|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2021 ETRI 오픈 API 활용사례 공모전 참가신청서 | | | | |
| **팀명** | | **뭘해야하조** | | |
| **팀원** | | **이현준, 이서우, 이아현** | | |
| **소속 기관** | | **일반(세종대 졸), 숭실대** | | |
| **공모안 제목** | | **톺핑있슈(Topping Issue)** | | |
| **공모**  **내용 요약** | | 특정 사건이 발생하게 되면, 사건이 전개되고 누적됨에 따라 파생주제들이 생겨나게 되고, 사건이 점점 더 복잡해지게 된다.  본 프로젝트에서는 이러한 특성을 갖는 현실의 이슈들에 대한 정보를 핵심 단어와 주변 단어로 구분하여 사용자에게 제공하는 것을 목적으로 하였다.  사용자는 제공받은 정보를 토대로 과거 발생한 사건에 대한 흐름을 보다 용이하게 습득할 수 있고, 이를 토대로 다양하고 방대한 시사 문제를 파악할 수 있을 것이다. | | |
| **개인정보**  **수집 및**  **활용동의** | | □ 개인정보 수집 및 이용 목적 : 응모자 관리  √ ‘2021 ETRI 오픈 API 활용사례 공모전’ 관련 심사를 위한 개인정보 수집 및 이용  □ 수집 대상 개인정보 항목  √ 성명, 생년월일, 휴대폰, 이메일, 소속기관, 및 직위 등 세부 정보  √ ‘2021 ETRI 오픈 API 활용사례 공모전’에 제출한 활용사례와 모든 관련 사항  □ 개인정보 보유 및 이용 기간  √ 심사/상금지급 및 기타 관련법령상 필요한 관리가 요구되는 시점까지 보유 및 이용  √ 수상한 수기를 홍보에 활용하는데 필요한 시점까지 보유 및 이용  √ 개인정보 수집 동의를 거부할 권리 : 개인정보 수집 및 활용에 대한 동의서의 제출을 거부할 권리가 있으나, 미동의 시 공모전 응모가 거부될 수 있음. 개인정보는 공모전 및 입상작의 홍보 이외의 다른 목적으로 사용하지 않습니다.  ※ 제공된 개인정보의 이용을 거부하고자 할 때에는 개인정보 관리책임자를 통해 열람, 정정, 삭제를 요구할 수 있음 | | |
| **본인은 한국전자통신연구원의 ‘개인정보 수집 및 이용’에 관한 설명을 모두 이해하였으며, 이에 동의합니다. (해당란에** √ **표시)**  **동의(**√**) 비동의( )** | | |
| **팀원 1**  **(팀장)** | **성 명** | 이현준 | **생년월일** | 1994.03.21 |
| **휴대폰** | 010-5481-3736 | **E-mail** | [hyunjun.bruce.lee@gmail.com](mailto:hyunjun.bruce.lee@gmail.com) |
| **주소** | 서울시 강남구 역삼로 90길 30, 101 | | |
| **구분** | **직장인( ), 일반(**√**), 학생( 대 / 고 / 중 )** | | |
| **직장/학교명** | 세종대학교 | **직위/학년** | 졸업 |
| **팀원 2** | **성 명** | 이서우 | **생년월일** | 1994.03.21 |
| **휴대폰** | 010-5481-3736 | **E-mail** | [tegfsl@gamil.com](mailto:tegfsl@gamil.com) |
| **주소** | 서울특별시 구로구 중앙로 5길 24, 2동 404호 | | |
| **구분** | **직장인( ), 일반( ), 학생( 대**√ **/ 고 / 중 )** | | |
| **직장/학교명** | 숭실대학교 | **직위/학년** | 4학년 |
| **팀원 3** | **성 명** | 이아현 | **생년월일** | 1998.05.08 |
| **휴대폰** | 010-2889-5816 | **E-mail** | [ahyeon\_0508@naver.com](mailto:ahyeon_0508@naver.com) |
| **주소** | 서울특별시 서초구 동광로 8길 6, 203호 | | |
| **구분** | **직장인( ), 일반( ), 학생( 대**√ **/ 고 / 중 )** | | |
| **직장/학교명** | 숭실대학교 | **직위/학년** | 4학년 |
| **「2021 ETRI 오픈 API 활용사례 공모전」공지사항을 숙지하고 본 공모에 참가하고자 합니다.**  **2021년 10월 18일**  **참가자 이현준 (서명)**  **참가자 이서우 (서명)**  **참가자 이아현 (서명)**  한국전자통신연구원 귀하 | | | | |

**톺핑있슈 (topping issue)**

**예전 이슈? 최근 이슈?, 궁금해? 여기 톺핑있슈~!**

**이현준, 이서우, 이아현**

**목차**

1. **공모안 개요**
2. **활용사례 결과물(API)의 기본 개념**
3. **활용한 API의 종류 및 상세 활용 내용**
4. **활용사례 결과물의 동작 과정**
5. **활용사례 결과물의 시연 예제**
6. **활용사례 결과물의 독창성 및 우수성**
7. **활용사례 결과물의 활용 범위 및 시장성**
8. **활용 데이터의 종류 및 특징, 확보 방법**

**1. 공모안 개요**

본 공모안의 주제를 정하기 위하여 최근 시사이슈를 탐색하던 도중, 생소한 이슈를 보게 되었다. 조원들 중 그 누구도 해당 이슈에 대하여 상세히 알고있는 사람이 없어, 이에 대한 정보를 찾아보게 되었다.

그러나, 특정 이슈에 대한 정보가 정리되어 제공되는 서비스가 존재하지 않았고, 웹 검색 만으로 특정 이슈를 둘러 싸고 있는 다양한 파생 주제를 파악하는데 어려움이 있었다.

이에 특정 이슈와 그에 대한 파생주제들을 파악하기 쉽게 사용자에게 제공하는 서비스를 기획하게 되었다.

파생 주제들을 어떻게 파악할 수 있을 것인가에 대하여 기술적 측면을 고민하던 중, BERT의 CLS토큰이 문맥상의 의미를 어느정도 보유하고 있다는 점에서 착안하여 S-BERT를 기반으로 한 문장 임베딩을 통해 해당 과제를 달성할 수 있을 것이라고 생각하였다.

따라서, 본 프로젝트에서는 BERT를 기반으로 기간내 이슈의 변화량과 이슈에 대한 파생 주제의 변화를 서비스화 하기로 하고자 한다.

