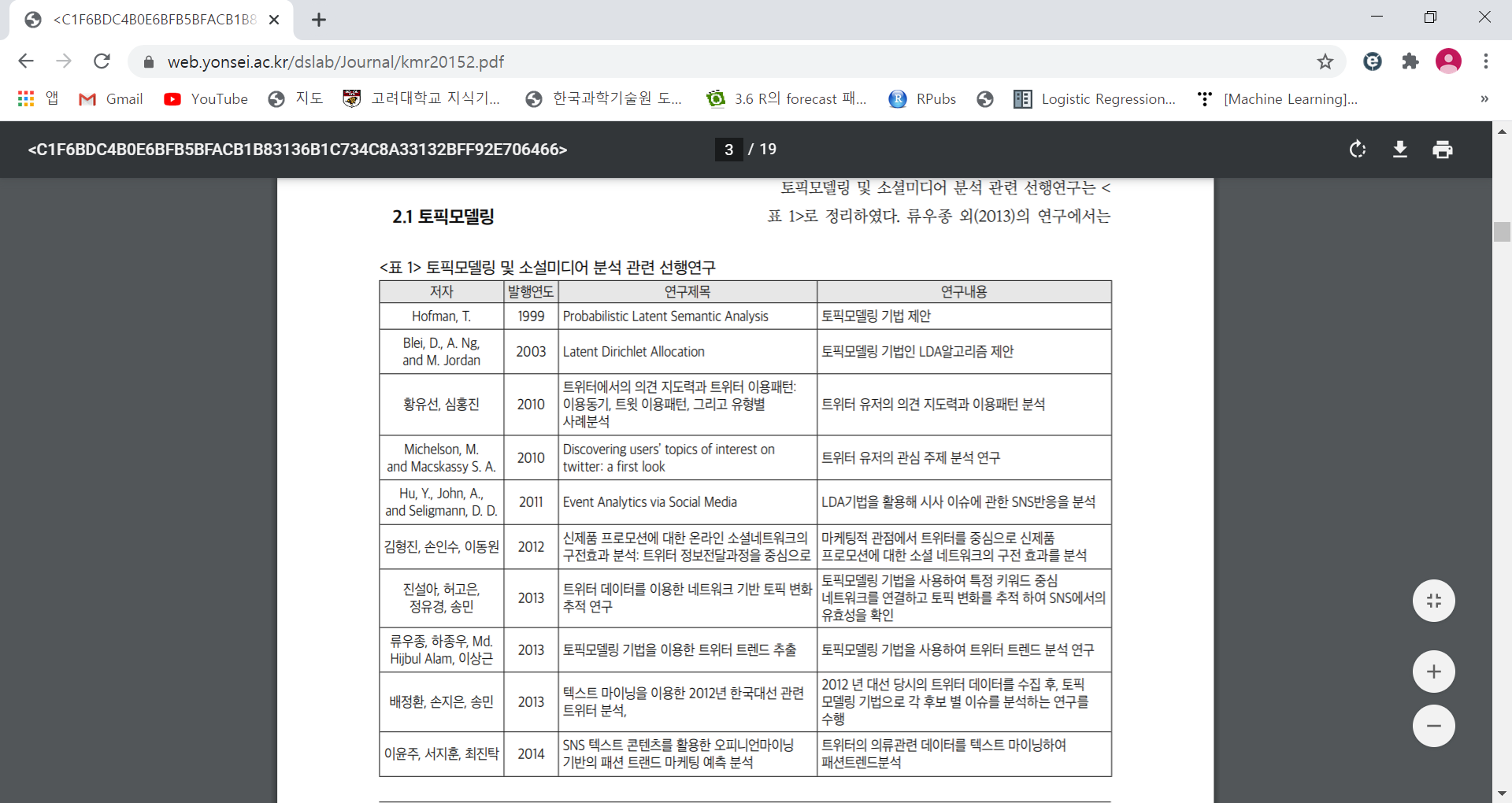
|  |
| --- |
| **LDA(Latent Dirichlet Allocation)기반 토픽 모델링을 활용한 감정 분석기 개발**  조상훈  고려대학교, 디지털금융공학과  csh301@korea.ac.krr  **Development of Sentiment Analysis with LDA Topic Modeling**  Sanghun Cho  Korea University, Digital Finance Engineering  **요 약**  LDA Topic Modeling은 여러 문장에 걸쳐 자주 등장하는 단어의 그룹을 찾는 확률적 생성 모델로 단어나 문서의 숨겨진 주제를 찾아내어 키워드별로 주제끼리 묶어주는 비지도 학습 알고리즘이다. 이 연구는 토픽모델링이 기본적으로 주어진 데이터셋으로 긍정/부정을 예측할 뿐 아니라 대화에서 화자가 전달하고자 하는 의도 및 주제를 같이 파악할 수 있어 감성분석을 비즈니스에 활용하는 모델로 고려하기에 더욱 적합할 수 있겠다는 가정에서 출발하게 되었다. BERT, ELETRA등 최신 언어분석 모델을 통한 빠르고 효율적인 감정분석 기술에 토픽모델링을 더하여 고객의 감정 및 발화 의도를 세부 카테고리별로 분류하고 감정 어휘 사전을 정교화 하여 CS관련 응대를 할 수 있다면 금융업 민원처리 방식의 큰 변화가 예상된다.  주제어: LDA(Latent Dirichlet Allocation)토픽 모델링, BERT ELECTRA 감정분석, 금융 민원 |

**1. 서론**

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

이러한 토픽모델링의 필요성은 기본적으로 주어진 데이터셋으로 긍정/부정을 예측할 뿐 아니라 대화에서 화자가 전달하고자 하는 의도 및 주제를 같이 파악할 수 있다면 감성분석을 비즈니스에 활용하는 모델로 고려하기에 더욱 적합할 수 있겠다는 가정에서 출발하게 되었다.

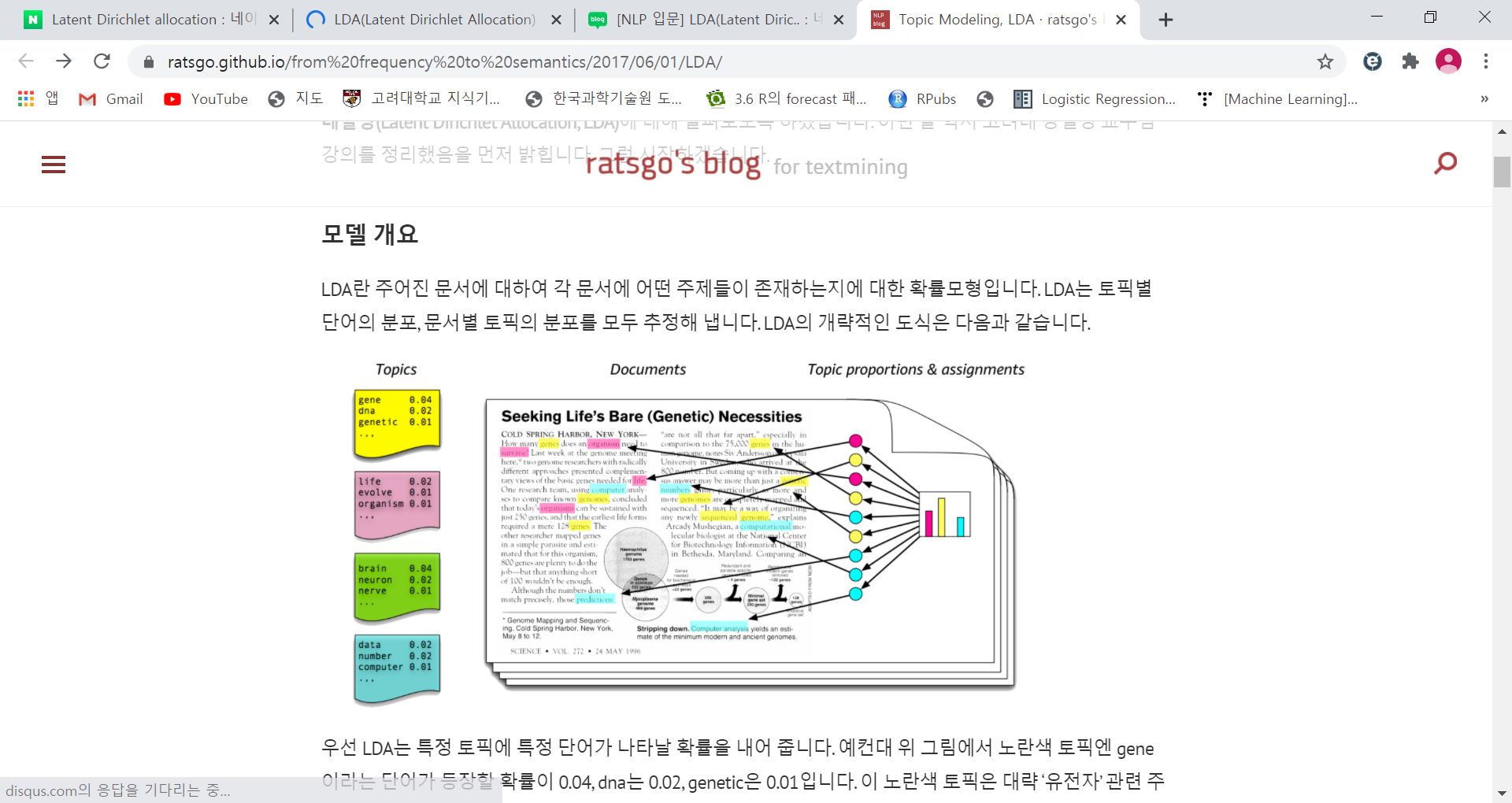
이미 여러 선행연구 등을 통해 감정분석과 토픽모델링을 결합하여 기업에서 제품개발시에 고객관점에서 제품에 대한 주요 토픽을 정의하고, 개별 토픽에 대한 기회분석을 정량화 하며, 따라서 변화하는 고객 니즈의 모니터링과 제품개발 컨셉 도출에 활용하는 연구를 활발히 진행 중이다.



이를 금융업에 적용하여 고객 상담 분석을 통한 상품 추천 및 고객 민원 감정분석 모니터링 등에 활용하면 금융산업에서 가장 중요한 관점 중에 하나인 CS(Customer Service)관련 고객응대의 빠른 문제해결에 도움을 줄 수 있을 것이라는 전제에서 연구를 시작하게 되었다.

