**여름방학 3주차 회의록 (21.07.18)**

4조 김태희, 안세훈, 정영훈

**7월 3주차 계획**

1. 크롤링 오류 보완

- 광고와 리뷰 업데이트로 html 구조 변화 -> 크롤링 에러

- 페이지 넘김 제한 횟수 증가

- 크롬 드라이버 메니저로 driver 오류 해결

2. 전처리 실습

- 맞춤법 검사, 토큰화 등을 이용해 다양한 문장 전처리 해보기

- 다양한 토크나이저 써보며 비교

3. 모델링 사전조사 및 실습

- Pororo 가상환경 구축 해보기

- 긍부정 분석, 주제 분류, 예상 별점 등 실습

**7월 3주차 활동 결과 요약**

1. 크롤링 도메인(네이버 플레이스)한정 완전 구현

2. 전처리 실습 완료

3. 모델링 환경 구축 및 실습 완료

-> 소스코드: <https://github.com/ashpurple/Graduation-Project/blob/main/Modeling/modeling.py>

**전처리 실습**

김태희

# -\*- coding: utf8 -\*-  
  
from konlpy.tag import Okt  
from hanspell import spell\_checker  
from konlpy.tag import Kkma  
from soynlp import DoublespaceLineCorpus  
from soynlp.word import WordExtractor  
from soynlp.tokenizer import LTokenizer  
import kss  
from kss import split\_sentences  
  
  
  
texts = [ "조용하고 공부에 집중하기 좋은 곳 이예요 !!!!!", "서현 스터디 카페중에 젤 조음", "진짜 분위기 좋다 알바생분들도 친절 ㅠㅠ", "편안한 분위기여서 공부 잘되요", "내부도 적당히 크고 자리마다 무선 충전기 설치 되어 있는 등 시설이 되게 좋아요!!제공되는 것도 많구요 건물이 낡아서 화장실이 후지긴 한데 어쩔 수 없죠ㅠ음료 포함 정기권 끊었는데 음식 종류도 많구 아직 많이 먹어보진 않았지만 맛있어요!볶음밥이랑 베이커리류도 팔아서 식사 해결도 가능해요ㅎㅎ다만 음료는 만들어주시는 알바 분들에 따라 맛 차이가 큰거 같아요. 그리고 창가 쪽에 앉으면 비둘기 등등의 각종 새들이 자꾸 와서 달려 있는 간판 위에 앉고 부딪히고 막 그래서 소리가 좀 나요ㅠㅠ구구구하고 우는 소리도 들리고 날아가는 소리,부딪히는 소리 전부 잘 들리는게 조금 거슬린다는 점 감안해야 해요!" ]  
  
i=0  
for x in texts :  
 i=i+1  
 print("<" ,i,">\n")  
 print(" 원래 문장 : ", x)  
  
 t= kss.split\_sentences(x)  
 print("문장토큰화 : ",t)  
  
 spelled\_sent = spell\_checker.check(x) #t(토큰화된 문장)넣으면 에러  
 hanspell\_sent = spelled\_sent.checked  
 print(" 맞춤법 검사 후 : ", hanspell\_sent)  
  
 okt = Okt()  
  
 print("\*\*konlpy\_okt\*\*")  
 print(" <형태소 추출> ")  
 print(okt.morphs(hanspell\_sent))  
 print(" <품사 태깅> ")  
 print(okt.pos(hanspell\_sent))  
 print(" <명사 추출> ")  
 print(okt.nouns(hanspell\_sent))  
  
 kkma = Kkma()  
  
 print("\*\*konlpy\_kkma\*\*")  
 print(" <형태소 추출> ")  
 print(kkma.morphs(hanspell\_sent))  
 print(" <품사 태깅> ")  
 print(kkma.pos(hanspell\_sent))  
 print(" <명사 추출> ")  
 print(kkma.nouns(hanspell\_sent))  
  
 print("\*\* soynlp\_ LTokenizer \*\*")  
 corpus = DoublespaceLineCorpus("study\_cafe\_review.csv") # 학습할 문서  
  
 word\_extractor = WordExtractor()  
 word\_extractor.train(corpus)  
 word\_score\_table = word\_extractor.extract() # 학습 완료  
  
 scores = {word: score.cohesion\_forward for word, score in word\_score\_table.items()}  
 l\_tokenizer = LTokenizer(scores=scores)  
 print(l\_tokenizer.tokenize(hanspell\_sent, flatten=False))  
  
 print("\n\n")

원래 문장 : 내부도 적당히 크고 자리마다 무선 충전기 설치 되어 있는 등 시설이 되게 좋아요!!제공되는 것도 많구요 건물이 낡아서 화장실이 후지긴 한데 어쩔 수 없죠ㅠ음료 포함 정기권 끊었는데 음식 종류도 많구 아직 많이 먹어보진 않았지만 맛있어요!볶음밥이랑 베이커리류도 팔아서 식사 해결도 가능해요ㅎㅎ다만 음료는 만들어주시는 알바 분들에 따라 맛 차이가 큰거 같아요. 그리고 창가 쪽에 앉으면 비둘기 등등의 각종 새들이 자꾸 와서 달려 있는 간판 위에 앉고 부딪히고 막 그래서 소리가 좀 나요ㅠㅠ구구구하고 우는 소리도 들리고 날아가는 소리,부딪히는 소리 전부 잘 들리는게 조금 거슬린다는 점 감안해야 해요!

