자연언어처리 프로젝트 기획안

디지털금융공학 석사과정

2019516004 조상훈

**1. 프로젝트 명: LDA(Latent Dirichlet Allocation)기반 토픽 모델링을 활용한**

**감정 분석기 개발**

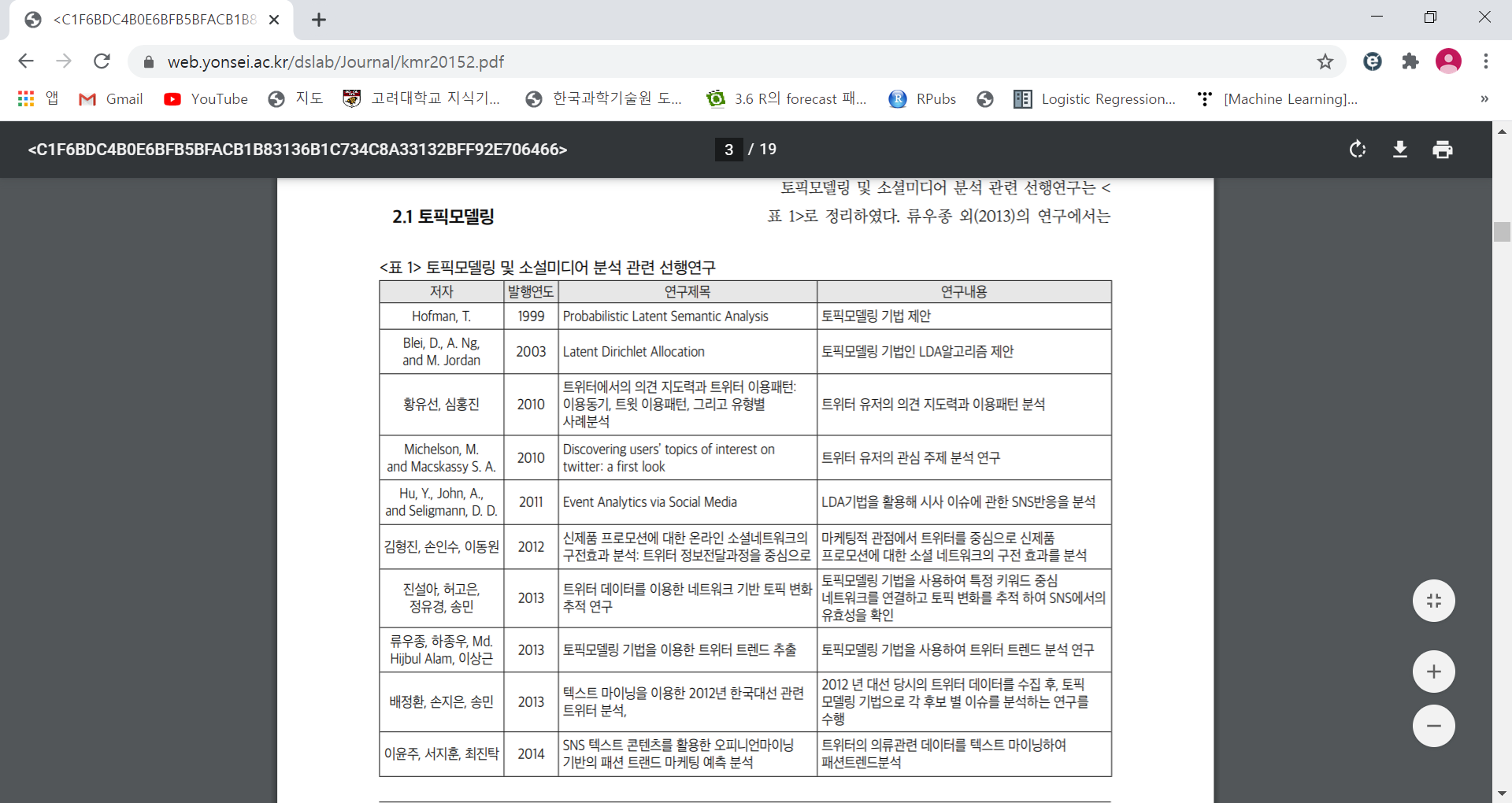
**2. 프로젝트 소개 (필요성, 기존 연구, 배경 지식 등)**

○ **토픽 모델링(Topic modeling)**은 레이블이 없는 텍스트 문서에 토픽을 할당하는 광범위한 분야이다. 예를 들어 대량의 뉴스기사에서 데이터셋을 분류하는 일등을 들 수 있다. 만약 기사데이터에 어떤 페이지나 카테고리에 게재되었는지 정보가 없다면 토픽모델링을 통해 스포츠, 금융, 세계뉴스, 정치 등 카테고리 레이블을 할당하여 분류할 수 있다. 머신 러닝 기초에서 생각해보면 비지도 학습(Unsupervised Learning)의 한 분야인 클러스터링(Clustering)과 비슷하게 생각 할 수 있다.