**2. LDA토픽모델링**

**2.1 LDA 토픽 모델링이란?**



**<그림 1> LDA 모델링 개념도식**

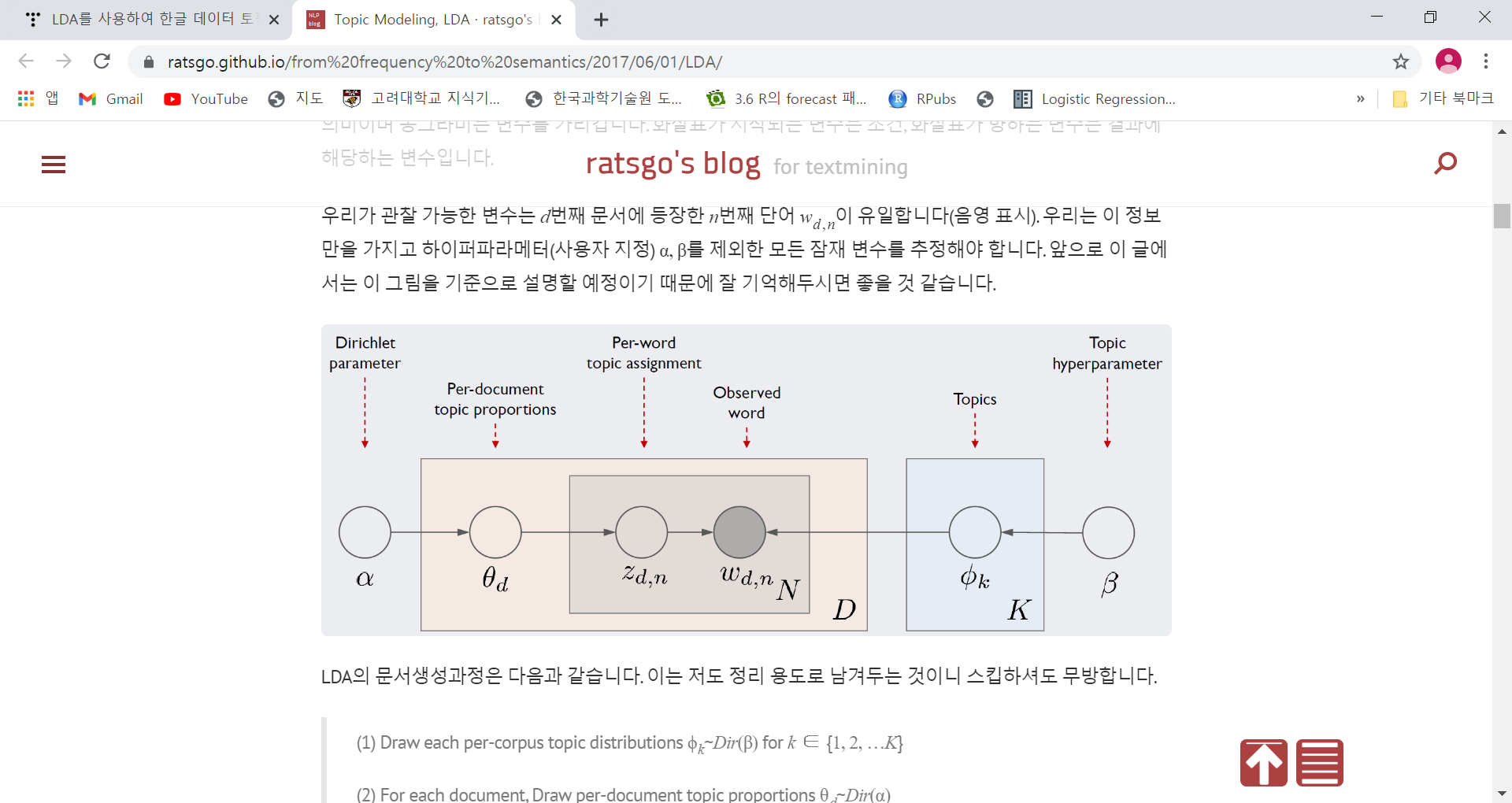
이번 프로젝트에서 적용하고자 하는 LDA(Latent Dirichlet Allocation) Topic Modeling은 여러 문장에 걸쳐 자주 등장하는 단어의 그룹을 찾는 확률적 생성 모델로 단어나 문서의 숨겨진 주제(Topic)을 찾아내어 문서와 키워드별로 주제끼리 묶어주는 비지도 학습 알고리즘이다.

추출한 원문에는 다양한 내용이 담겨있을 수 있고 이러한 주제들을 일일이 수기로 분류하기 어렵기 때문에 LDA 분류방법을 통해 전반적인 데이터 구조를 먼저 파악하는 것이 중요하다. LDA 토픽모델링은 단순히 주제분류 뿐만 아니라 주제에 포함되는 키워드를 보여주고 그 키워드로 해당주제를 해석하고 정의할 수 있다.

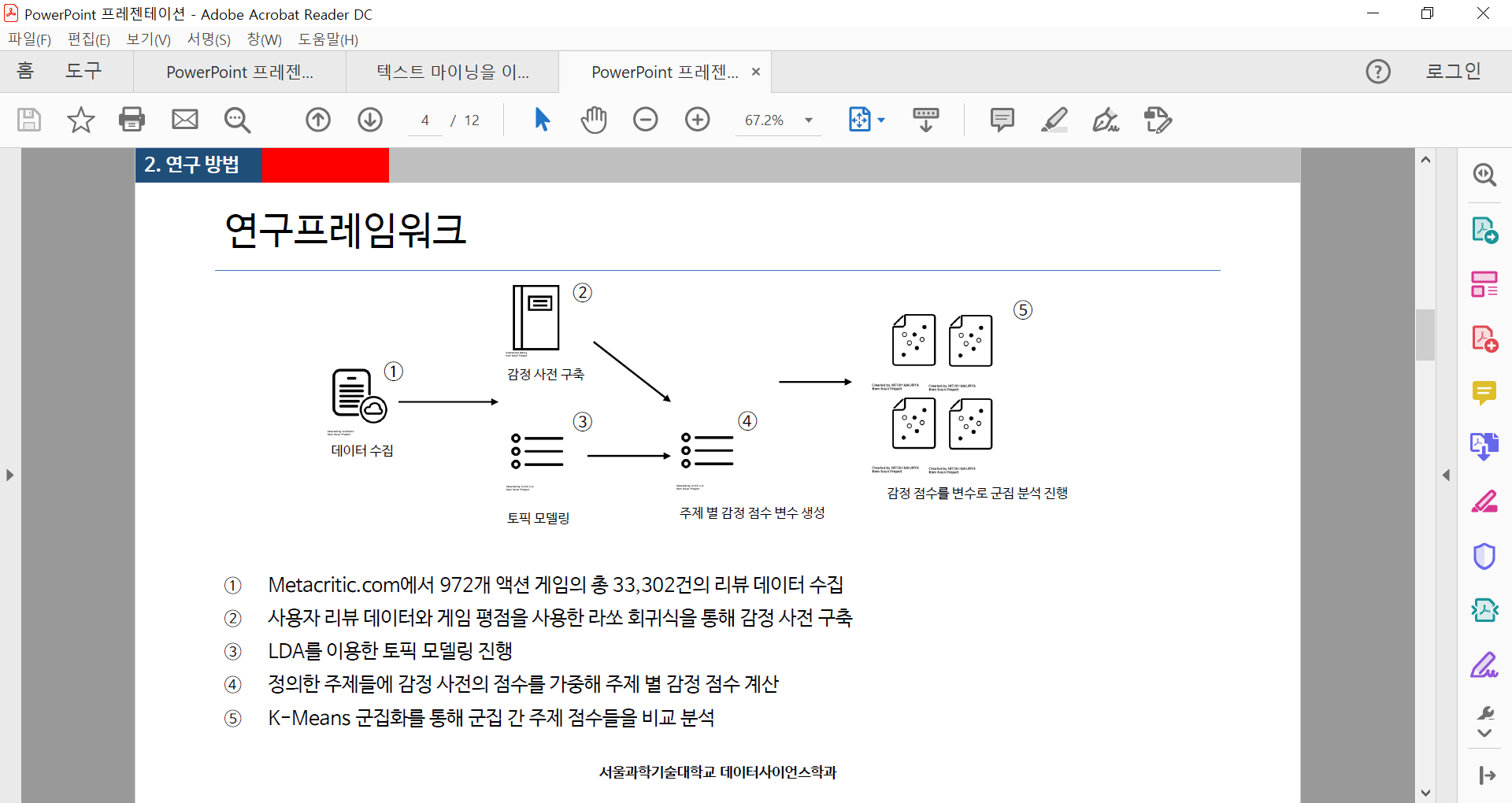
각 문서를 여러 단어가 혼합된 것으로 가정하면 토픽은 자주 등장하는 단어들로 나타낼 수 있다. LDA는 일종의 BOW(Bag-of-Word)모델로 입력 받은 BOW행렬을 문서-토픽 행렬, 단어-토픽 행렬 두 개의 행렬로 분해한다.

이 두 행렬을 곱해서 가능한 작은 오차로 BOW입력 행렬을 재구성할 수 있도록 LDA가 BOW행렬을 분해한다. LDA가 BOW행렬에서 찾은 토픽이 즉 관심 대상이다.

단점은 Topic의 수를 LDA의 Hyperparameter로 미리 수동으로 지정해야 한다는 점이나 모델링을 통한 예측을 통해 토픽의 최적 값을 찾아 대입할 수 있고 평가를 통해 정확도 높은 적정 Topic수를 찾는 것이 가능하다.

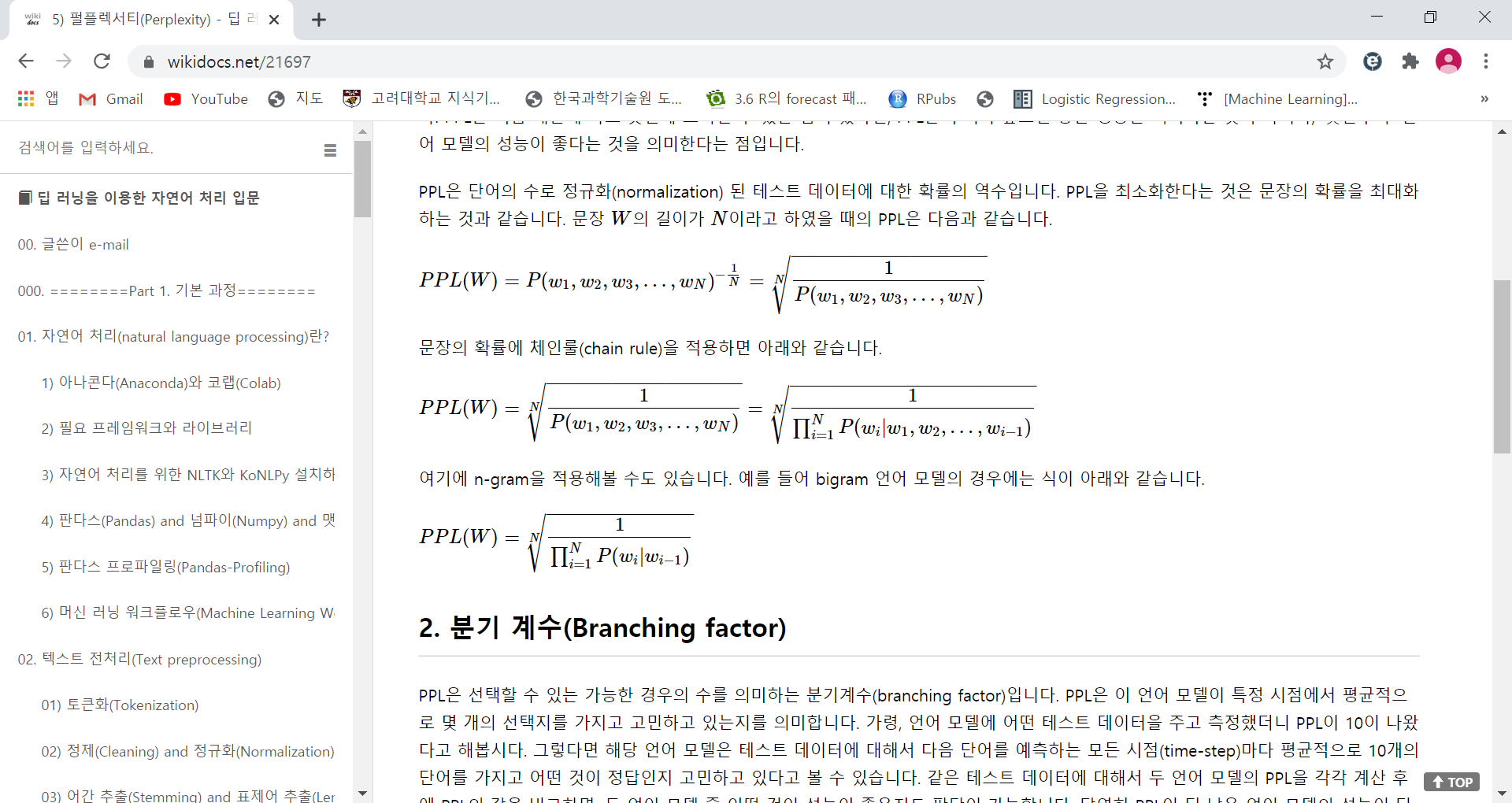


**<그림 2> LDA 모델 아키텍쳐**

**2.2 프레임워크**

**2.3 토픽 모델링 적용 및 성능 평가**

LDA모델 성능 평가 지표는 **Perplexity,Topic Coherence**를 이용 측정함.