맞춤법 검사 후 : 내부도 적당히 크고 자리마다 무선 충전기 설치되어 있는 등 시설이 되게 좋아요!! 제공되는 것도 많고요 건물이 낡아서 화장실이 후지긴 한데 어쩔 수 없죠ㅠ음료 포함 정기권 끊었는데 음식 종류도 많고 아직 많이 먹어보진 않았지만 맛있어요! 볶음밥이랑 베이커리류도 팔아서 식사 해결도 가능해요ㅎㅎ다만 음료는 만들어주시는 알바 분들에 따라 맛 차이가 큰 거 같아요. 그리고 창가 쪽에 앉으면 비둘기 등등의 각종 새들이 자꾸 와서 달려 있는 간판 위에 앉고 부딪히고 막 그래서 소리가 좀 나요ㅠㅠ구구구하고 우는소리도 들리고 날아가는 소리, 부딪히는 소리 전부 잘 들리는 게 조금 거슬린다는 점 감안해야 해요!

<형태소 추출>

['내부', '도', '적당히', '크고', '자리', '마다', '무선', '충전기', '설치', '되어', '있는', '등', '시설', '이', '되게', '좋아요', '!!', '제공', '되는', '것', '도', '많고요', '건물', '이', '낡아서', '화장실', '이', '후지', '긴', '한데', '어쩔', '수', '없죠', 'ㅠ', '음료', '포함', '정기권', '끊었는데', '음식', '종류', '도', '많고', '아직', '많이', '먹어', '보진', '않았지만', '맛있어요', '!', '볶음밥', '이랑', '베이커리', '류', '도', '팔아서', '식사', '해결', '도', '가능해요', 'ㅎㅎ', '다만', '음료', '는', '만들어주시는', '알바', '분', '들', '에', '따라', '맛', '차이', '가', '큰', '거', '같아요', '.', '그리고', '창가', '쪽', '에', '앉으면', '비둘기', '등등', '의', '각종', '새', '들', '이', '자꾸', '와서', '달려', '있는', '간판', '위', '에', '앉고', '부딪히고', '막', '그래서', '소리', '가', '좀', '나요', 'ㅠㅠ', '구구구', '하고', '우는', '소리', '도', '들리고', '날아가는', '소리', ',', '부딪히는', '소리', '전부', '잘', '들리는', '게', '조금', '거슬린다는', '점', '감안', '해야', '해', '요', '!']

<품사 태깅>

[('내부', 'Noun'), ('도', 'Josa'), ('적당히', 'Adjective'), ('크고', 'Verb'), ('자리', 'Noun'), ('마다', 'Josa'), ('무선', 'Noun'), ('충전기', 'Noun'), ('설치', 'Noun'), ('되어', 'Verb'), ('있는', 'Adjective'), ('등', 'Noun'), ('시설', 'Noun'), ('이', 'Josa'), ('되게', 'Adverb'), ('좋아요', 'Adjective'), ('!!', 'Punctuation'), ('제공', 'Noun'), ('되는', 'Verb'), ('것', 'Noun'), ('도', 'Josa'), ('많고요', 'Adjective'), ('건물', 'Noun'), ('이', 'Josa'), ('낡아서', 'Adjective'), ('화장실', 'Noun'), ('이', 'Josa'), ('후지', 'Noun'), ('긴', 'Verb'), ('한데', 'Eomi'), ('어쩔', 'Adverb'), ('수', 'Noun'), ('없죠', 'Adjective'), ('ㅠ', 'KoreanParticle')

**모델링 실습**

안세훈

목표: 한국어 자연어 처리 모델을 이용한 실습을 통해 전반적인 작동 원리를 익힌다.

**1. 환경 조성**

- Anaconda 설치  
- 가상 환경 만들고 활성화  
- 가상환경 내에서 Pytorch, Cuda, Fairseq, pororo 설치  
- pandas와 같은 데이터 분석 라이브러리 설치