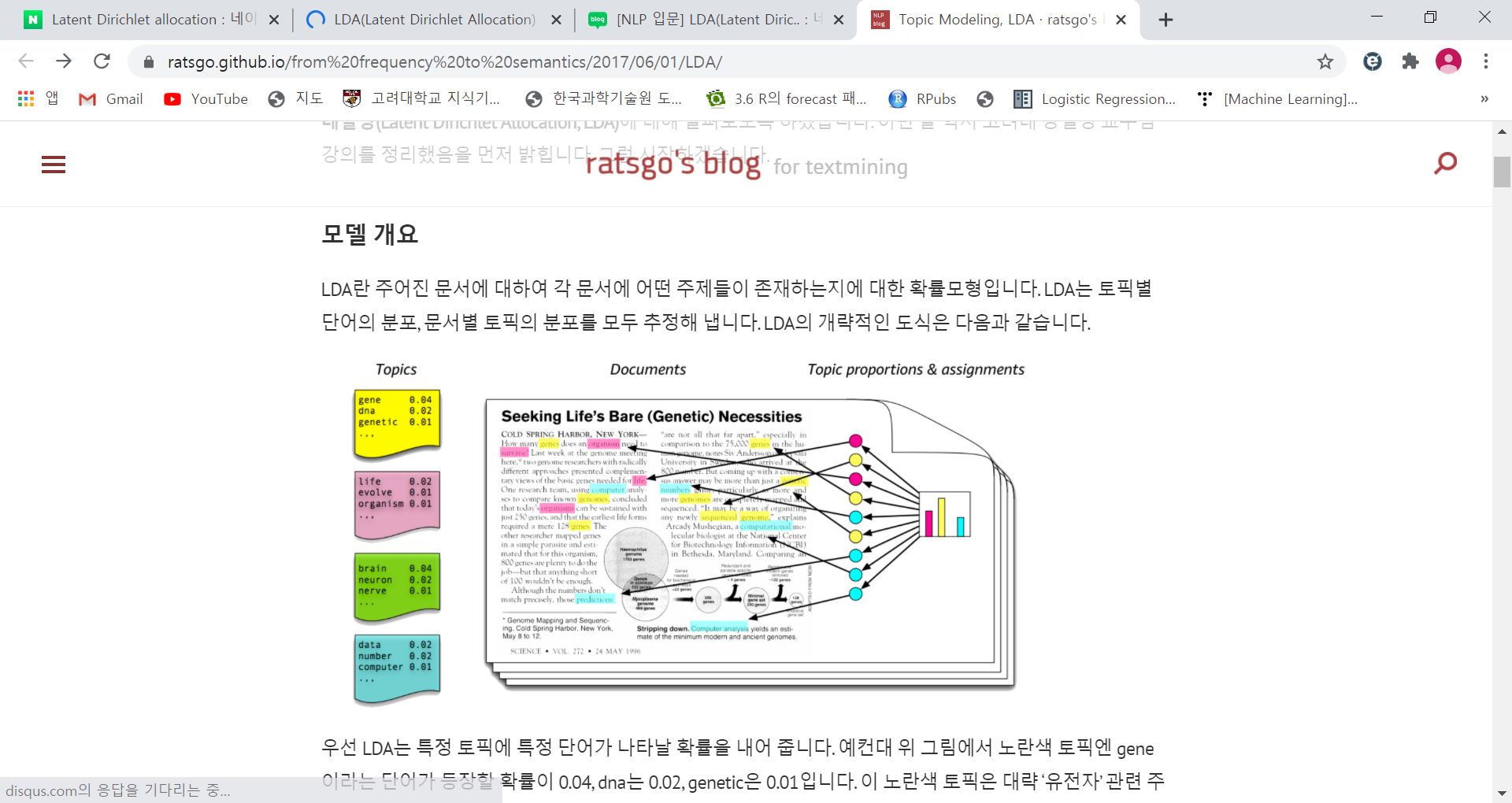
○ 이러한 **감정분석의 필요성**은 기본적으로 주어진 NSMC/Friends 데이터셋으로 분류되는 긍정/부정 또는 7가지 감정과 더불어 대화에서 화자가 전달하고자 하는 의도 및 주제를 같이 파악할 수 있다면 감성분석을 비즈니스에 활용하는 모델로 고려하기에 더욱 적합할 수 있겠다는 가정에서 출발하게 되었다.