1. **Perplexity**: 확률 모델이 어떤 값을 얼마나 잘 예측하는지 평가하기 위한 내부 평가 지표. (보통 줄여서 PPL로표시) 수치가 '낮을수록' 언어 모델의 성능이 좋다는 것을 의미함.
2. **Topic Coherence**: 토픽이 얼마나 의미론적으로 일관성 있는지 판단하는 지표로 해당 모델이 얼마나 실제로 의미 있는 결과를 내는지 확인하기위해 사용함. 토픽모델링 결과로 나온 주제에서 상위 N개 단어를 뽑아 유사도를 계산. 값이 높을수록 일관성 있는 키워드로 결과가 나와 모델의 성능이 높다는 것을 의미함

**3. BERT, ELECTRA 모델을 통한 감정분석 연구**

**3.1 BERT모델의 특징**

BERT는 Sebastian Ruder가 언급한 NLP’s ImageNet에 해당하는 가장 최신 모델 중 하나로, 대형 코퍼스에서 Unsupervised Learning으로 General-Purpose Language Understanding 모델을 구축하고pre-training Supervised Learning으로 Fine-tuning 해서 QA, STS등의 하위downstream NLP 태스크에 적용하는 Semi-supervised Learning 모델이다. ULMFiT이 가능성을 보여주었고, 이후 ELMo, OpenAI GPT등이 놀라운 성능을 보여주면서 그 진가를 인정받았다. BERT의 경우 무려 11개의 NLP 태스크에서 state-of-the-art를 기록하면서 뉴욕 타임즈의 지면을 장식하기도 했다

다만 많은 계산양을 요구하고 문맥정보를 활용하기 어려운 단 점이 있어 그 문제를 해결하기 위한 방법으로 pre-train을 통한 MLM(Masked Language Model), NSP(Next Sentence Prediction)방식이 제시되고 있다.

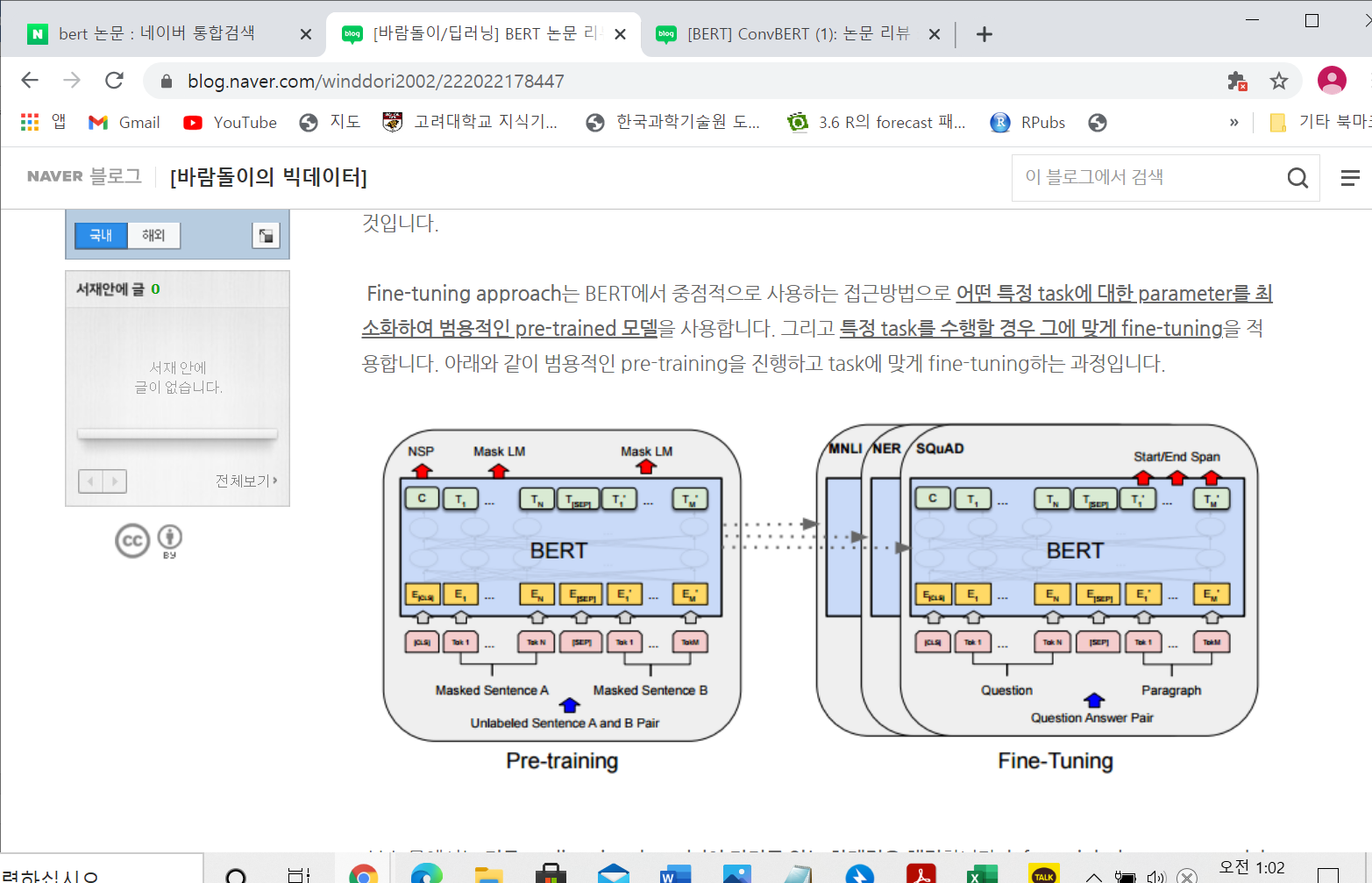
**3.1.1 Pre-trained BERT**

Pre-trained된 모델들은 성능측정에 매우 효율적이다.

Pre-training은 Feature-based, Fine-tuning 두가지 접근 방법으로 나뉘며 Feature-based approach는 대표적으로 ELMO가 있다. Feature-based의 핵심은 어떠한 특정 task를 해결하기 위한 architecture를 구성하며 pre-trained representations를 부가적인 feature로 사용하는 것입니다.

Fine-tuning approach는 BERT에서 중점적으로 사용하는 접근방법으로 어떤 특정 task에 대한 parameter를 최소화하여 범용적인 pre-trained 모델을 사용한다. 그리고 특정 task를 수행할 경우 그에 맞게 fine-tuning을 적용한다.

일반적으로 범용적인 pre-training을 진행하고 task에 맞게 fine-tuning하는 과정을 거친다.

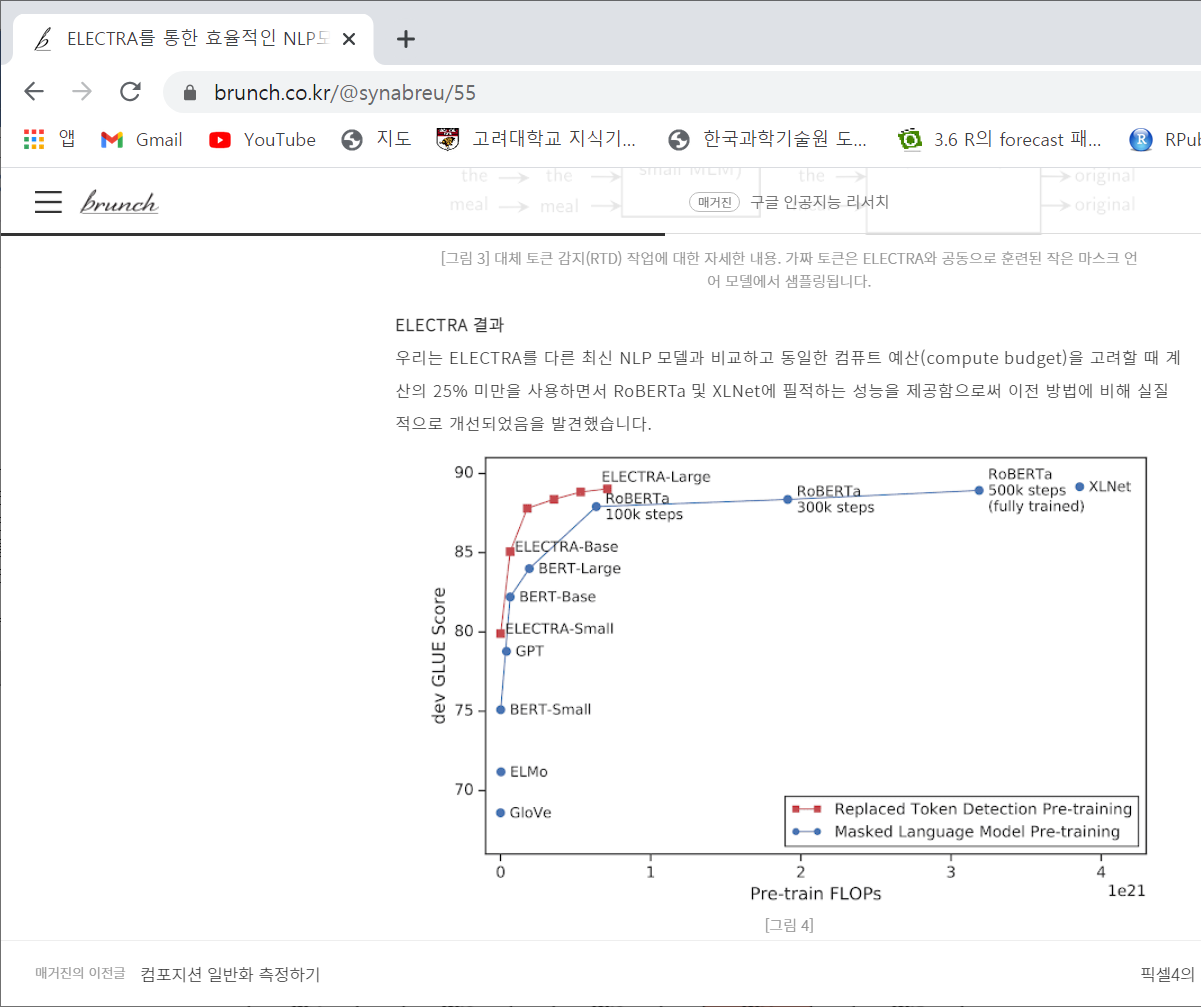


**<그림 3> BERT Pre-training to Fine-tuning 도식도**

**3.2 ELECTRA모델의 특징**

BERT의 성능과 효율을 개선하기 위해 다른 발전된 모델들이 제시되고 있고 그 중에 하나가 ELECTRA이다.

ELECTRA모델은 BERT의 이점을 제공하지만 훨씬 더 효율적으로 배우는 언어사전훈련에 대한 다른 접근방식을 취한다. 토큰교체를 정확하게 분류하는 인코더를 효율적으로 학습하는 방법으로 동일한 컴퓨팅 예산으로 BERT를 능가하는 효율을 보여줄 수 있다. ELECTRA는 최근 논문[1]에 의하면 계산량의 1/4미만을 사용하더라도 NLP 벤치 마크에서 ROBERTa 및 XLNet의 성능과 일치시킬 수 있다고 한다.