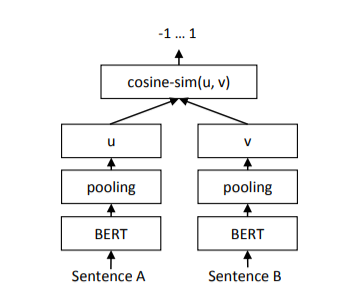
**2. 활용사례 결과물(API)의 기본 개념**

**2.1 한국어 BERT 언어 모델 - 004\_bert\_eojeol\_tensorflow**

본 프로젝트에서는 ETRI에서 제공하는 BERT모델을 하기의 두가지 방향으로 이용하였다.

**2.1.1 News Clustering (Cosine similarity based)**

본 과제에서 뉴스 기사 분류의 목적은 단순 정치, 경제, 사회와 같은 대분류가 아닌 ‘Semantic Textual Similarity’ 즉, 문맥상의 의미를 중심으로 정치와 같은 대분류 내에서 발생하는 다양한 사건을 중심으로 분류를 진행하는 것을 목적으로 했다.

Bert 모델은 공개된 이래 많은 주목을 받았고 현재까지 NLP 분야의 State of Art 중 하나로서 다양한 과제에 활용되고 있다. 본 News Clustering 에서는 이러한 BERT모델을 이용하여 뉴스 기사의 제목을 Embedding하고, 제목에 대한 Sentence vector를 이용하여 뉴스 기사 분류를 진행하였다.

모델 구조의 경우, Sentence-Bert: Sentence Embeddings using Siamese BERT-Network(2019)에 소개된 Bert 모델을 이용한 Siam network에 Cosine Similarity Layer를 얹은 구조를 이용하였다.

학습 데이터의 경우, 현재 공개된 2종(KLUE, KorNLU)의 STS data set을 이용하였으며, 각 Date set은 Label을 본 모델의 out put인 cosine similarity와 매칭시키기 위하여 조정한 후 이용되었다. (8. 활용데이터의 종류 및 특징 참고)

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (3p)

학습 시 Loss Function으로는 계산된 Cosine simularity 간의 error를 반영하기 위해 MSE를, optimizer로는 Adam을 이용하였으며, 1e-5의 learning rate으로 KLUE-STS 데이터로만 1epoch(batch size = 20) 학습을 진행한 모델을 최종적으로 채택했다.

**2.1.2 Sentiment analysis**

본 과제에서 감성 분석은 사용자들에게 특정 사건에 대한 객관적인 정보 전달을 목적으로 이용된다. 네이버와 다음의 이용자 성향이 다르기에 두 사이트에서 제공되는 뉴스 서비스 댓글을 각각 분류를 한 후 하나의 사건에 대해 각 사이트의 사용자들이 어떻게 반응하고 있는지를 제시하는 것을 목표로 했다.

BERT모델을 이용한 감성 분류의 경우 이미 다양한 시도가 있었고, 이에 대한 모델과 학습데이터가 오픈소스로 공개되어 있다. 그 중 본 과제에서는 BERT모델의 NSP출력부분(NSP-Dense.output)에 Dense(activation = sigmoid)를 얹은 모델을 사용하였고, 훈련 데이터로는 “Naver Sentiment movie corpus V1.0”과, AI HUB에서 제공되는 “한국어 감정 정보가 포함된 단발성 데이터 대화 데이터 셋”을 과제에 맞게 조정하여 이용하였다. (8. 활용데이터의 종류 및 특징 참고)

학습 시 Loss Function으로는 BCE(binary cross entropy)를, optimizer로는 Adam을 이용하였으며, 1e-5의 learning rate으로 각 데이터를 1epoch(batch size = 10)씩 학습한 모델을 최종적으로 채택했다.

**2.2 언어 분석기술(문어) - 형태소 분석 API**

ETRI에서 제공해주는 형태소 분석 API의 경우, 상기 News Clustering의 결과를 보정하는 용도로 이용되었다.

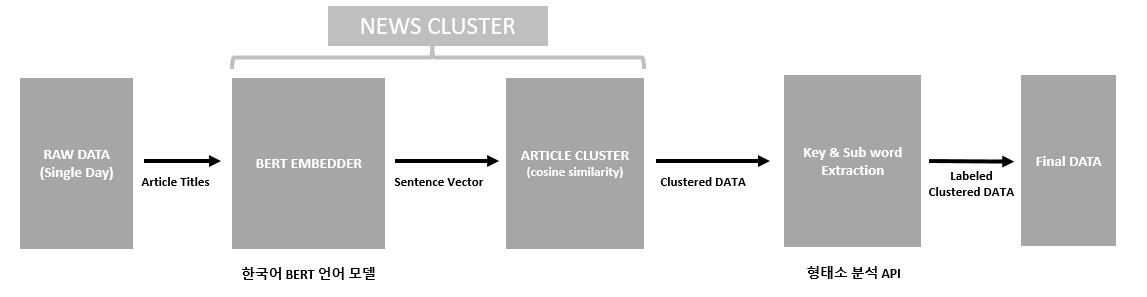
본 News Clustering 결과의 경우 의미론적 유사도를 계산하기에 주체가 동일하더라도 주체의 행위에 따라 별개의 군집으로 분류되었다. 이때 동일 주체를 하나의 주제로 구분한다. 그리고 이에 대한 여러 파생 주제를 소분류로 구분하기 위해 형태소 분석 API를 이용하여 주제에 해당하는 Key word를, 파생 주제에 해당하는 Sub word를 추출한다. (하기 예시 표 참고)

|  |  |  |
| --- | --- | --- |
| **기사 제목** | **Key word** | **Sub word** |
| 송두환 인권위원장 인사청문회 통과…'이재명 무료 변론' 공방 | 이재명 | 무료 변론 |
| 박주민 '이재명 무료변론' 의혹에 "이렇게까지 해야 하나" |
| 이재명 “초등생 저녁 7시까지 돌봄…공공어린이집 50% 이상으로” | 공공어린이집, 50%이상 |
| [속보] 이재명 "공공 어린이집 이용 비율 50%이상으로 늘리겠다" |