이미 여러 논문 등을 통해 감정분석과 토픽모델링을 결합하여 기업에서 제품개발시에 고객관점에서 제품에 대한 주요 토픽을 정의하고, 개별 토픽에 대한 기회분석을 정량화 하며, 따라서 변화하는 고객 니즈의 모니터링과 제품개발 컨셉 도출에 활용하는 연구를 활발히 진행 중이다. <표1>참조

**이를 금융업에 적용하여 고객 상담 분석을 통한 상품 추천 및 고객 민원 감정분석 모니터링 등에 활용하면 은행에서 가장 중요시하고 예민한 CS(Customer Service)관점의 문제해결에 도움을 줄 수 있을 것이다.**



**○ LDA 토픽 모델링이란?**



이번 프로젝트에서 적용하고자 하는 LDA(Latent Dirichlet Allocation) Topic Modeling은 여러 문장에 걸쳐 자주 등장하는 단어의 그룹을 찾는 확률적 생성 모델이다. 각 문서를 여러 단어가 혼합된 것으로 가정하면 토픽은 자주 등장하는 단어들로 나타낼 수 있다. LDA는 일종의 BOW(Bag-of-Word)모델로 입력 받은 BOW행렬을 문서-토픽 행렬, 단어-토픽 행렬 두 개의 행렬로 분해한다.

이 두 행렬을 곱해서 가능한 작은 오차로 BOW입력 행렬을 재구성할 수 있도록 LDA가 BOW행렬을 분해한다. LDA가 BOW행렬에서 찾은 토픽이 즉 관심 대상이다. 단점은 토픽의 개수를 LDA의 Hyperparameter로 미리 수동으로 지정해야 한다는 점이다.

**3. 프로젝트 개요 (기존 연구와의 차별점, 본인만의 독창적인 아이디어(선택사항), 전체 개요도)**

○ **기존 연구와의 차별점**

1) NSMC 기반의 한글 감정 분석기

▷As-is: 긍정1/부정0 이 정해져 있고, 텍스트가 들어왔을 때, 1/0 분류해 주는 것이 텍스트 안에 있는 감정을 긍정 부정으로 분류해주는 것과 동일하다.

▷To-be: LDA를 이용하여 단순히 긍정 부정이 아닌 토픽모델링을 추가 적용하여 영화리뷰를 여러가지 카테고리(영화장르, 주연배우 등)로 분류하여 세부 주제별로 감정분석 가능.

2) Friends 데이터 기반의 영어 감정분석기

▷As-is: 영어 대화 안의 감정을 7가지(neutral, joy, sadness, fear 등)로 분류하

여 다자간 대화의 발화 시 감정 분석에 이용한다.

▷To-be: LDA토픽모델링을 이용하여 문장안에 자주 등장하는 단어의 중요도에

따른 감정 및 상황을 토픽으로 분류

**[전체 개요도]**

**-1단계**: 데이터 전처리 -> 단어 태깅 -> 양태분석 -> 관계추출 -> 데이터요약

**-2단계**: 감정분석 모델 훈련/검증 Traing(75%) – Testing(25%)

**[NSMC]**

TABLE: id(작성자ID). Document(리뷰내용), label (0 negative, 1 positive)

File: rating\_train.txt(15만개), rating\_test.txt(5만개)