**<그림 4> 각 언어모델 별 성능 비교**

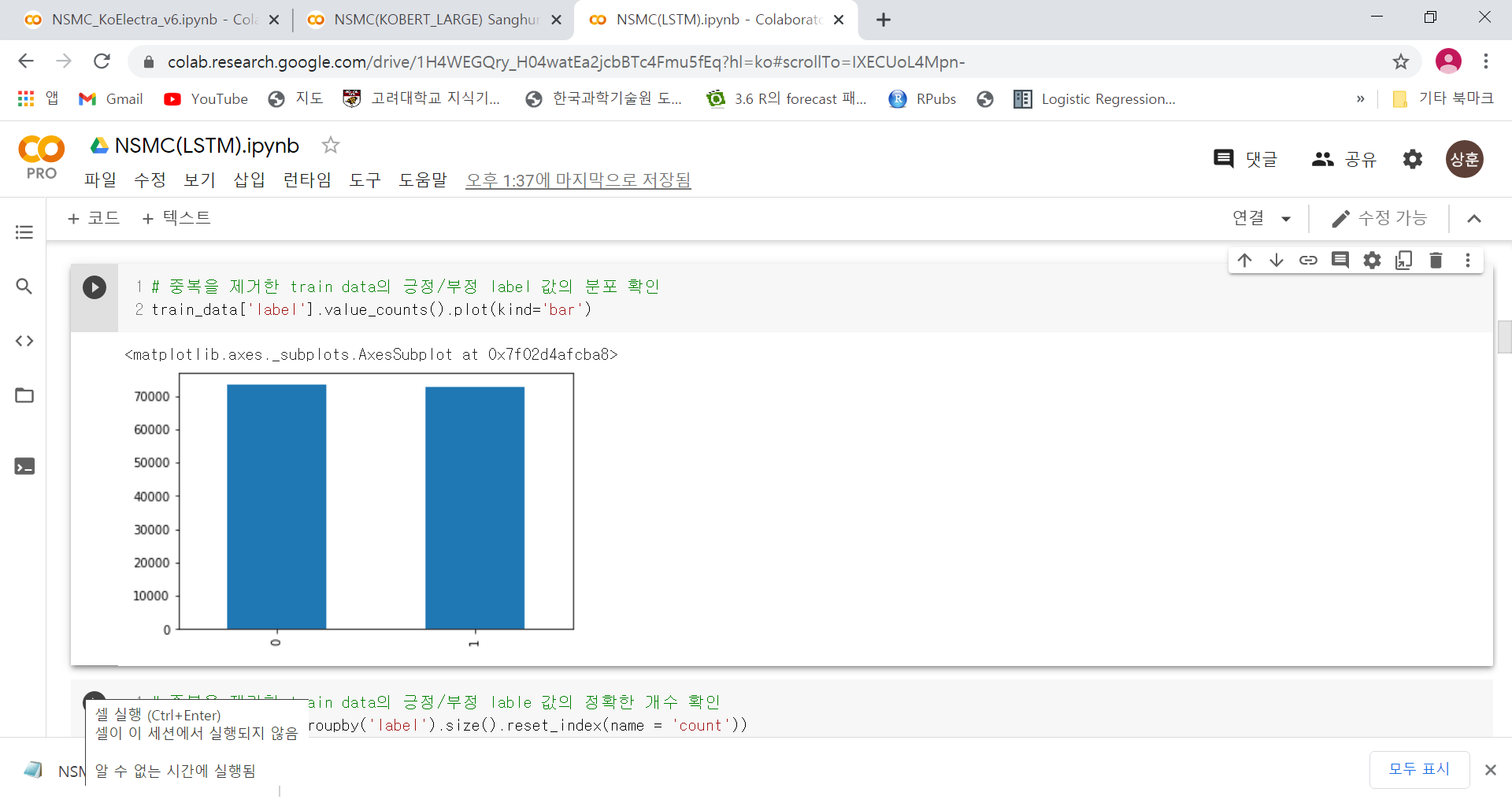
기존의 BERT가 MLM Pretrain Methods로 Masking 된 토큰만 예측하면 되지만 ELECTRA는 모든 input token에 대해 판별을 하기 때문에 문맥적 표현(Contextual representation)에 더 효과적이다. 따라서 BERT와 동일한 Model size, data, compute라면 더욱 좋은 문맥적 표현이 가능하다.

그리고 ELECTRA Base model에서도 BERT의 Large Model보다 조금 더 좋은 성능을 보여주고 있어 효율면에서도 더 뛰어나다고 볼 수 있다.

* 1. **실제 성능 비교 연구 결과**

**3.3.1 TEST 환경**

1. 데이터셋: 한국어 감정분석 데이터셋(NSMC), 영어 감정분석 데이터셋(Friends)



**<그림 5> NSMC 긍/부정 label 분포표(Train set)**

1. 작동환경: Google Colab pro(GPU tesla v100, CPU: Intel(R)Xeon(R)CPU @ 2.30GHz, 고용량 램 25.51GB)
2. Tensorflow와 Pytorch 모두 이용
3. 전처리: 한국어(KoNLPy Kkma, AutoTokenizer) 영어(Nltk, ElectraTokenizer)

성능 비교를 위해 LSTM, BERT, KOBERT, ELECTRA, KOELECTRA 모델을 테스트해 보았고 KOBERT는 KOELECTRA Base모델과의 비교를 위해 Fine-tuning된 Large/Base모델로 테스트해 보았다.

성능순서는 **ELECTRA(KOELECTRA)> BERT(KOBERT)> LSTM** 순이었다. LSTM은 구형 모델 성능비교용으로 NSMC기준 약 85.9%, Friends 기준 약 41.1%의 정확도를 보여 크게 변별력은 없다고 판단된다.

NSMC KOBERT Large모델(약 90.1%)은 KOELECTRA base모델(90.4%)과 비슷한 수준의 정확도를 보였으나 동일 Epoch기준 학습시간에서 KOELECTRA는 2시간, KOBERT는 13시간으로 많은 차이를 보였으며 KOBERT의 성능이 KOELECTRA보다 일정하지 않아 효율성, 안정성면에서 KOELECTRA가 더 뛰어남을 알 수 있었다.

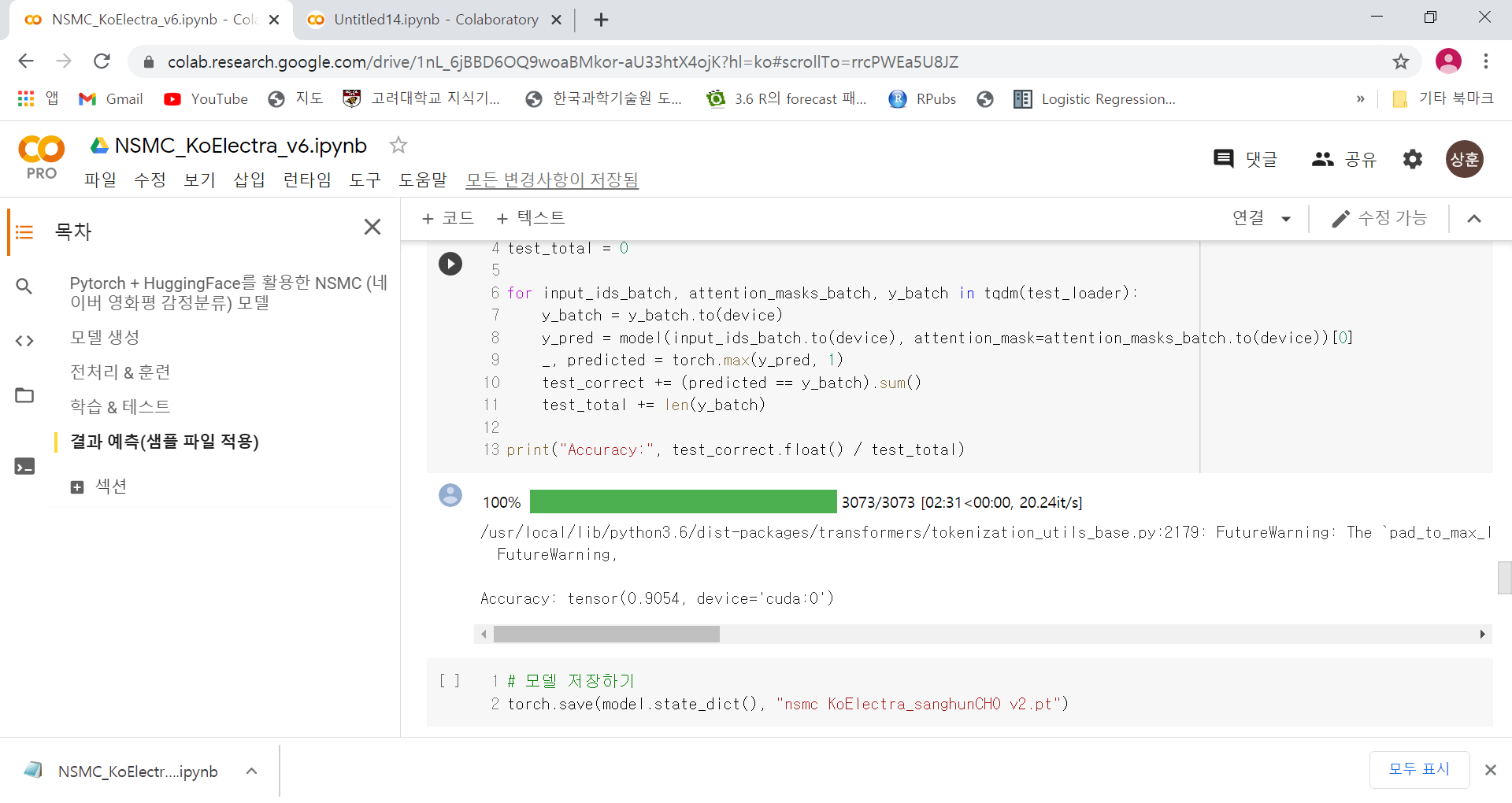
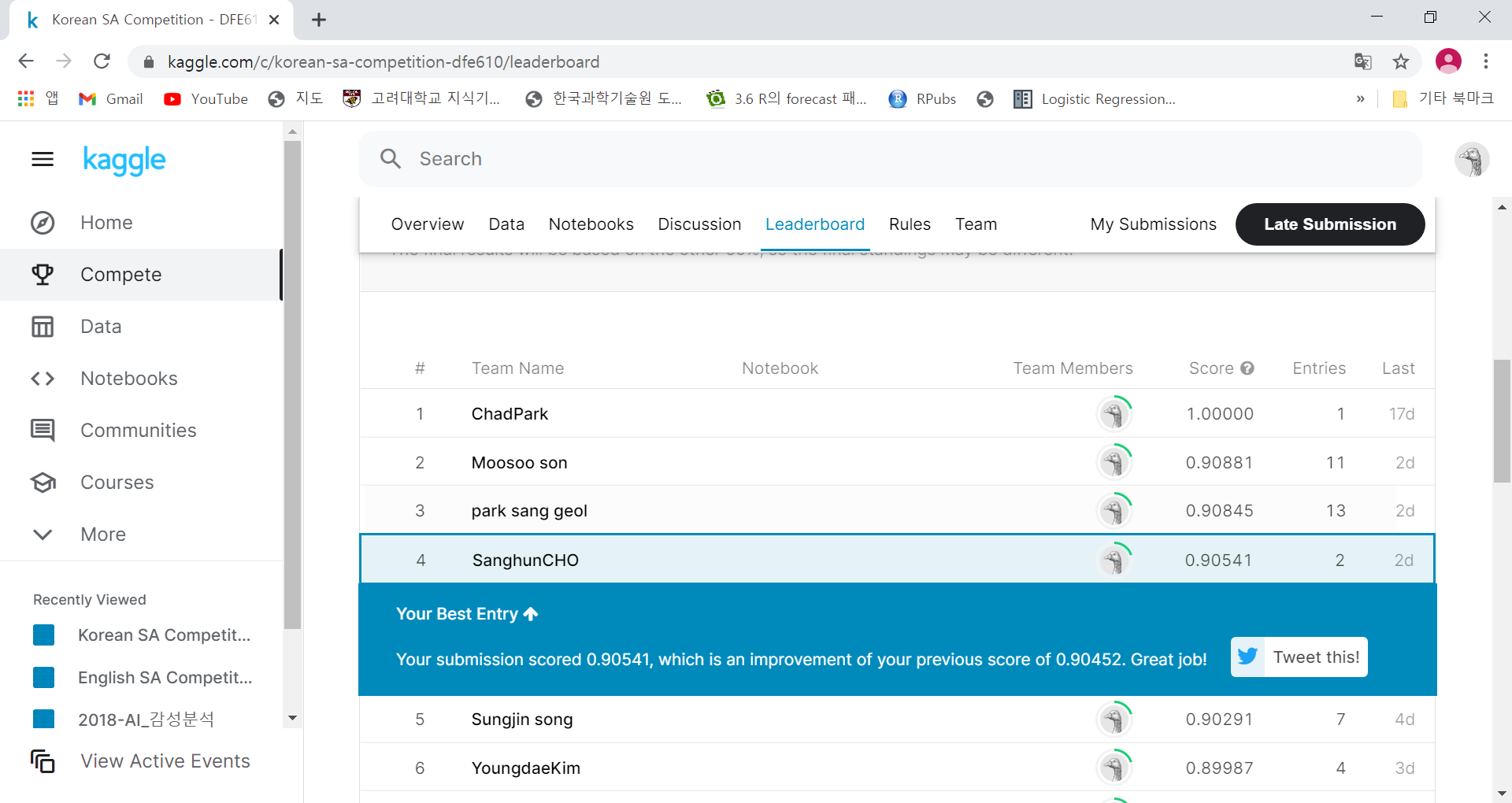
**[표2]한국어 모델 성능 비교(NSMC)**

|  |  |  |  |
| --- | --- | --- | --- |
| **모델명** | **Epoch 수** | **시도횟수** | **평균acc** |
| **KOELECTRA** | 10 | 6 | **0.9045** |
| **KOBERT-Large** | 20 | 2 | **0.9014** |
| **KOBERT-Base** | 5 | 16 | **0.8823** |
| **BERT (pre-trained)** | 4 | 6 | **0.8977** |
| **LSTM** | 15 | 3 | **0.8592** |

