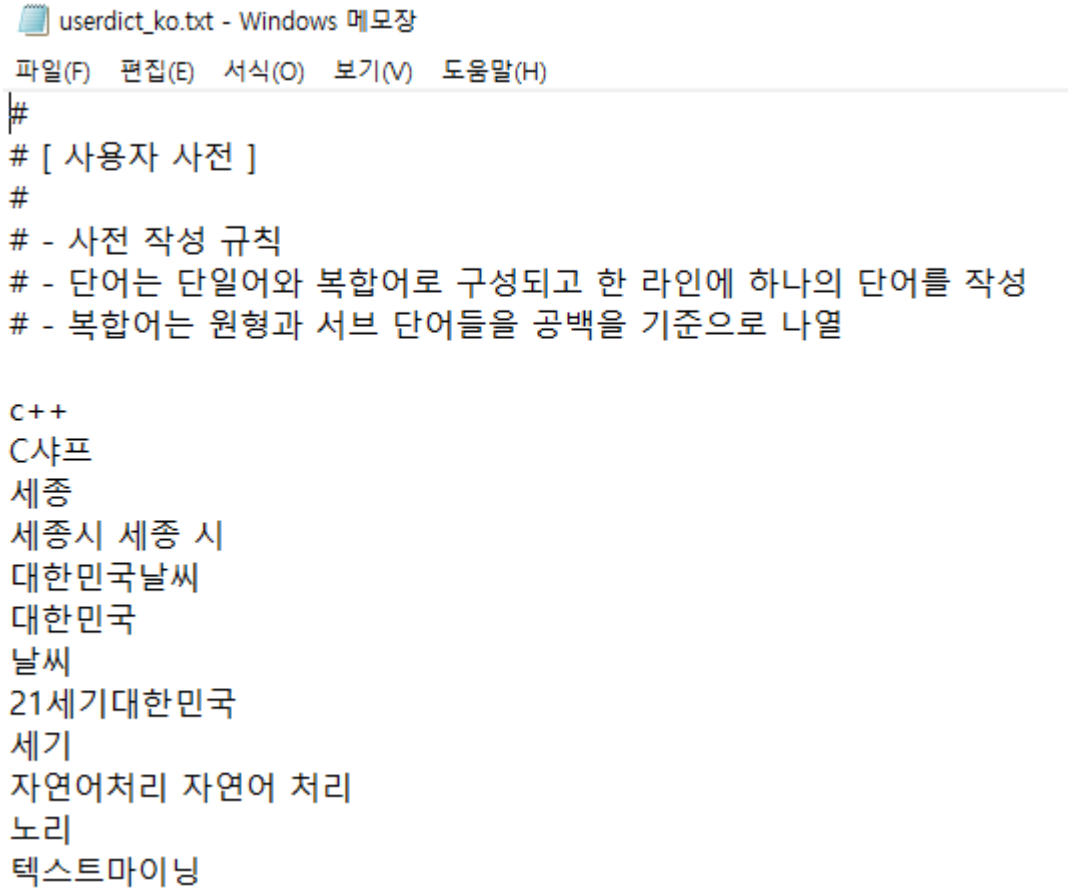
동의어 사전은 ~/pynori/resources/synonyms.txt.txt 에서 수정 (곧바로 적용 가능)

하면 적용이 된다.



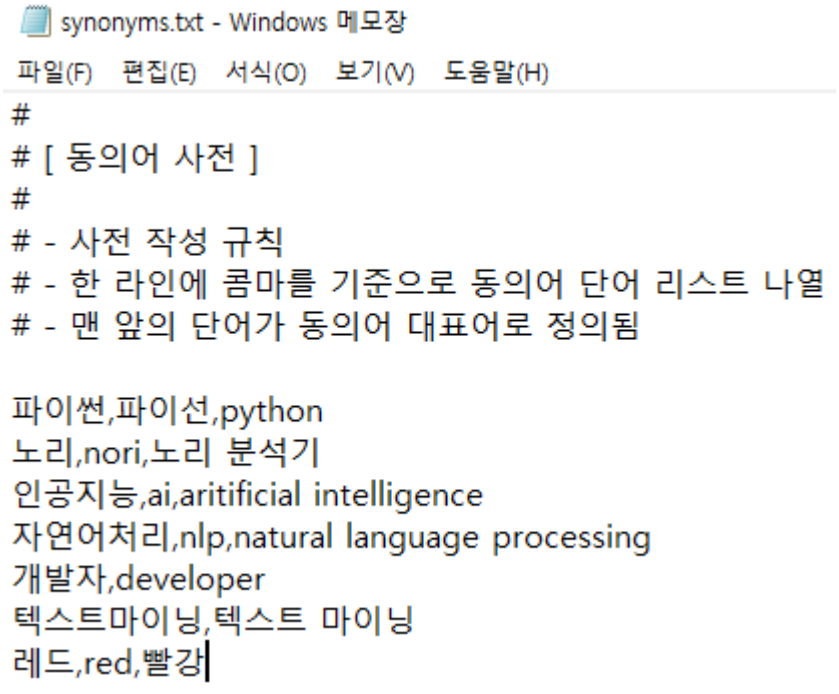
동의어에 "개발자,developer" 가 있으므로 개발자가 developer로 변환이 된 것이다.

참고로,



로 되어있다.

그렇다면 동의어 사전에 "레드,red,빨강"을 추가해본다.



[7]:
```
nori.set_option_tokenizer(decompound_mode='DISCARD', infl_decompound_mode='DISCARD')
nori.set_option_filter(mode_synonym='EXTENSION')
print(nori.do_analysis("레드 가방")['termAtt'])
```

```
['레드', 'red', '빨강', '가방']
```

[8]:
```
nori.set_option_tokenizer(decompound_mode='DISCARD', infl_decompound_mode='DISCARD')
nori.set_option_filter(mode_synonym='EXTENSION')
print(nori.do_analysis("red 가방")['termAtt'])
```

```
['레드', 'red', '빨강', '가방']
```

[9]:
```
nori.set_option_tokenizer(decompound_mode='DISCARD', infl_decompound_mode='DISCARD')
nori.set_option_filter(mode_synonym='EXTENSION')
print(nori.do_analysis("elasticsearch 개발자")['termAtt'])
```

```
['elasticsearch', '개발', '자', 'developer']
```

레드로 칠 경우 red와 빨강이 나오고

Red로 칠 경우 레드와 빨강이 나온다.