**[Friends]**

TABLE: Role, Utterance, Emotion(7가지) File: Train, Dev, Test Source

**-3단계**: 토픽모델링을 활용한 주제 분류 및 성능 검증(Perplexity, Topic Coherence)

4. **프로젝트 세부계획 (세부 모델, 세부 파라미터, 세부 알고리즘 구체적으로 서술)**

○ 학습 데이터: 네이버 영화리뷰 데이터(NSMC), Emotion lines Friends데이터

○ **모델구축 계획**

: BERT 등을 이용한 감정분석, Python Gensim 패키지를 이용한 LDA모델 생성

○ **개발 환경**: 구글 Colaboratory Jupyter notebook(GPU) 활용

○ **데이터 전처리 방법**

: 파이썬 KoNLPy 패키지 Kkma, Komoran, Mecab, 카카오 Khaiii 형태소 분석기

LSTM Sentiment Analysis 코드를 통한 이모티콘 감정분석 후 전처리 여부 결정

○ **학습모델 선정:** 가장 유명한 Hugging Face의 PyTorch BERT를 주로 사용

로지스틱회귀 모델도 학습하여 성능비교예정 (속도 향상을 위해 외부메모리 학습 기법을 이용하여 SGDclassfier의 partial\_fit메서드 이용)

○ **토픽 모델링 적용 및 성능 평가**

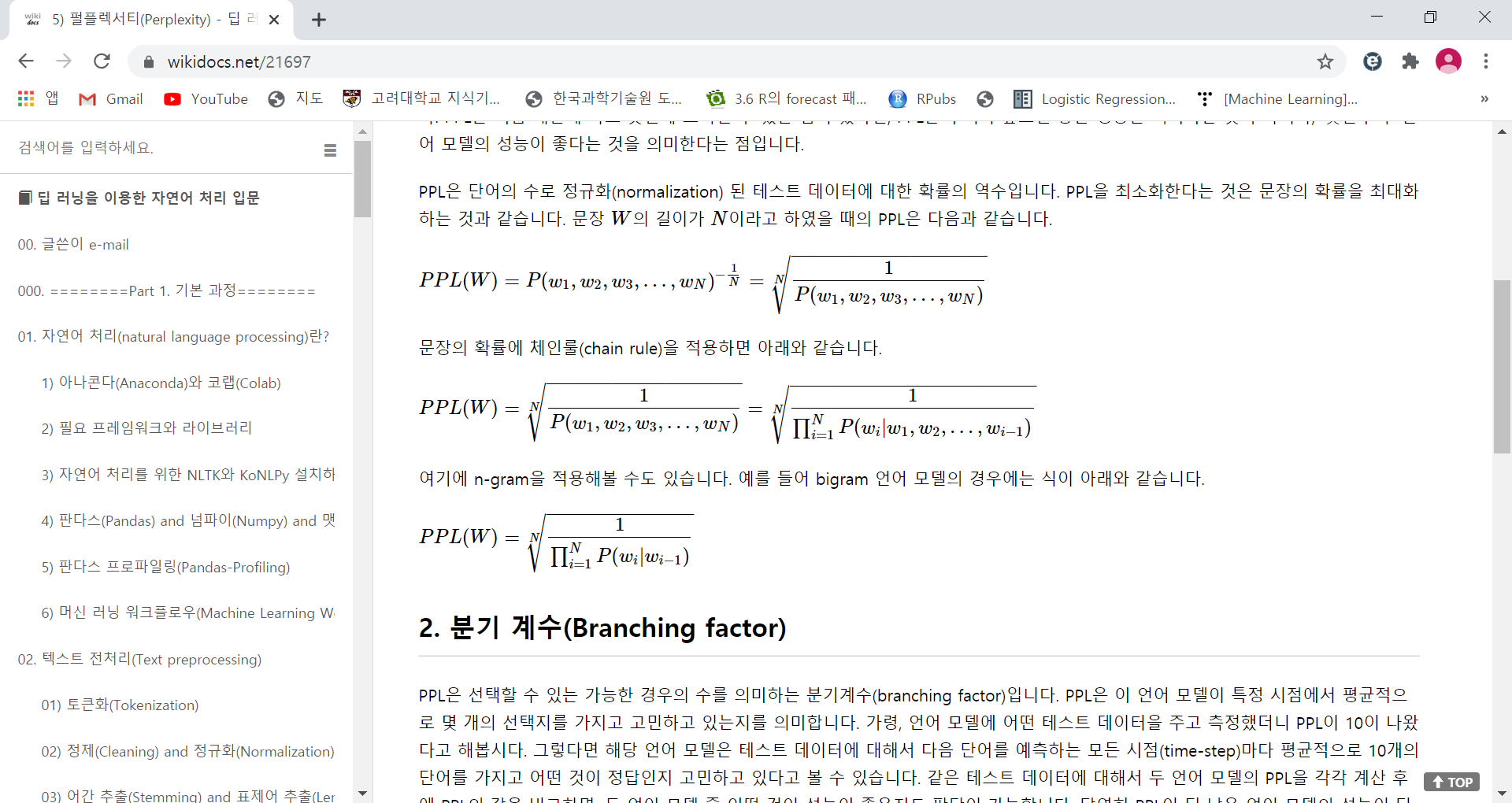
NSMC/Friends데이터 중 80%는 Training-data, 20%는 Testing-data로 사용하고