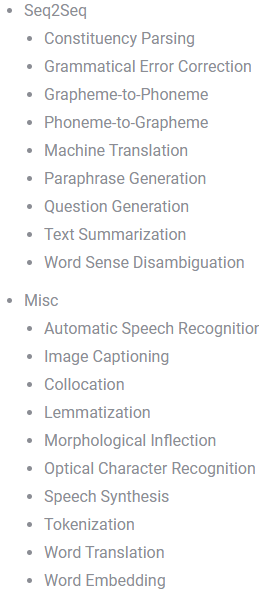
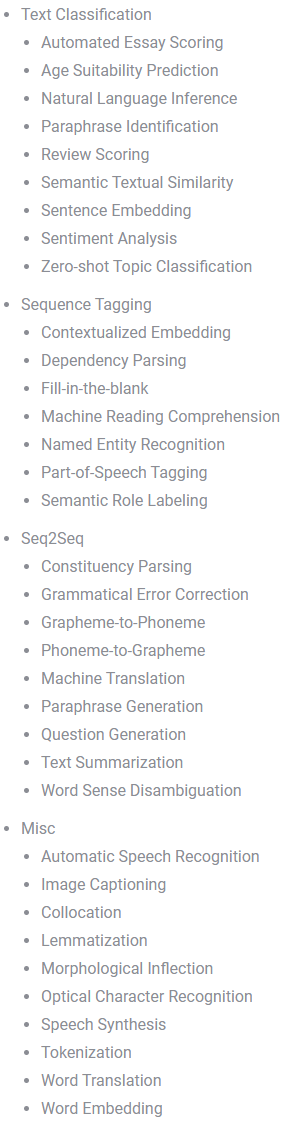
**2. 모델링**



**<오픈 소스 사용>  
Pororo:** 카카오브레인에서 개발한 다양한 자연어 태스크에 대응 가능한 통합된 형태의 자연어 프레임워크  
<https://kakaobrain.github.io/pororo/>

**장점**: 한국어 자연어 처리에서 여러 태스크들을 하나로 통합하여 구성한 유일한 프레임워크로 올해 초 공개되어 지속적인 성능 개선을 기대할 수 있다.

**지원 태스크**



**<실습 방법>**

**테스트 예제**: 스터디 카페 리뷰를 크롤링하여 (850 여개) 그 중 분류가 뚜렷한 케이스들을 샘플링 하였다

**사용 함수**: Pororo에서 제공하는 함수 중 리뷰분석에서 유용할 만한 함수를 골라 테스트해보았다  
-> 긍부정 감성분석, 예상 점수 분석, 형태소 분석, 주제 분석

**과정**: 엑셀 데이터를 데이터 프레임으로 만들기 -> 리뷰 열만 슬라이싱 -> 샘플 행 추출  
-> hanspell 을 이용해 맞춤법 교정 -> 다양한 함수에 파라미터로 전달

**<긍부정 감성분석>**

Pororo(task="sentiment", model="brainbert.base.ko.shopping", lang="ko")

Train Set: 네이버 쇼핑 리뷰 20만건, 긍부정 기준 ( 4,5 긍정 / 1,2 부정)  
<https://github.com/bab2min/corpus/tree/master/sentiment>

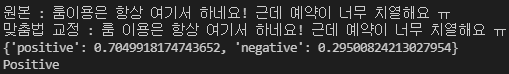
**input**

print(sa(review, show\_probs=True)) # 긍부정 비율

print(sa(review)) # 긍부정 판단

-> 맞춤법만 교정한 텍스트 데이터를 넣어주면 된다

**output**



-> 긍부정 비율과 긍부정 결과값 모두 출력이 가능하다

**활용방안**  
- 리뷰마다 긍정 리뷰 / 부정 리뷰 태깅 => 정렬 가능  
- 한 매장의 전체 긍부정도 통계  
- 추후 가짜 리뷰 판별 지표로도 활용 가능

**<리뷰 예상 점수 분류>**

Pororo(task="review", lang="ko")

Train Set: 외부 데이터

**Input**

print(rating(review))

**output**  
: 1점 ~ 5점사이 실수형

**활용방안**- 별점이 누락된 리뷰도 데이터로 사용 가능  
- 긍부정 분석의 보조지표로 활용 가능  
- 실제 별점과 비교했을 때 정확성이 떨어지는 단점이 있음

**<형태소 분석>**

Mecab(dicpath=r"C:\mecab\mecab-ko-dic")

Konlypy 패키지의 Mecab 사용

사용 이유: 토크나이저 중 소요시간이 가장 짧음, Pororo의 토크나이저는 mecab 기반



**Input**

print(mecab.morphs(review)) # 토큰화

print(mecab.pos(review)) # 품사 태깅

**output**



**<주제 분석>**

Pororo(task="zero-topic", lang="ko")

Train Set: KorNLI (Ham et al. 2020)  
모델: brainbert.base.ko.kornli

**input**

print(zsl(review, ["음식", "온도","청결","분위기"]))

-> 주제 항목을 사용자가 정의할 수 있다

**output**



**활용방안**

- 각 도메인마다 관련 속성을 미리 정의할 수 있다 -> 속성 정렬 가능  
- 속성을 바탕으로 매장의 속성 별 긍부정도 출력 가능

**<해결해야될 Issue>**

- pororo 내에서 라이브러리 인식 오류, gpu 할당 관련 런타임 에러

- 광고성 리뷰 판별 라이브러리 조사