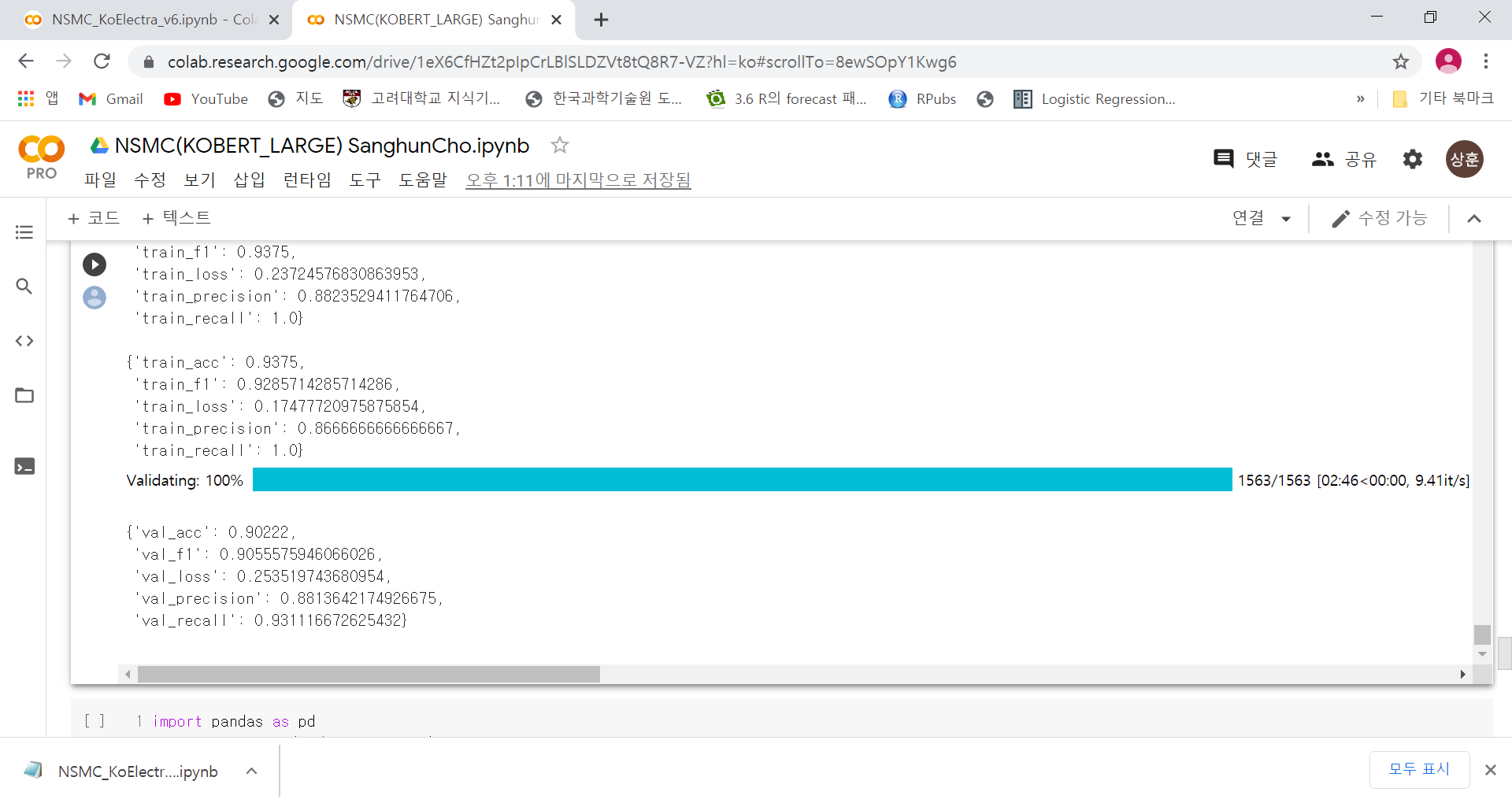
**[표3]영어 모델 성능 비교(Friends)**

|  |  |  |  |
| --- | --- | --- | --- |
| **모델명** | **Epoch 수** | **시도횟수** | **평균acc** |
| **ELECTRA** | 10 | 18 | **0.5528** |
| **BERT** | 5 | 10 | **0.5416** |
| **LSTM** | 20 | 4 | **0.4112** |