사이킷런의 CountVectorizer, LatentDirichletAllocation 클래스 활용하여 토픽

모델링을 적용한다.

제작된 LDA모델 성능 평가 지표는 Perplexity, Topic Coherence를 이용 측정함.

- Perplexity: 확률 모델이 어떤 값을 얼마나 잘 예측하는지 평가하기 위한 내부 평가 지표. (보통 줄여서 PPL로표시) 수치가 '낮을수록' 언어 모델의 성능이 좋다는 것을 의미함.



-Topic Coherence: 토픽이 얼마나 의미론적으로 일관성 있는지 판단하는 지표로

해당 모델이 얼마나 실제로 의미 있는 결과를 내는지 확인하기

위해 사용함. 토픽모델링 결과로 나온 주제에서 상위 N개 단어

를 뽑아 유사도를 계산. 값이 높을수록 일관성 있는 키워드로

결과가 나와 모델의 성능이 높다는 것을 의미함

**5. 프로젝트 추진일정**

○ 11월 1주: NSMC/Friends 데이터 전 처리 (Khaiii, re, nltk 패키지 등 활용)

○ 11월 2주: Logistic Regression, BERT를 이용한 FRIENDS 훈련 및 테스트

○ 11월 3주: Logistic Regression모델 BERT, KOBERT 등을 이용한 NSMC 훈련 및

테스트

○ 11월 4주~12월 1주: LDA를 이용한 데이터셋 분해 및 토픽 분류

(사이킷런 CountVectorizer, LatentDirichletAllocation 클래스 활용)