**3 활용한 API의 종류 및 상세 활용 내용**

**3.1 News Clustering (Cosine similarity based)**

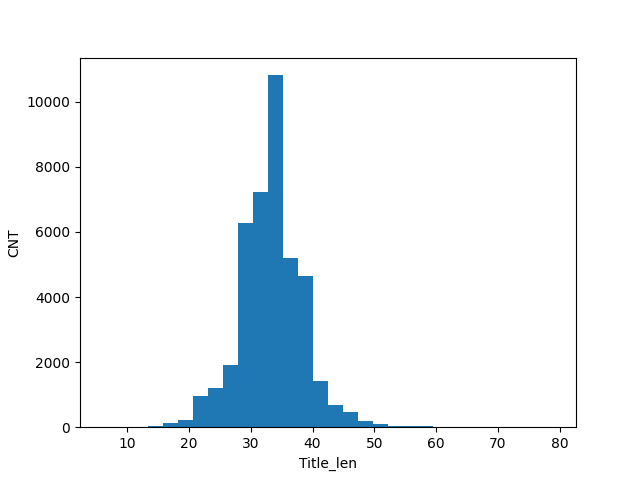
* 한국어 BERT 언어 모델 - 004\_bert\_eojeol\_tensorflow
* 형태소 분석 API



**3.1.1 BERT EMBEDDER**

위 파이프 라인에서 기사의 제목에 대한 Sentence Vector를 구하는 모델의 경우 Fine Tuning DATA를 기준으로 아래의 4가지 모델을 후보 군으로 설정하였으며, 각 Fine Tuning의 epoch, learning rate, batch size는 1, 1e-5, 20으로 통일하여 훈련을 진행하였다.

* **Base Model**: 별도의 Fine Tuning을 진행하지 않고 pre Training된 ETRI-BERT의 embedding 출력 중 CLS토큰에 해당하는 값을 Sentence vector로 이용
* **KLUE, KorNLU, KLUE+KorNLU Model**: 각 데이터 셋을 이용하여 S-BERT 구조로 Fine Tuning을 진행한 모델의 Embedding 출력 값을 Sentence vector로 이용



상기 후보 모델이 공통으로 갖는 Hyper parameter인 SEQ\_LEN의 경우 기사 제목들의 길이에 대한 99% 백분위 수(47)를 반올림 한 값인 50으로 결정하였다.

**3.1.2 ARTICLE CLUSTER**

문장 간 유사도 판별의 기준 점수의 경우 Train을 진행한 데이터 셋과, 실제 서비스에 사용될 데이터 셋 간의 불일치로 인하여 기존 Train data set의 Label값을 기준으로 결정할 수 없었다.

따라서 서비스에 사용될 데이터를 기준으로 “두 문장이 같은 의미이다” 라고 판단할 수 있는 기준 코사인 유사도 점수를 다음의 과정을 통해 각 후보 embedding모델별로 재정의 하였다.

**평가 방식**: 기준 코사인 유사도를 다르게 설정하며 출력한 결과값에 대하여 정성 평가를 진행

**기준 DATA**: 2021-04-13일자 정치-청와대, 경제-금융 면의 기사 각 100개 (random seed = 1004)

**평가 기준**: min(오 분류 개수 + 미 분류 개수)

**조절 방법**: 오분류 < 미분류 => 기준 하향 조정, 오분류 > 미분류 => 기준 상향 조정

**오 분류**: 군집화가 되었지만, 해당 군집을 구성하는 대다수의 기사들에 대한 요점과 다른 기사

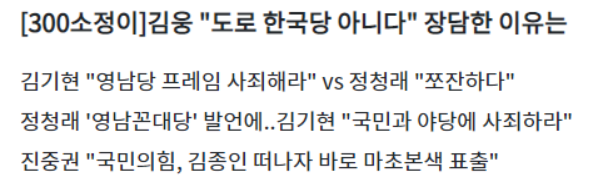
**미 분류**: 분류가 진행되지 않아 0번 라벨을 부여받은 기사 중, 해당 기사의 요점으로 군집화가 진행된 군집이 있는 기사, 0번 라벨 내에서 새로운 군집을 형성하는 것이 타당한 기사

위의 과정으로 도출된 모델 별 최적의 기준 유사도를 적용한 News Cluster의 성능은 다음과 같았다. ()내의 숫자는 결정된 기준 코사인 유사도

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **모델 명** | **BASE(0.95)** | **KLUE(0.65)** | **KorNLU(0.7)** | **KLUE+KorNLU(0.7)** |
| **오분류+미분류** | **25** | **8** | **13** | **11** |

상기 표의 결과에 따라 본 서비스에서는 **KLUE Model의 Embedding**과 **0.65의 유사도** 기준을 이용하였다.

**결과 예시**

도로 한국당, 영남당, 영남꼰대당, 국민의힘 4개의 단어가 동일한 정당을 지칭하지만 표기되는 단어가 다르다. 하지만 본 모델의 경우 그림의 4개 기사를 같은 군집으로 분류하였다.

**3.1.3 KEY & SUBWORD EXTRACTION**

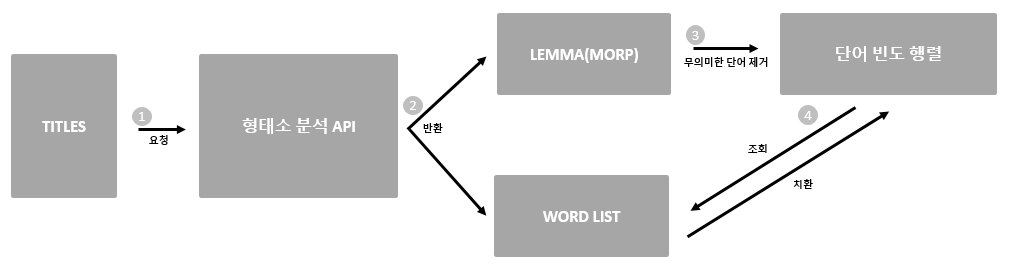
본 과제에서의 Key & Sub word extraction단계에서는 문맥상의 유사도로 분류된 뉴스 기사 군집에 대하여 사건의 중심이 되는 핵심 키워드와 해당 사건과 연관된 주변 단어들을 추출하는 것을 목적으로 정의하였다.

상기의 목적을 달성하기 위하여 형태소 분석 API를 이용하여 Key & Sub word를 추출한 후, 이를 상용 단어로 변환하는 과정을 진행하였다.

Key & Sub word를 집계함에 있어, 한국어의 경우 동일 단어일 지라도 사용되는 형태가 다양하고, 이에 따른 단어의 변이 또한 빈번하게 발생하는 특성을 지니고 있기에 형태소 분석 API를 이용하여 뉴스 제목에 등장하는 단어들을 표제어로 변환한 후 단어를 집계하였다.