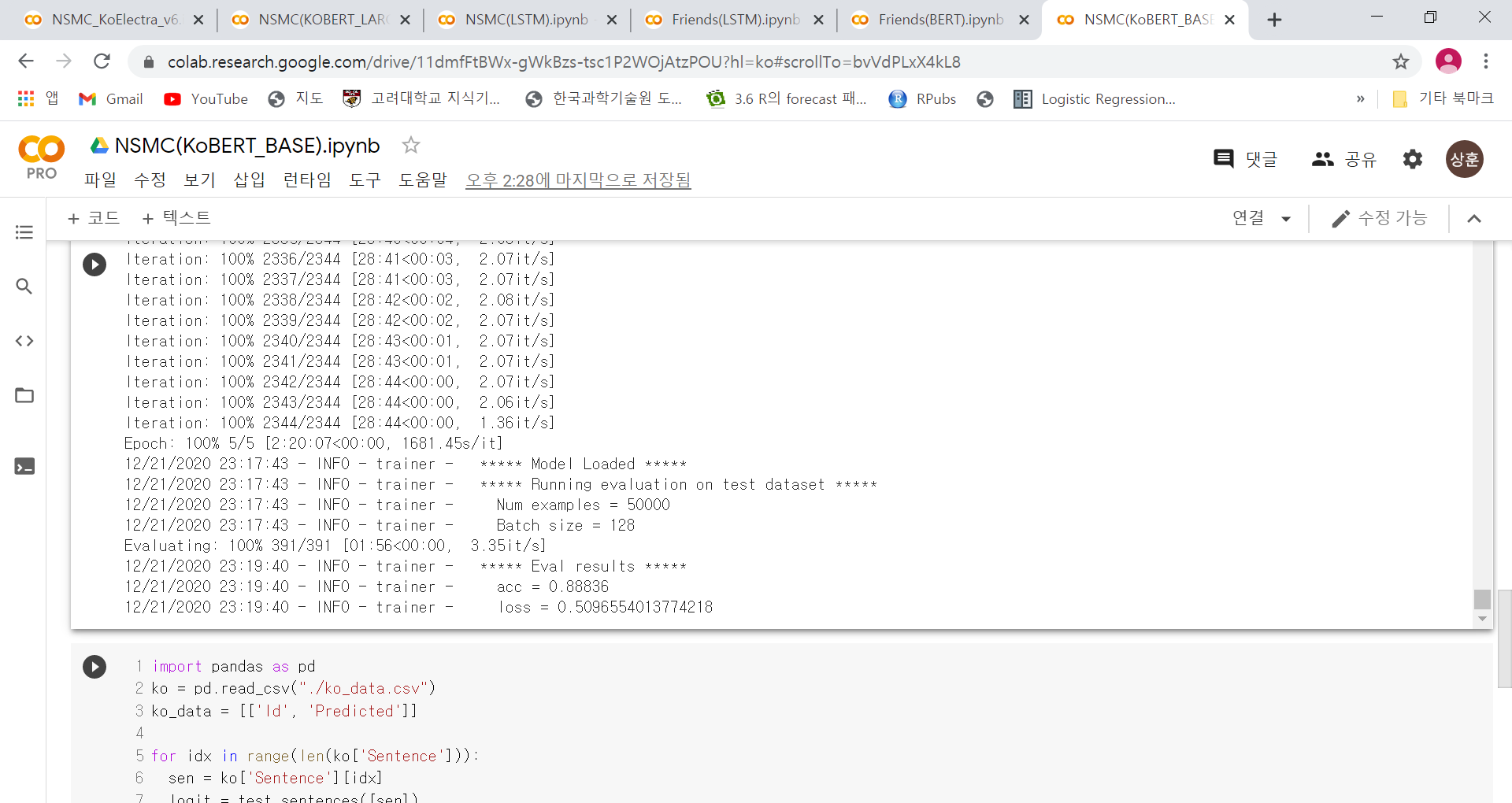
**[KOELECTRA – NSMC]**

****

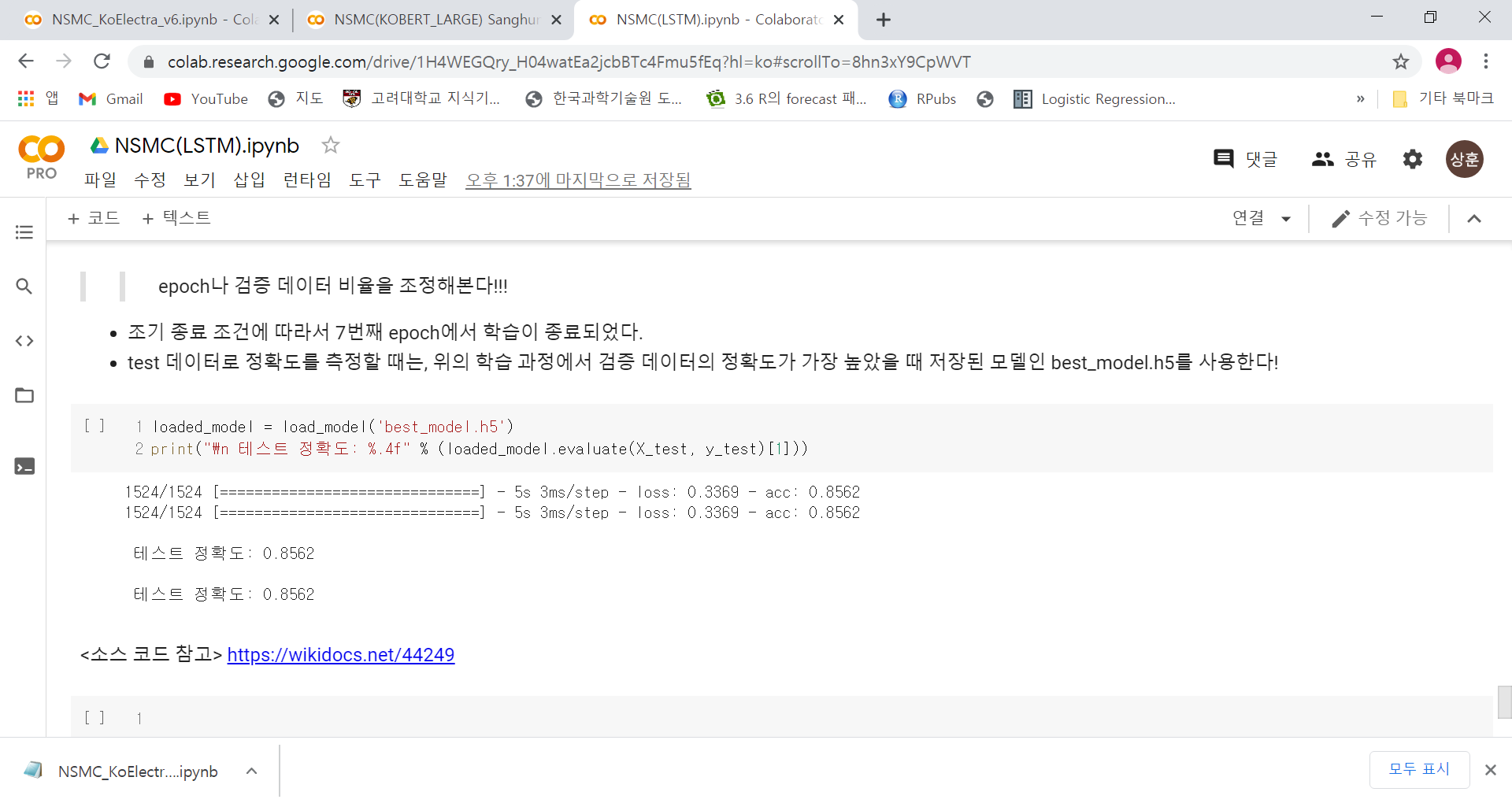
**[KOBERT(Large) – NSMC]**

****

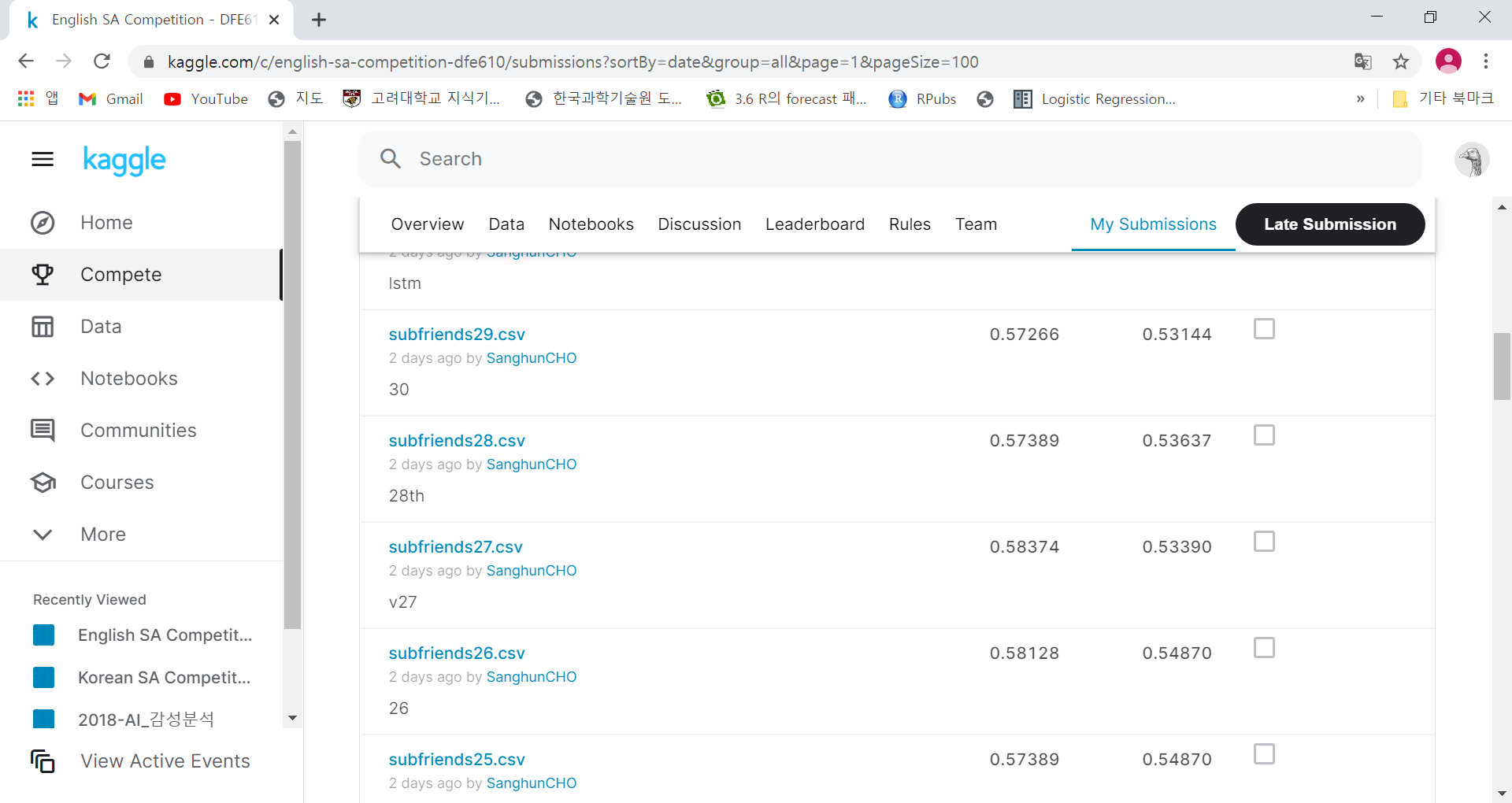
**[KOBERT(Base) – NSMC]**

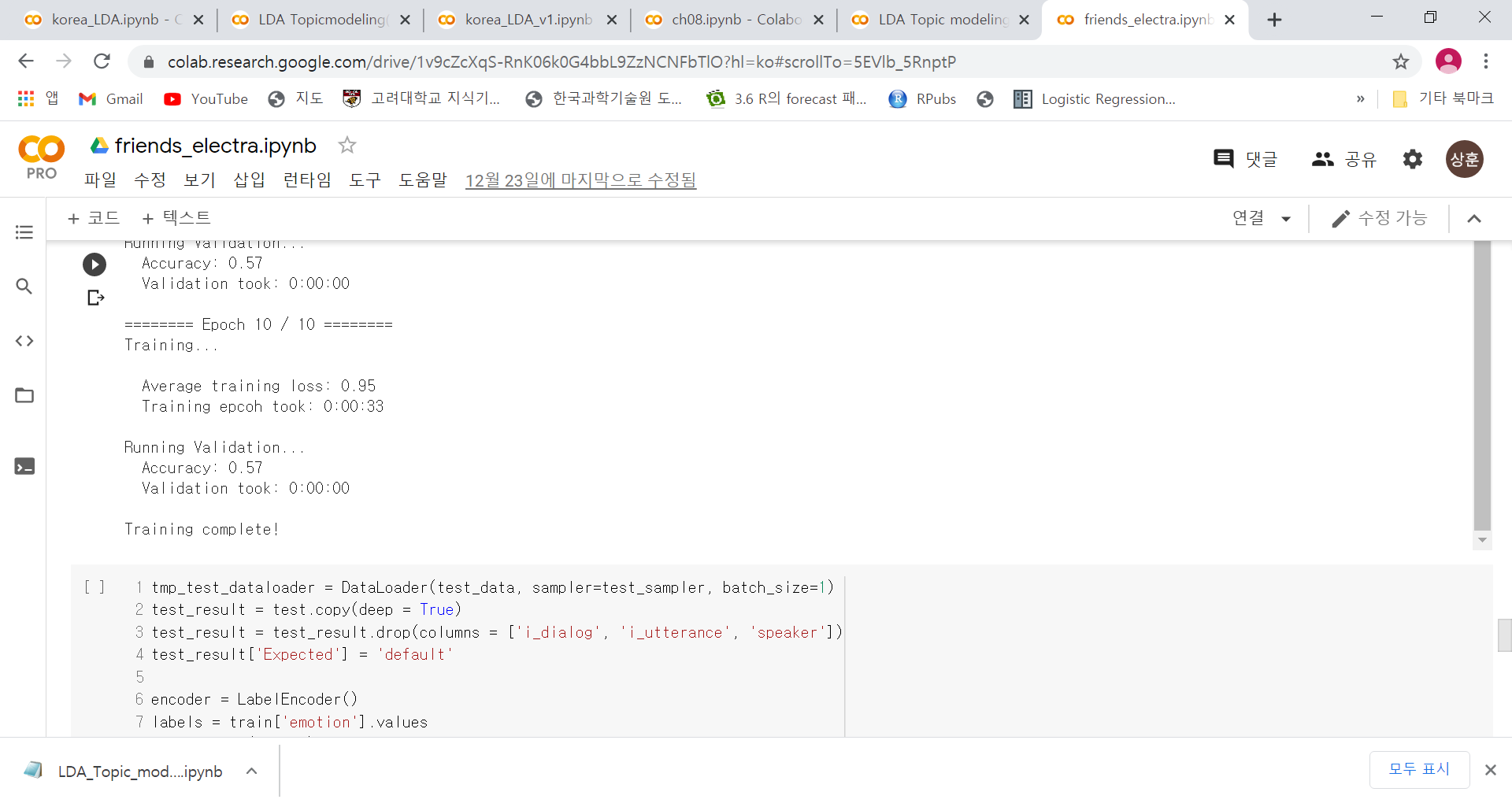
****

**[LSTM – NSMC]**

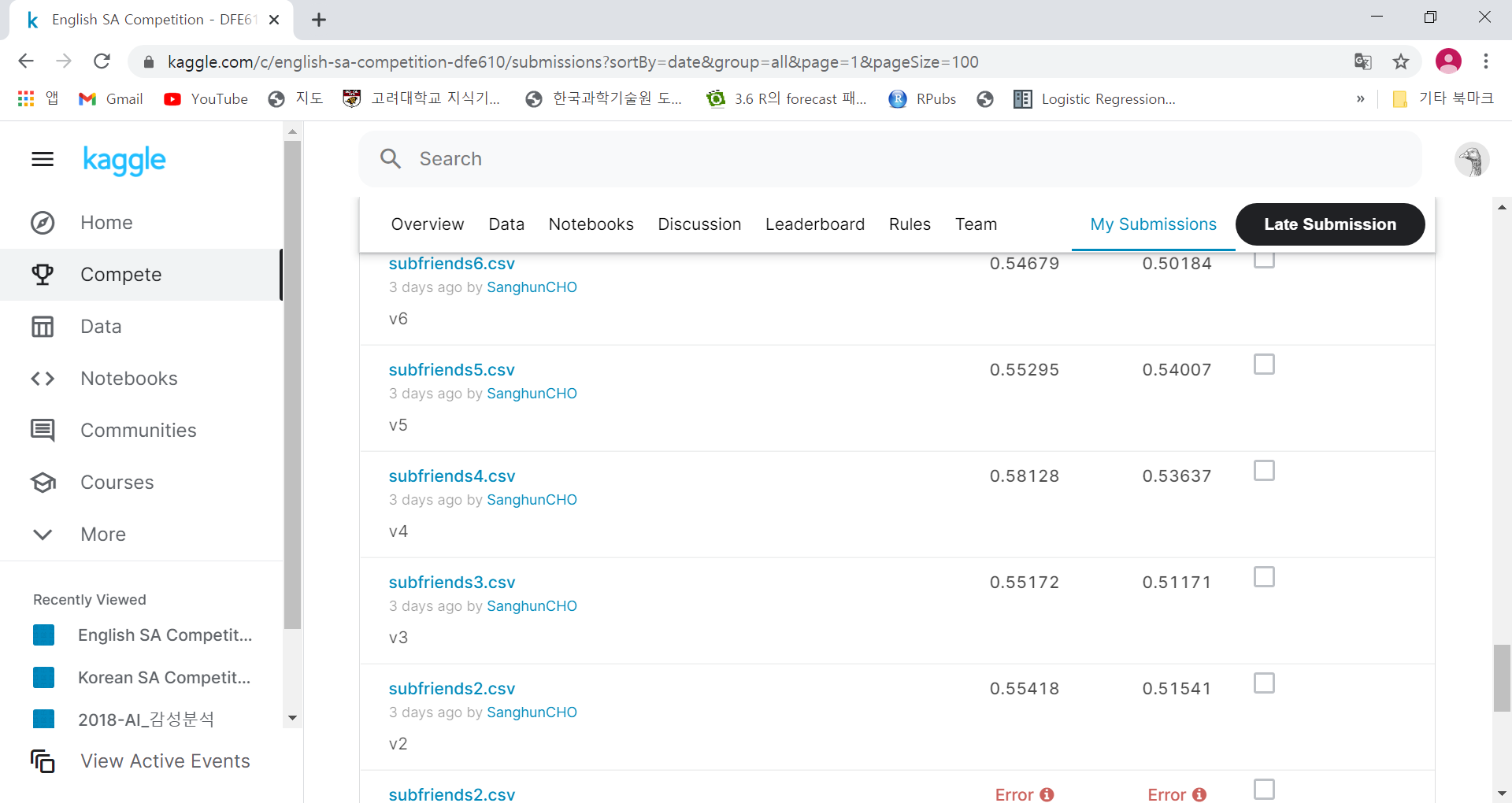
****

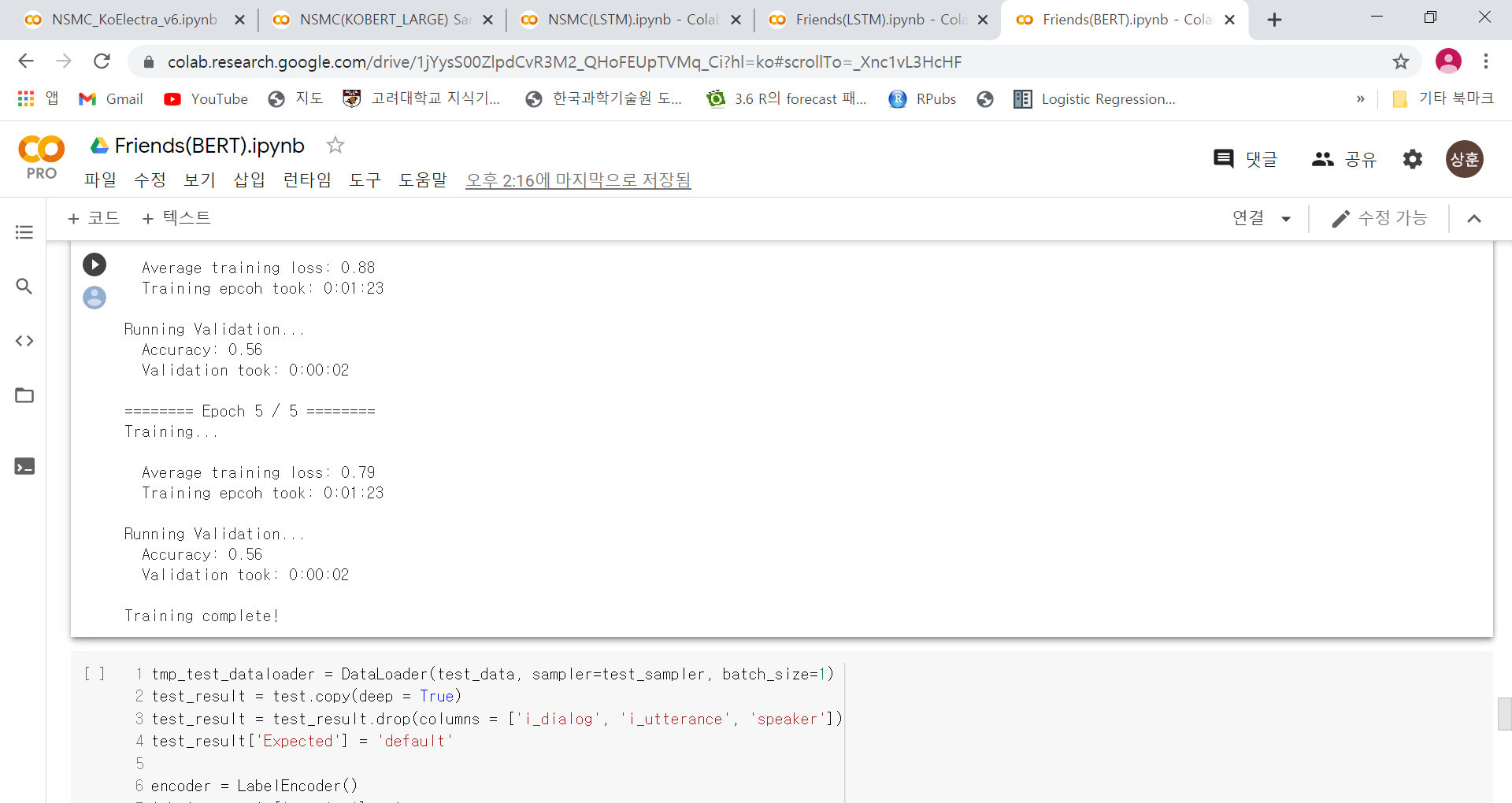
**[ELECTRA – Friends]**

****

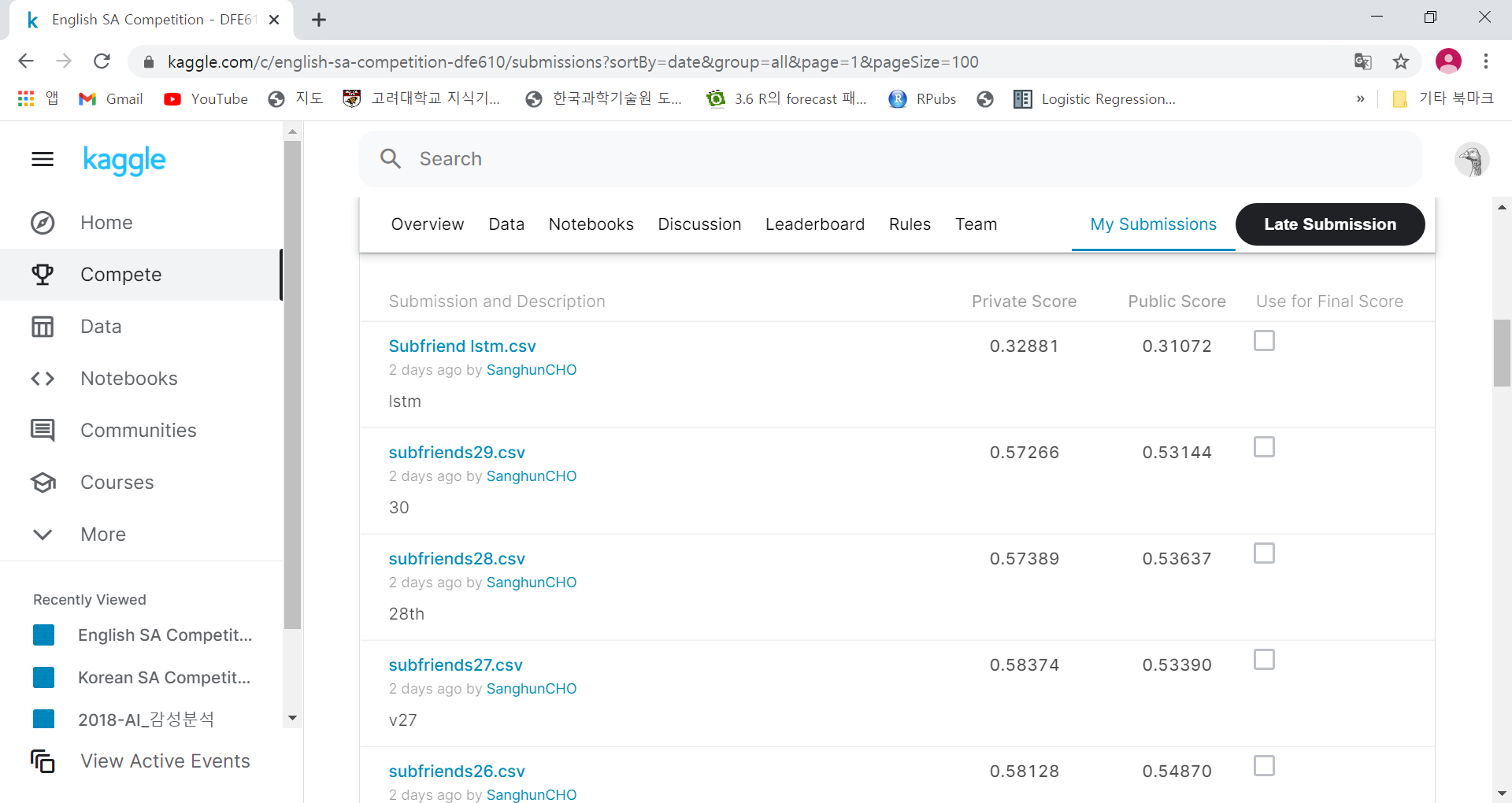
****

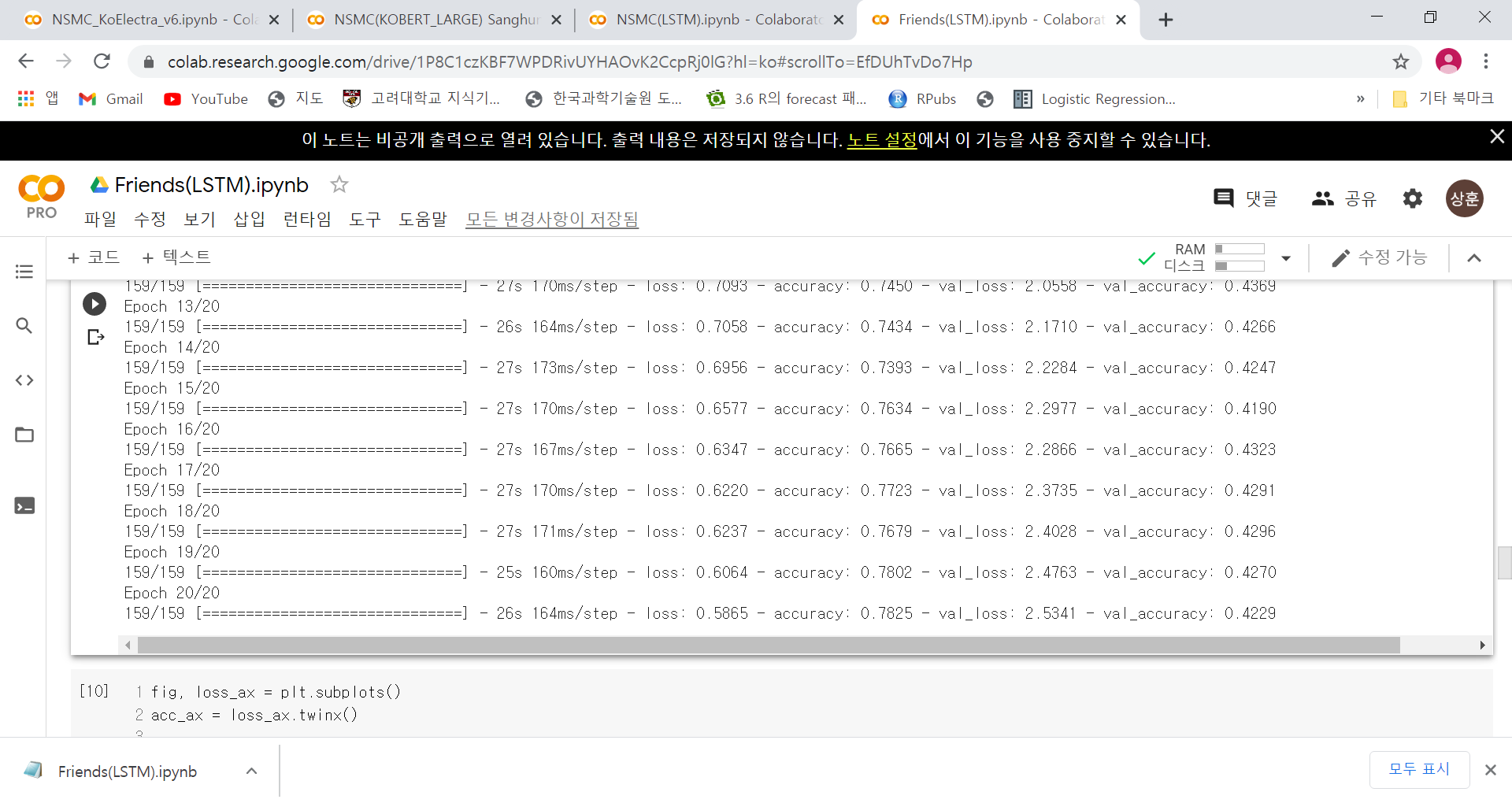
**[BERT – Friends]**

****

****

**[LSTM – Friends]**





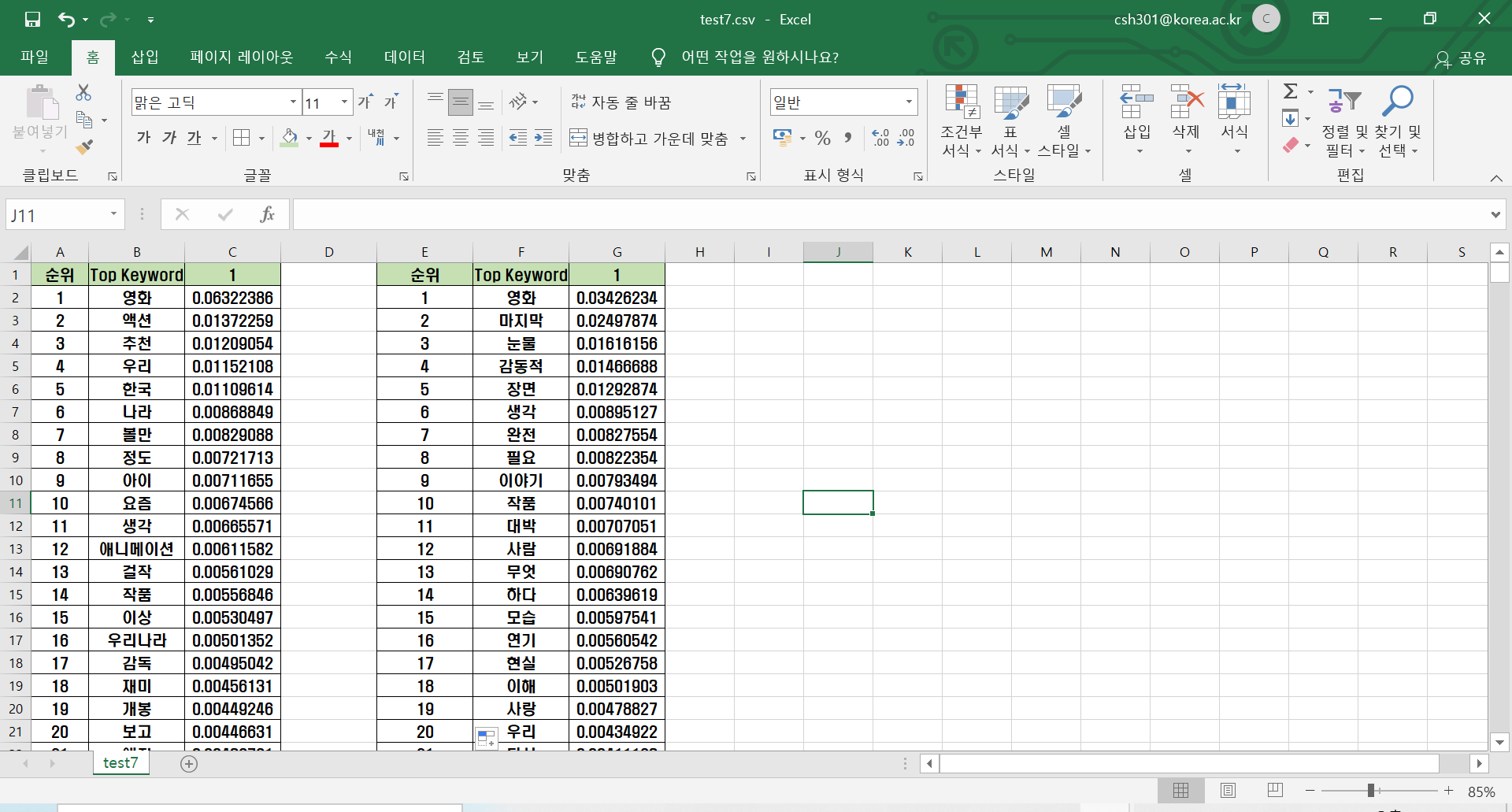
