[Day-01-Lecture-03]

관련 논문

> **권기성, 최운호, 김동건 (2022), "문학 작품의 거리 측정을 활용한 야담의 이본 연구",「한국고전연구 57집」87~120쪽, 한국고전연구학회.**

제공 자료(야담집 서두 단락)

okss > okssw_00

- BCGY_003.txt, ... ZSUN_003.txt (36개 파일)
- 각 파일은 30행으로 이루어져 있음. 30개의 내용 단락
- 내용 표시 중 x는 내용 누락(낙장 등)

[Drill]

- BCGY_003.txt를 엑셀에서 열어 봅시다.
- 구분기호는 '#'로 지정해야 하며 원본 파일의 코드는 65001: 유니코드(UTF-8)으로 지정해야 합니다.
- **NKES_003.txt ~ ZKES_003.txt 를 Excel에서 열어서 서로 대응되는 단락을 비교, 관찰해 봅시다**.
- VSCode.dev 에서 동일한 텍스트 파일을 열어 봅시다.

---

okss > okssw_01

- 각 30행씩 내용이 담겨 있는 36개 파일을 하나로 합했습니다.
- okss_004_001.txt
  - 30 x 36 = 1,080행의 파일이 있습니다.
- okss_004_002.txt
  - 1,080 행에서 2개씩 짝을 만들어서 Pairwise Comparison을 위한 쌍을 만들면 모두 582,660개의 단락 쌍이 만들어집니다.
  - 이 중에서 서로 단락번호가 동일한 것만 걸러내겠습니다.
    - 36개 작품에서 2개씩 고르면 모두 630개의 쌍이 만들어집니다.
    - 630개 쌍을 30행을 대상으로 하면 모두 18,900개의 비교쌍이 만들어집니다.

```python
#python
#
# okss_004_002.py
#


if __name__ == "__main__":
```

```
    # File reading
    with open("okss_004_001.txt", "r", encoding="utf-8") as f_in:
        m_lines = [l.strip() for l in f_in.readlines()]

    #print(len(m_lines))

    m_pairs = []
    for idx_i in range(len(m_lines)):
        list_cur_i = m_lines[idx_i].split('#')
        str_cno_i = list_cur_i[2]

        for idx_j in range(idx_i+1, len(m_lines)):
            list_cur_j = m_lines[idx_j].split('#')
            str_cno_j = list_cur_j[2]

            if str_cno_i == str_cno_j:
                m_pairs.append( [m_lines[idx_i], m_lines[idx_j]] )

    #print(len(m_pairs))

    # Tab으로 분리된 텍스트로 비교 대상 쌍을 출력
    with open("okss_004_002.txt", "w", encoding="utf-8") as f_out:
        for elt in m_pairs:
            str_left = '\t'.join(elt[0].split('#'))
            str_right = '\t'.join(elt[1].split('#'))
            print(str_left, str_right, sep='\t', file=f_out)

```

- 새로 생성된 파일 okss_004_002.txt를 엑셀에서 열어봅시다. **구분 기호는 탭(T)입니다**.
- okss004_002.txt 파일에서 불필요한 정보를 제거하고 이제 pairwise comparison을 위한 파일을 만들어 봅시다.
- 모든 과정을 자동화하거나 pandas dataframe 등을 이용하면 좋겠지만, 자료가 어떻게 가공되어 가는지 확인하기 위해서 단계적으로 변경해 보도록 하겠습니다.

```
#!python
#
#-*-coding=utf-8
#
# okss_004_003.py
#


if __name__ == "__main__":

    with open("okss_004_002.txt", "r", encoding="utf-8") as f_in:
        m_lines = [l.strip('\n') for l in f_in.readlines()]
        m_lines_out = []
        with open("okss_004_003.txt", "w", encoding="utf-8") as f_out:
            for line in m_lines:
                cur_line = line.split('\t')
```

```
                if len(cur_line) < 12:
                    list_elt = [cur_line[0], cur_line[5], cur_line[6], ""]
                else:
                    list_elt = [cur_line[0], cur_line[5], cur_line[6],
cur_line[11]]

                print('\t'.join(list_elt), file=f_out)
```

- 18,900 행의 데이터가 잘 저장되었나요?