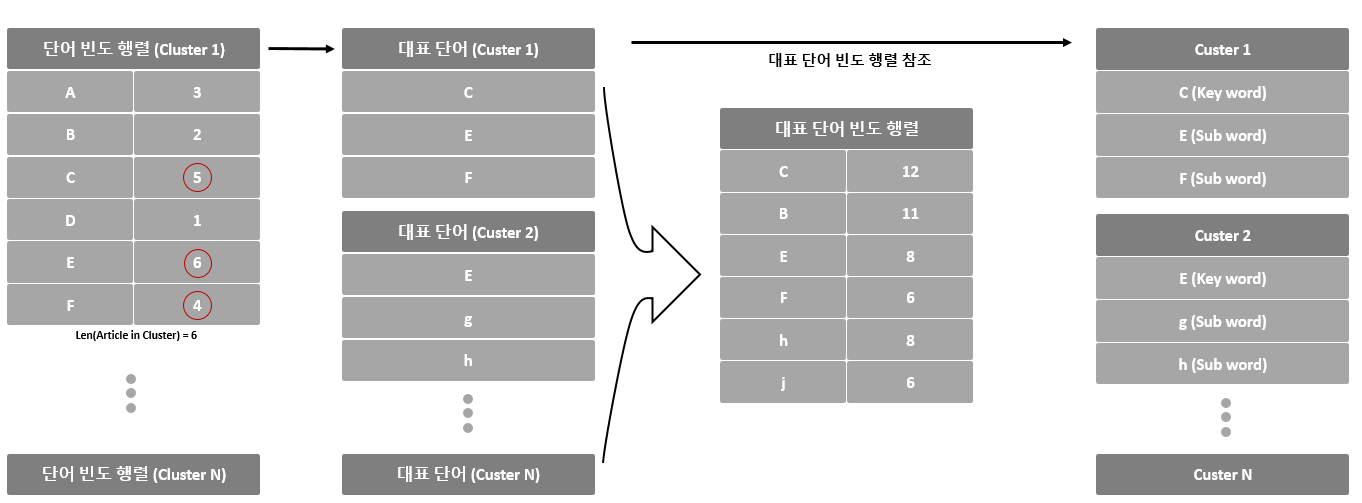
단, 추출된 Key & Sub word가 본 서비스에서 사용자에게 직접적으로 노출되기에 품사 정보를 이용하여 의미를 담지 않은 품사(XS, VC 등)를 제거한 후, 표제어로 집계된 Key & Sub word를 형태소 분석 API에서 표제어와 함께 제공되는 Word단위 구분을 이용하여 사용 단어의 형태로 복원하였다.

이와 동시에, ‘국민의힘’, ‘국민의당’ 과 같은 합성어를 형태소 분석 API에 통과 시킬 시 ‘국민’, ‘의’, ’힘’ / ‘국민’, ‘의’, ‘당과 같이 개별 단어 단위로의 변환이 이루어지는 것을 확인하여, Word단위 구분을 통해 하나의 단어로 변환하였다.



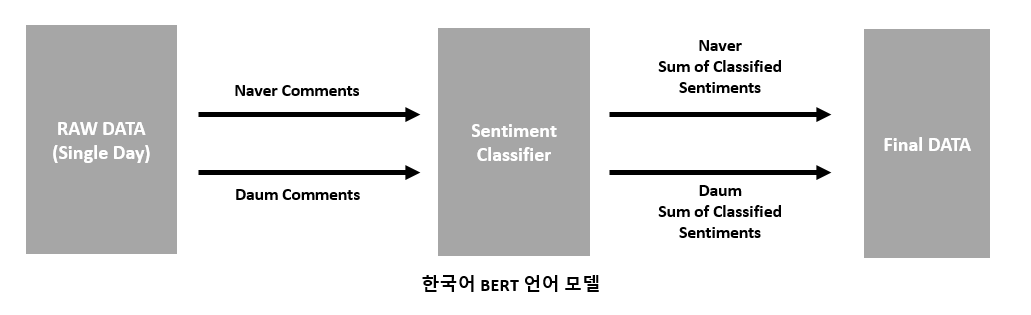
위의 작업을 거친 뒤, 특정 군집에 포함된 기사들에 대한 단어 빈도 행렬을 구성하여 등장 빈도가 군집내 포함된 기사 개수의 절반 이상인 단어들을 해당 군집을 대표하는 단어들로 추출하였다.

전체 군집의 추출된 대표 단어들을 대상으로 단어 빈도 행렬을 구성한 뒤, 단어 등장 빈도를 점수로 이용하여, 특정 군집을 대표하는 단어들 중 가장 점수가 높은 단어를 Key word로, 이 외의 단어들을 Sub word로 다시금 추출하였다.



**3.2 Sentiment analysis**

* 한국어 BERT 언어 모델 - 004\_bert\_eojeol\_tensorflow



댓글에 대한 감성 분류의 목적은 본 서비스를 이용하는 사용자들에게 해당 주제에 대한 다른 사람들의 의견을 최대한 객관적으로 제공하는 것을 목적으로 하였다.

이에 따라 사용자들의 성향이 나뉜다고 알려진 한국의 양대 포털사이트를 별개로 구분하여 각각의 댓글에 대하여 독립적으로 감성 분류를 진행하였다.

**3.2.1 Sentiment Classifier**

감성 분류 모델의 경우 Fine Tuning DATA를 기준으로 아래의 3가지 모델을 후보군으로 설정하였으며, 각 Fine Tuning의 epoch, learning rate, batch size는 1, 1e-5, 10으로 통일하여 훈련을 진행하였다.

* **AI HUB, Naver Movie Review**: 각각의 데이터 셋을 Train test split 진행 후 이용
* **AI HUB + NMR**: 순차적으로 1epoch식 훈련, test set은 두 데이터를 혼합 이용하였다.

상기 후보 모델이 공통으로 갖는 Hyper parameter인 SEQ\_LEN의 경우 각 사이트에서 제한하는 최대 댓글의 길이인 300자에 CLS토큰, SEP토큰을 더하여 302자로 통일하였다.

각 모델 별 성능은 다음과 같았다.