**<그림 6> 모델별 Trial 결과 화면**

**4. LDA(Latent Dirichlet Allocation)토픽모델링 연구결과**

**4.1 NSMC LDA 토픽모델링**

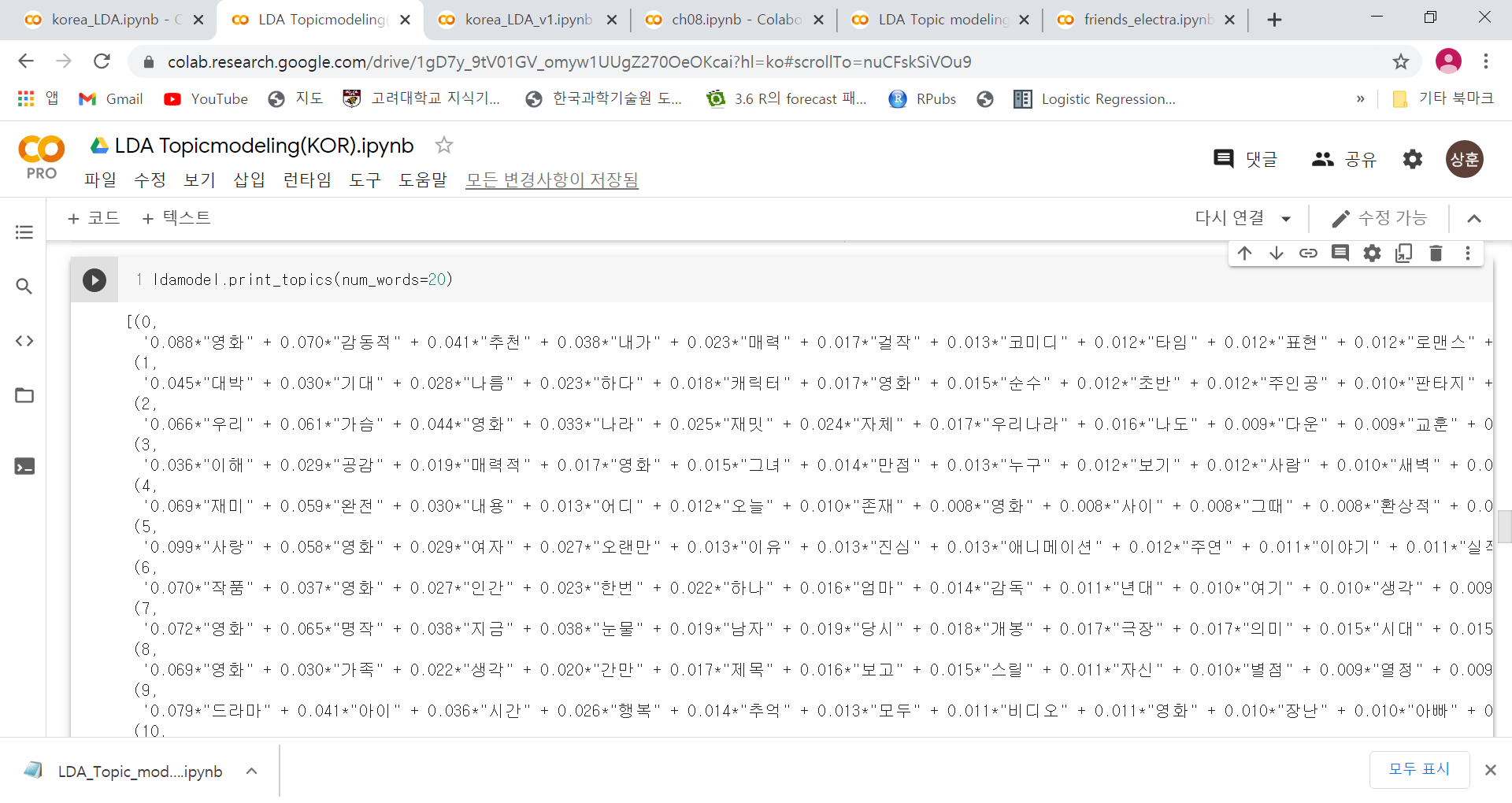
**4.1.1 테스트 환경**

1. 데이터셋: nsmc`ratings.txt`: 200K reviews
2. 작동환경: Google Colab pro
3. Gensim 패키지 이용
4. 전처리: KoNLPy Kkma

**4.1.2 토픽 모델링 결과**

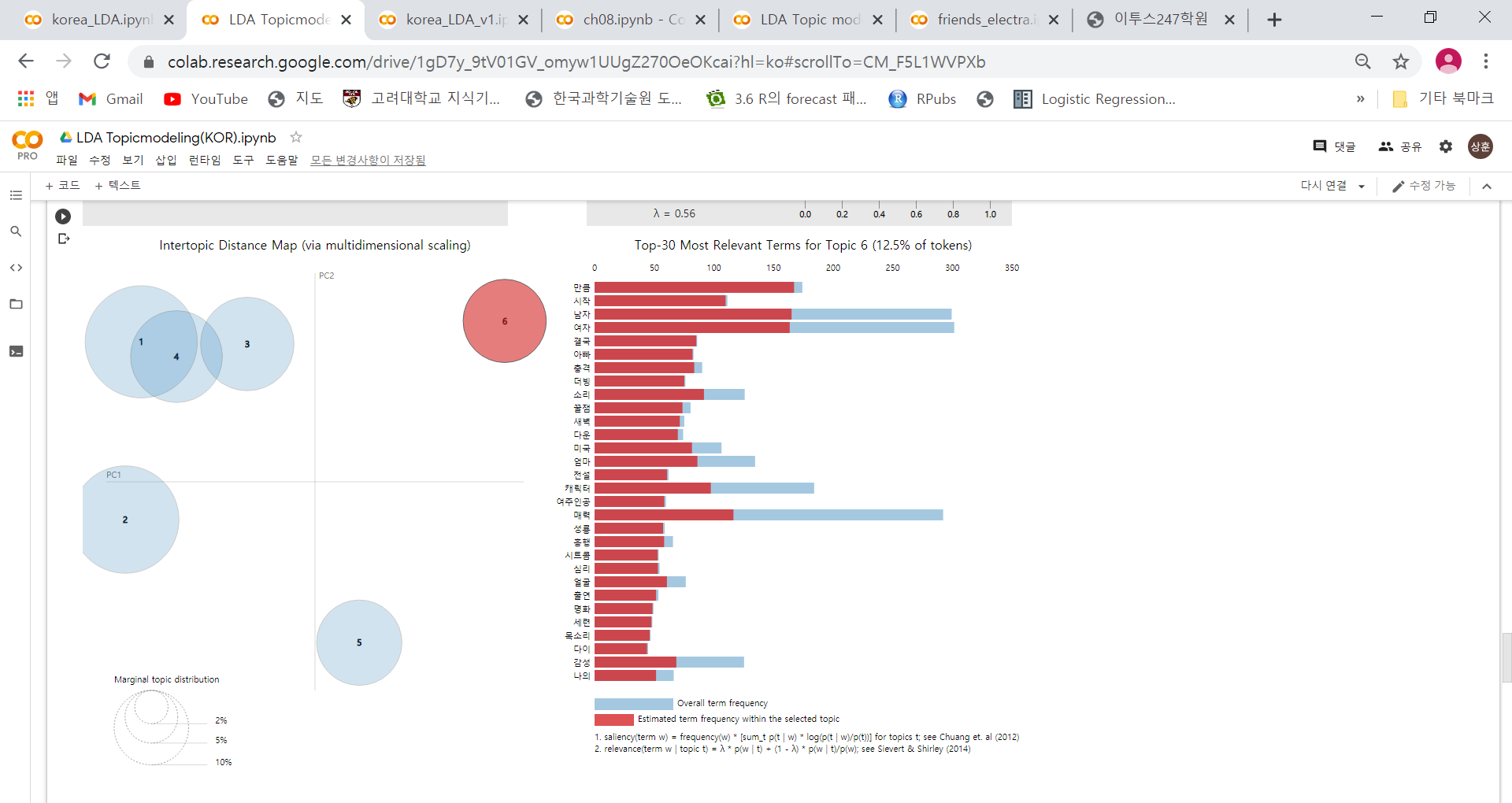
**<그림 7> 상위 20개 Keyword 및 가중치**

* 상위 Keyword 100개 이내에 각 장르(액션, 드라마, 애니메이션, 로맨스, 사극 등)가 확인되어 추후 긍정 부정 데이터와 결합하여 국내 영화 트렌드 장르별 선호도 등 다양한 조합의 분석 가능
* 명사, 형용사를 통한 감정 어휘 분석(Emotion Trigger)[2]을 통해 기존 긍정 부정 분석에서부터 각 감정 별 명사 사용량 분석, 연령별 감정 표현 사용 단어 분석 등 감정분석 세분화로 연구 발전 가능



**<그림 8> 상위 키워드의 단어 별 가중치 확인**

* Keyword 상위권 단어 중 중복 단어 및 ‘영화’와 같이 감정분석에 의미 없는 단어들에 대한 가중치조정 필요