○ 12월 2주: 학습결과 확인 및 결론 도출 후 최종 리포트 작성

6. **프로젝트 예상결과**

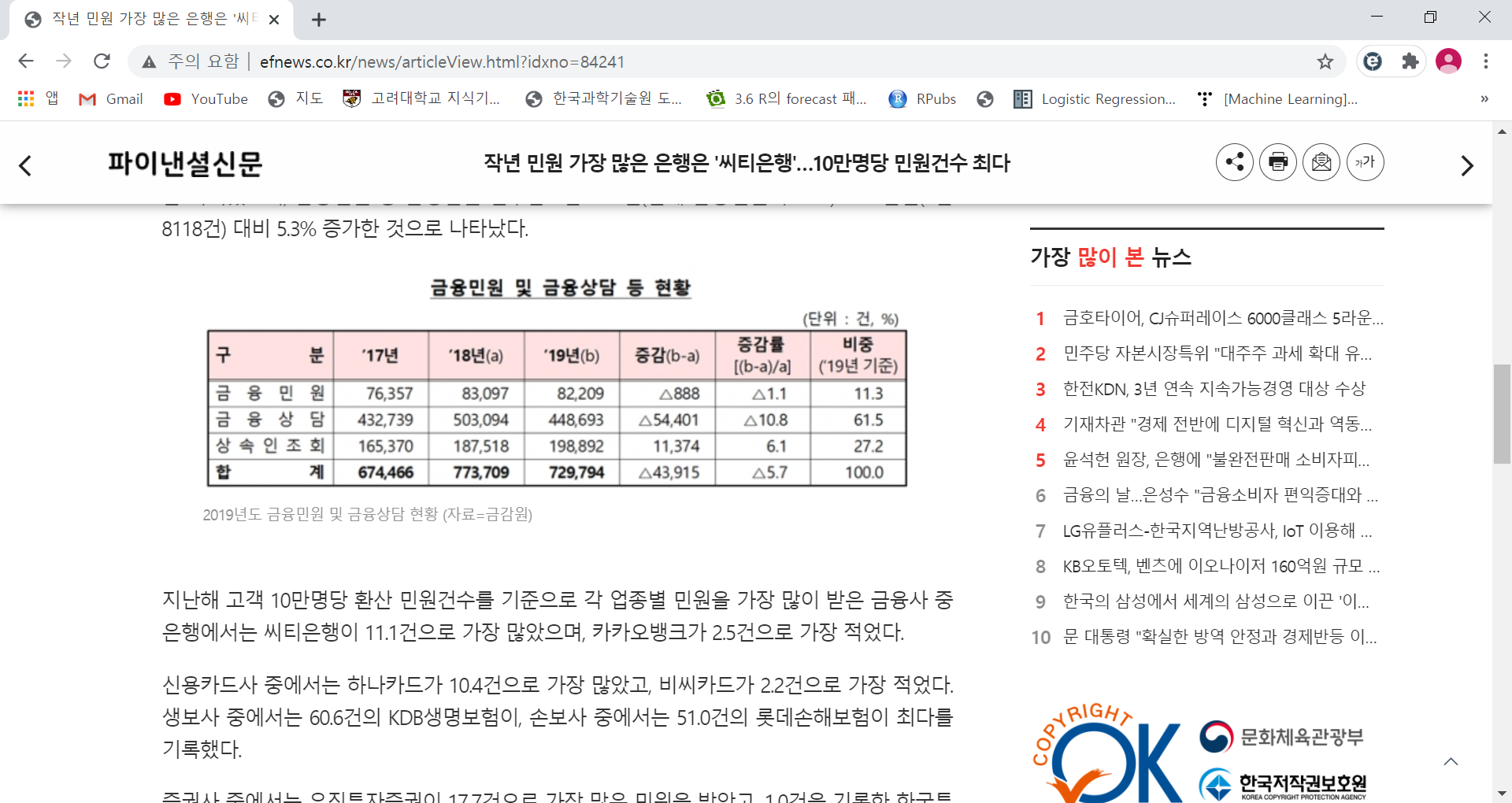
○ 로지스틱 회귀모델(싸이킷런 SGDClassifier의 Partial\_fit메서드 활용한),

PyTorch BERT, KOBERT 사용시 88%이상(기존 분석 참고)의 정확도 예상됨.

○ LDA모델링시 Hyperparameter 조정을 통해 토픽의 개수 및 확률 실험필요

○ 향후 현업 적용 가능성

: 2019년 한 해 금융권 민원은 총 72만건에 달한다. 코로나이후 2020년 상반기 민원은 은행만 6,107건으로 작년 상반기보다 30%이상 증가하였다.



앞으로도 경기침체와 고객들의 금융지식 심화로 민원은 더욱 증가할 예정이며 대책이 필요한 상황이다. 따라서 고객 상담 음성분석을 통한 마케팅 및 최적의 상담 멘트 선정 등 고객 상담데이터의 자연어 처리를 통한 잠재 메시지 파악을 위한 분석이 필요할 것이다. 본 프로젝트와 같이 감정분석 및 토픽모델링을 통해 고객의 감정 및 발화 의도를 세부 카테고리별로 분류해서 빠르게 응대할 수 있다면 고객민원처리 만족도가 크게 향상될 것이라고 예측된다.

**[참고문헌]**

Latent Dirichlet Allocation, David M. Blei, Andrew Y.Ng, and Michel I. Jordan, Journal of Machine Learning Research 3page : 993-1022, Jan 2003

Efficient Estimation of Word Representations in Vector Space, T.Mikolov, K.Chen, G.Corrado, and J.Dean, arXiv preprint arXiv:1301,3781, 2013

Topic Modeling, LDA(고려대학교 강필성 교수님 강의 정리) ratsgo’s Blog for textmining 2017

<https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/>

Latent Dirichlet Allocation: Towards a Deeper Understanding, <http://obphio.us/pdfs/Ida_tutorial.pdf>

Khaiii, https://github.com/kakao/khaiii, May 2019

이유진, 김세빈, 홍현석, 김장원 특허 문서를 위한 형태소 분석기 비교평가. 한국정보기술학회 종합학술발표논문집, 2019년 6월 호, 264-265. 2019

<https://www.sentv.co.kr/news/view/579490> “코로나에 은행 민원 30% 급증…“상환 유예” 서울경제

금융**민원** 주요 **통계 “금감원” www.fss.or.kr**