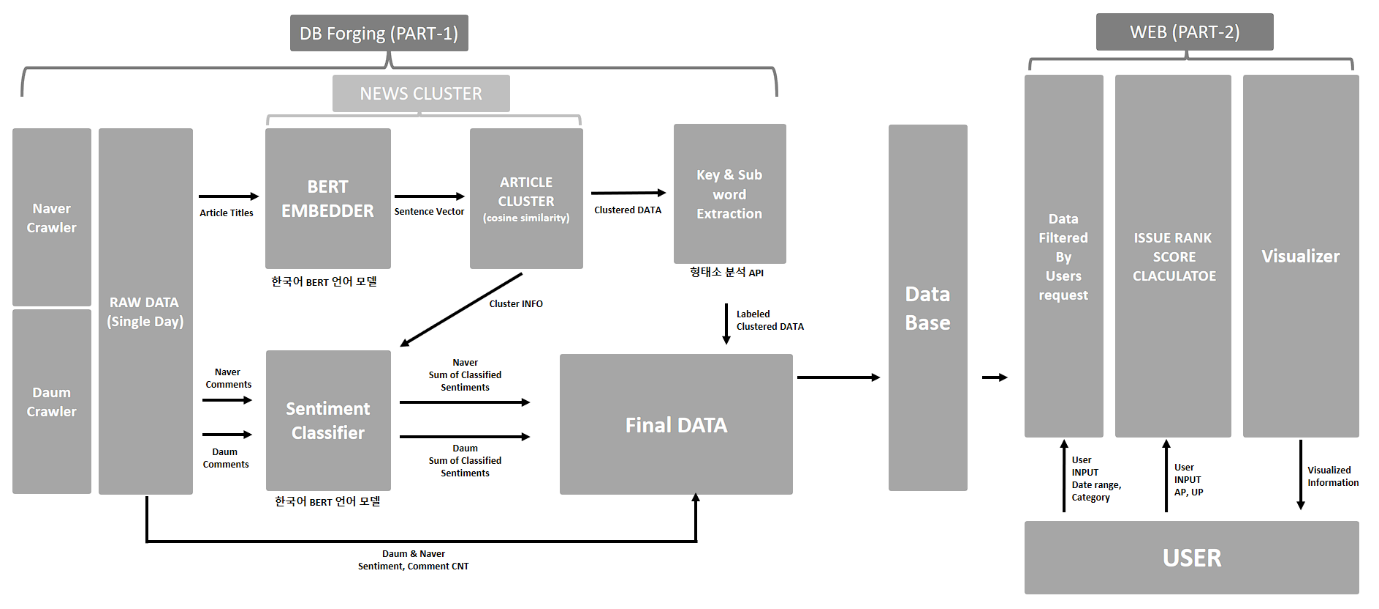
|  |  |  |  |
| --- | --- | --- | --- |
| 모델 명 | AI HUB | NAVER | AI + NAVER |
| Validation ACC | 0.9354 | 0.8982 | 0.9371 |

상기 표의 결과에 따라 본 서비스에서는 **AI + Naver Model**을 이용하였으며, 서비스에 사용될 실제 댓글들에 대하여 정성평가를 진행한 결과 모호한 댓글에 대해 평가가 정확하게 이루어지고 있지 않음을 발견하였다.

이에 모델의 Sigmoid출력 값을 0.5를 기준으로 한 2진 분류가 아닌, 중립 구간(0.4~0.6)을 지정하여 3진 분류를 진행하였다.

**4. 활용사례 결과물의 동작 과정**

**톺핑있슈(Topping ISSUE) 서비스 전체 흐름도**

****

본 프로젝트에서 구축한 모델을 이용하여 서비스화를 진행함에 있어, 유저의 인풋을 받은 후 모든 작업을 진행할 경우 사용자의 대기시간이 매우 길어지는 문제점이 있었다. 이 문제를 해결하기 위해, 위 그림처럼 DB를 기준으로 서비스 프로세스를 2분할하여 사용자의 요청이 발생했을 때 최대한 적은 양의 연산을 거칠 수 있게 설계하였다.

본 서비스는 양대 포털사이트에 달리는 댓글을 수집하여 정보를 제공해 주기에, 기사에 사용자들의 댓글이 달릴 시간이 필요하다. 따라서 요청일 2일전까지의 기사들에 대하여 집계한 정보를 제공한다.

이슈량 정도를 측정함에 있어, 자체적으로 정의한 계산식을 이용했다.

**ISSUE RANK SCORE**: **AP x (군집내 기사 개수) + UP x (군집내 댓글 수 + 군집내 공감 수)/2**

* 본 점수는 특정 사건에 대하여 언론에서 주목하는 정도와, 이에 대해 네티즌이 주목하는 정도를 측정하여 사용자의 기호에 맞게 언론 가중치(AP)와 네티즌 가중치(UP)를 설정하여 데이터를 조회할 수 있게 구성하였다.
* 언론주목도에 해당하는 “군집내 기사 개수”는 기간내 발생 기사를 모집단으로 MIN MAX 스케일 조정된 값이며, 네티즌 주목도 역시 이와 동일하게 댓글 수와 공감 수를 각각 모집단으로 MIN MAX 스케일 조정을 했다.

본 서비스의 흐름은 다음과 같다.

**PART - 1 (Daily Base Update)**

* 2개의 포털사이트를 대상으로 2일 전의 뉴스와 댓글에 대한 크롤링을 진행한다.
* 앞서 기술한 과정을 바탕으로 군집화를 진행하고 핵심 단어와 주변 단어를 추출한다.
* 군집화 된 기사들에 대하여 댓글의 감정을 집계한다.
* DB데이터를 업데이트 한다.

**PART – 2 (When User request)**

* 사용자로부터의 요청이 발생했을 시, 앞서 기술한 과정을 토대로 구축된 DB에서 요청 기간에 대한 데이터를 로드 한다.
* 로드 된 데이터에 대하여, 사전에 추출된 Main key word를 기준으로 랭킹 스코어를 계산한 후, 공통의 Main key word를 갖는 Sub word에 대한 랭킹스코어를 계산한다.
* 계산된 스코어를 기준으로 다음 3종의 데이터를 사용자에게 제공한다.
  + Main랭킹 기반 전체 기간 이슈량 순위 (list 형식)
  + 시간의 흐름에 따른 Main랭킹과 각 Main랭킹에 대한 Sub word순위 (그래프)

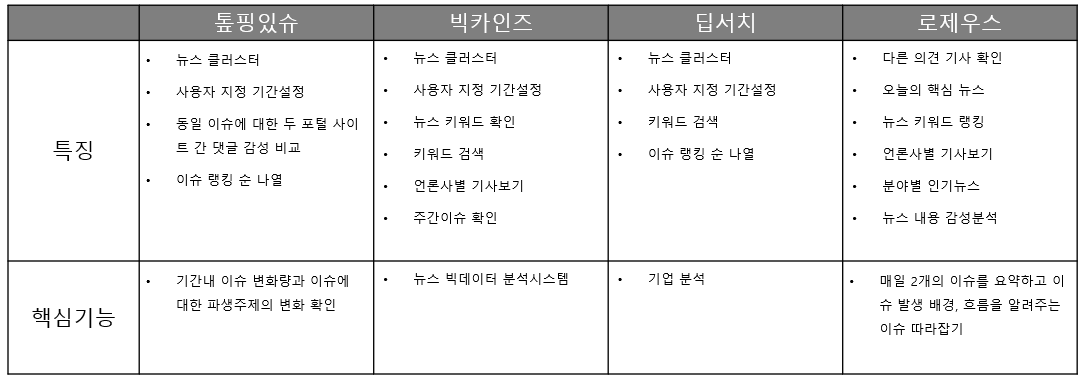
**5. 활용사례 결과물의 시연 예제**

**별첨 영상 참고(2021\_ETRI\_OPENAPI활용사례공모전\_뭘해야하조\_시연영상.mp4)**



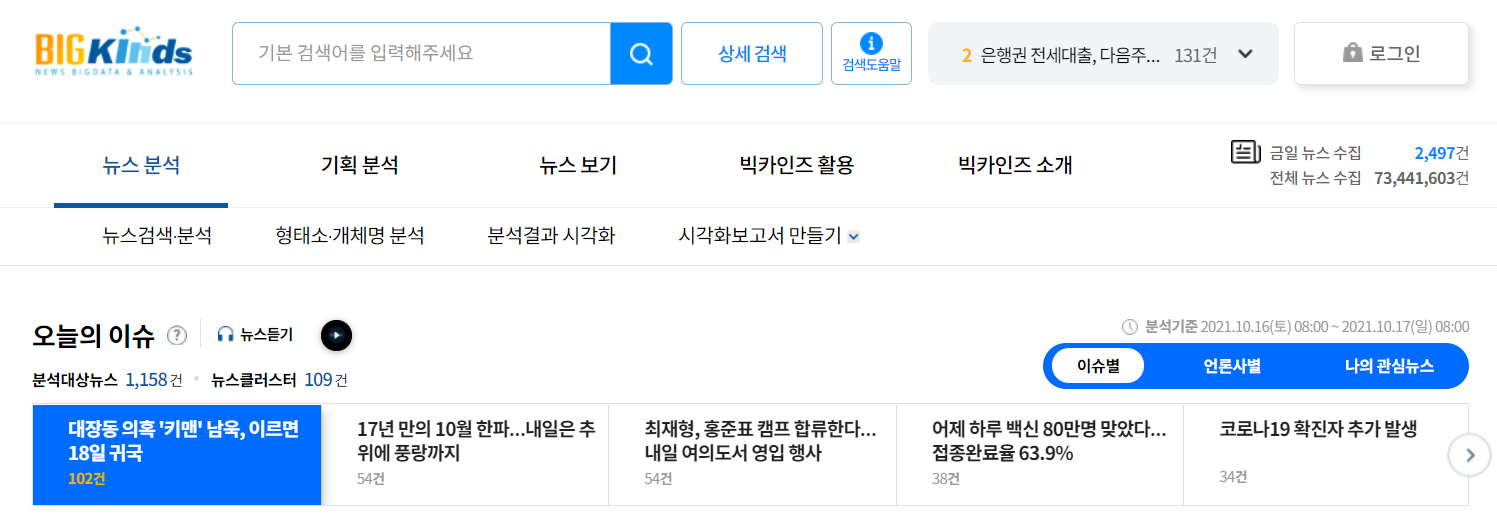
**6. 활용사례 결과물의 독창성 및 우수성**

결과물의 독창성과 우수성을 설명하기에 앞서 기존에 존재하는 뉴스 서비스 사이트들과 본프로젝트의 서비스인 톺핑있슈를 비교하여 각 서비스들의 특징과 핵심기능을 나열해 보았다.



**[빅카인즈]**

오늘의 이슈(전날 오전 8:00시 ~ 당일 오전 8:00시)에 대한 상위 10개의 클러스터 확인 가능하지만, 당일에 대한 정보만 조회 가능 하다.



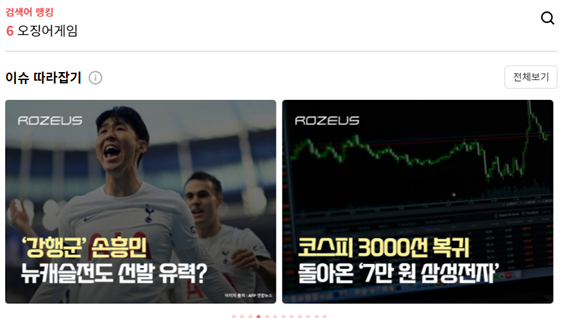
**[딥서치]**

사용자 지정 기간 설정과 이슈 정도에 따른 클러스터 결과에 대한 확인이 가능하지만 클러스터의 기사를 클릭하면 기사 원문 링크로 이동하기 때문에 사용자들이 가장 많이 뉴스를 확인하는 포털에 남긴 댓글이나 반응을 확인할 수 없어 실제 이슈 정도를 파악하는 데에 어려움이 있다.



**[로제우스]**

이슈 따라잡기 영역에서 최근 이슈에 대한 긍정, 부정기사를 모아 보여주기는 하지만 클러스터 목록 자체는 제공하지 않는다.



**6.1 서비스적 측면의 독창성 및 우수성**

본 프로젝트는 기존 포털사이트에서 제공하는 뉴스 페이지처럼 현재 날짜에 해당하는 이슈들을 랭킹순으로 보여주는 것에서 벗어나 사용자의 요청에 따른 기간에 대한 이슈랭킹을 보여주며 각 이슈에 대한 파생 주제의 변화 또한 확인할 수 있다.

또한 사람들이 가장 많이 이용하는 두 개의 포털 사이트에서 데이터를 수집하여 같은 뉴스에 대한 두 사이트의 댓글 감성 분석의 결과를 보여준다. 위에서 설명한 바와 같이 두 사이트 간 이용층의 성향이 다르다고 알려져 있기에 두 사이트를 모두 활용함으로써 동일 기사에 대한 다양한 반응들을 수집할 수 있어 보다 객관적인 댓글 감성 분석 전달이 가능하다. 이를 토대로 사용자는 클러스터링 된 주제에 대한 두 사이트 간의 반응 비교를 할 수 있다.

**6.2 기술적 측면의 독창성 및 우수성**

기존 실행되었던 뉴스 기사 분류와는 다르게, 정치, 경제와 같은 대분류 분할이 아닌 특정 사건에 대한 파생 주제들을 BERT와 STS데이터셋을 이용하여 분류를 시도하였다. 또한 대표적으로 알려진 K-means, K-medoids, H-clustering과는 다르게 분석가의 주관이 개입되지 않고, 가변 길이를 갖는 군집화 방법을 채택하여, 일련의 과정이 완전 자동화가 가능하다는 점에서 독창성과 우수성이 있다고 생각한다.

**7. 활용사례 결과물의 활용 범위 및 시장성**

**7.1 기대효과 및 활용 분야**

본 서비스를 통하여 평소 관심 갖지 않던 분야에 대한 인사이트와 정보를 얻을 수 있다.

누구나 포털 사이트 메인 페이지에 실린 뉴스 기사를 보던 중 특정 분야의 기사를 처음 접하게 되거나, 생소한 분야에 관심을 갖게 되는 상황이 발생할 수 있다. 이와 같은 경우 새로운 분야에 대한 이전 이슈의 흐름들을 파악하기에는 방대한 양일 것이다. 이에 사용자는 본 서비스를 통하여 정리된 정보를 제공받고, 빠르게 해당 분야에 대한 흐름을 파악할 수 있을 것이다.

이와 같은 장점을 토대로 평소 시사에 관심을 두지 않던 대학생, 취업준비생이라면 이 서비스를 활용하여 분야별 원하는 기간에 따른 이슈의 흐름을 파악하여 공모전, 프로젝트 등의 주제에 활용할 수 있고, 해당 분야의 상식을 쌓는 데에 도움을 얻어 면접 등에 활용할 수도 있다. 혹은 특정 프로젝트를 진행하게 되었을 경우를 비롯한 다양한 상황에서 이 서비스를 통해 인사이트를 얻는 데에 도움을 받을 수 있을 것이다.

**7.2 향후 발전 방안**

**7.2.1 서비스 측면의 향후 발전 방안**

제공하는 데이터에 대하여 현재 서비스 중인 기간보다 더 이전부터 최근까지의 뉴스기사를 수집해 사용자가 기간 설정을 함에 있어 더 다양한 기간을 설정해 이용할 수 있도록 서비스를 확대할 계획이다. 보다 더 긴 기간에 대한 서비스를 통해 사용자는 더 다양하고 많은 시사 이슈에 대한 정보를 제공받게 될 것이다.

현재 제공하는 뉴스 클러스터링 결과에서 키워드와 서브워드들을 보여주는 만큼 향후 키워드 검색 기능을 추가하여 검색한 키워드에 대한 이슈 랭킹을 제공할 수 있다. 사용자는 키워드 검색을 통해 어떤 사안에 궁금증이 생겼을 때 키워드를 검색하고 이에 대한 이슈의 흐름을 그래프와 클러스터 목록으로 한눈에 파악할 수 있을 것이다.

**7.2.2 기술적 측면의 향후 발전 방안**

BERT모델의 MLM모델을 이용한 분야 맞춤 fine tuning을 진행할 시, 성능이 향상된다는 연구 결과가 다수 존재한다. 본 프로젝트에서는 컴퓨팅 자원의 한계로 인하여 진행하지 못하였지만, 수집된 각 지면의 기사 내용을 바탕으로 MLM모델 튜닝을 진행하여 BERT Embedder에 해당 지면의 내용을 반영한 후, 본 프로젝트에 이용한다면 보다 더 정확하게 분류를 진행할 수 있을 것이라고 생각한다. 이와 더불어, Key word, Sub word를 추출함에 있어 동의어, 유의어에 대한 내용을 반영한다면 더더욱 정확한 분류가 가능할 것이라고 생각된다.

**8. 활용 데이터의 종류 및 특징, 확보 방법**

**8.1 News Clustering**

|  |  |  |
| --- | --- | --- |
| **Data Nam** | **KLUE - STS** | **KorNLU - STS** |
| **Label Format** | 0~5 (float) | 0~5 (float) |
| **Data Size** | 11,668(train) | 5,749(train)+1,500(Dev) |
| **출처** | KLUE-GIT | KorNLU-GIT |

**8.1.1 STS(Semantic textual similarity)데이터 선정 배경**

* 앞서 언급되었듯이 본 과제에서는 사건을 중심으로 기사들이 군집화 되기를 원하였기에, pre-trained된 BERT에 문장 간 의미의 유사도에 대한 데이터 셋인 STS 데이터셋으로 fine tuning을 진행하여, 보다 더 문장의 의미론적 요소를 반영한 embedded vector를 얻기 위해 STS 데이터를 Fine tuning Data로 선정하였다.

**8.1.2 KLUE – STS**

* 기존 라벨 값에 대한 의미를 기반으로 라벨을 재 분류 하였다. 문장 간 코사인 유사도가 0~1사이의 범위를 갖기에 기존 0~5의 범위(실수)로 정의 되어있는 라벨을 0(0~2)과 1(3~5)로 재분배 하였다. (Figure 2 참고)

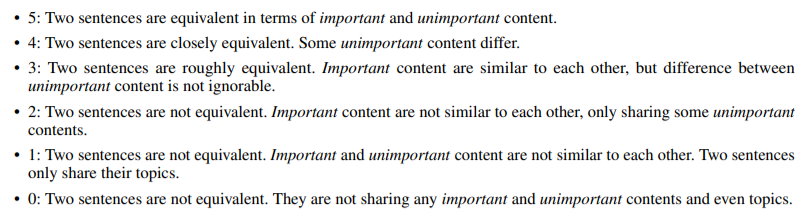


Figure 1) KLUE: Korean Language Understanding Evaluation (17p)

* 데이터의 경우 AIRBNB(구어체 리뷰), POLICY(뉴스), PARAKQC(benchmark data set)로부터 가져온 문장들로 구성되어 있으며, 뉴스데이터에 대한 내용이 포함 되어있다는 점에서 본 과제의 fine tuning data로 적합하다고 판단하였다.

**8.1.3 KorNLU – STS**

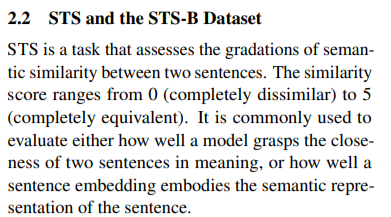
* 기존 라벨 값이 유사도를 기준으로 0~5 사이 실수의 형태로 부여되어 있어 기존 라벨의 값을 Cosine Similarity가 갖을 수 있는 범위(0~1)로 조정하였다. (Figure 3 참고)
* 데이터의 경우 이미지 캡션, 뉴스 헤드라인, 유저 포럼으로부터 수집된 영어 문장을 기계번역과 전문가 번역을 이용하여 한국어로 바꾼 문장들로 구성되어있다. 이 역시 KLUE의 데이터와 동일하게 뉴스 헤드라인이 포함 되어있다는 점에서 본 과제의 fine tuning data로 적합하다고 판단하였다.

Figure 2) KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding (2p)

**8.2 Sentiment Analysis**

|  |  |  |
| --- | --- | --- |
| **Data Nam** | **Naver Sentiment movie corpus V1.0** | **한국어 감정 정보가 포함된 단발성 데이터 대화 데이터 셋** |
| **Label Format** | 0,1 (긍부정 binary) | 기쁨, 슬픔, 놀람, 분노, 공포, 혐오, 중립 |
| **Data Size** | 150,000 (train) | 38,594 (All) |
| **출처** | e9t - GIT | AI-HUB (요청 데이터) |

**8.2.1 댓글 데이터 선정 배경**

* 댓글의 경우 최대한 많은 형식의 댓글에 대한 정보를 학습시키기 귀하여 하기의 두 데이터 셋을 선정하였다.

**8.2.2 Naver Sentiment movie corpus V1.0**

* 기존 라벨 값 들이 본 과제의 분류 기준인 긍정, 부정으로 분류되어있기에 별도의 변형 없이 이용하였다.
* 데이터의 경우, 2016년에 공개되었으며 네이버 영화에 달린 댓글들로 구성되어있다. 현재와 5년간의 차이가 존재하는 데이터이기에 현재 사용되는 구어체와의 부분적 차이(신조어 등)가 존재하지만 큰 차이가 존재하지는 않기에 감성 분류의 fine tuning data set으로 채택하였다.

**8.2.3 한국어 감정 정보가 포함된 단발성 데이터 대화 데이터 셋**

* 기존 라벨 값들이 7개의 감정(기쁨, 슬픔, 놀람, 분노, 공포, 혐오, 중립)으로 분류되어있기에, 본 과제의 분류 기준인 긍정(기쁨) 부정(혐오, 분노)에 맞게 데이터를 선별하여 이용하였다.
* 데이터의 경우 SNS 및 온라인 댓글에 대한 크롤링을 기반으로 구축되었다. 분류의 기준이 본 과제와 완전하게 부합하지 않지만, 상기 언급한 네이버 댓글 데이터 보다 최신의 데이터이기에(2019년) 신조어에 대한 훈련이 가능하며, 선택적 이용으로 본 과제에 적합한 데이터셋으로 변형이 가능하기에 감성 분류의 fine tuning data set으로 채택하였다.

**8.3 Service Data**

* 서비스에 사용되는 데이터의 경우, 2개의 포털 사이트에 대하여 크롤러를 구성한 후, 2021년 01월 01일부터 201년 08월 31일까지의 모든 기사를 각 지면별로 크롤링 하였다.

**8.3.1 Data 상세**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data type** | **ex** |
| **Date** | Date time | yyyy-mm-dd |
| **Title** | String | “[27일부터 전세대출 ‘잔금일 전’ 오른 만큼만 가능](https://news.naver.com/main/read.naver?mode=LSD&mid=shm&sid1=101&oid=374&aid=0000261055)” |
| **Context** | String | “기사 내용” |
| **Context\_len** | Integer | 300 |
| **Comments** | two-dimensional array | Missing Value: “NA”  Else: [[‘text’,좋아요,싫어요]. …] |
| **Comment\_cnt** | Integer | 400 |
| **Sentiment** | Dictionary | {"좋아요":0, "훈훈해요":0, "슬퍼요":0, "화나요":0, "후속기사 원해요":0} |
| **Link** | String | http://www... |
| **Press** | String | 언론사명 |

**8.3.2 Naver Daum 전처리**

크롤링 데이터에서 우선 Link 열을 기준으로 중복 행 중 가장 마지막 행만 남기도록 중복 기사를 제거하였다. 데이터 수집과정에서 중복되어 수집한 경우도 있고 실제 포털 사이트에 중복으로 존재하는 경우도 있었기에 Link를 통한 중복 기사 제거가 필요했다. 크롤링 과정 상 하루 안에서는 시간의 역순으로 데이터프레임에 저장되므로 중복 행 중 가장 먼저 나온 행을 남기기 위해 가장 마지막 행만 남기도록 했다.

또 실제 원문의 Link는 다르지만 Date, Title, Context, Press 가 모두 일치하는 경우 같은 기사로 판단했다. Context까지 비교하므로 위의 네 열의 값이 일치하면 실제 같은 기사임을 확인할 수 있었다. 네 열이 모두 같은 값을 가지는 행에 대하여 Comments와 Sentiment를 합하였다. 기사에 대한 사용자의 반응들을 더 정확히 수집하기 위해 진행한 과정이다.

Title\_clean 컬럼을 추가하여 Title에 포함된 특수문자를 제거한 Title을 따로 저장하였다. 전처리 진행 후 두 포털 사이트의 크롤링한 데이터를 join과정에서 Title열의 값을 사용하는데 두 포털에서 사용된 특수문자가 일치하지 않아 join의 결과가 제대로 나오지 않아 이 과정을 추가로 하게 되었다.

Context\_len이 300보다 작은 경우 대부분 내용이 없는 포토뉴스에 불과하여 join과정에 사용하지 않기 위해 두 포털의 크롤링 데이터에서 각각 Context\_len 이 300보다 작은 행은 제거하였다.

**8.3.3 Naver Daum join**

본 서비스에서 클러스터 기사에 대한 두 포털의 반응을 비교하여 보여주기 때문에 두 포털에서 수집한 데이터에 모두 존재하는 기사에 대한 데이터만 최종 데이터로 사용하기 위해 두 포털에 대한 데이터를 inner join하여 이용하였다.