**<그림 9> 각 토픽 별 분포도**



**<그림 10> Most Relevant Terms for Topic**

* 각 토픽 별 관계 파악하여 문맥을 파악한 감정분석에 활용가능. 감마 값 조정을 통해 결과치 세부 조정 필요.

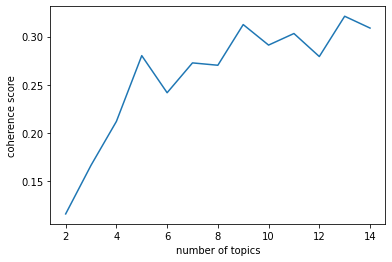
**4.1.3 토픽모델링 성능 평가**

**[Perplexity]**

**<그림 11> Perplexity 성능 평가 그래프**

* Perplexity는 동일 모델 내 파라미터에 따른 성능 평가할 때 주로 사용하는 지표로 측정값이 낮을수록 우수한 성능으로 평가됨.
* 선정된 토픽 개수마다 학습시켜 가장 낮은 값을 보이는 구간을 찾아 최적화된 토픽의 개수 선정 가능함.
* **측정 내용 분석 결과 토픽 개수 12~13개일 때 가장 성능이 좋은 것으로 측정됨.**
* 보완점: Perplexity 가 낮다고 해서 무조건 결과가 해석 용이하다는 의미가 아니므로 성능평가 모델을 종합적으로 고려할 필요가 있음.

**[Topic Coherence]**



**<그림 12> Topic Coherence 성능 평가 그래프**

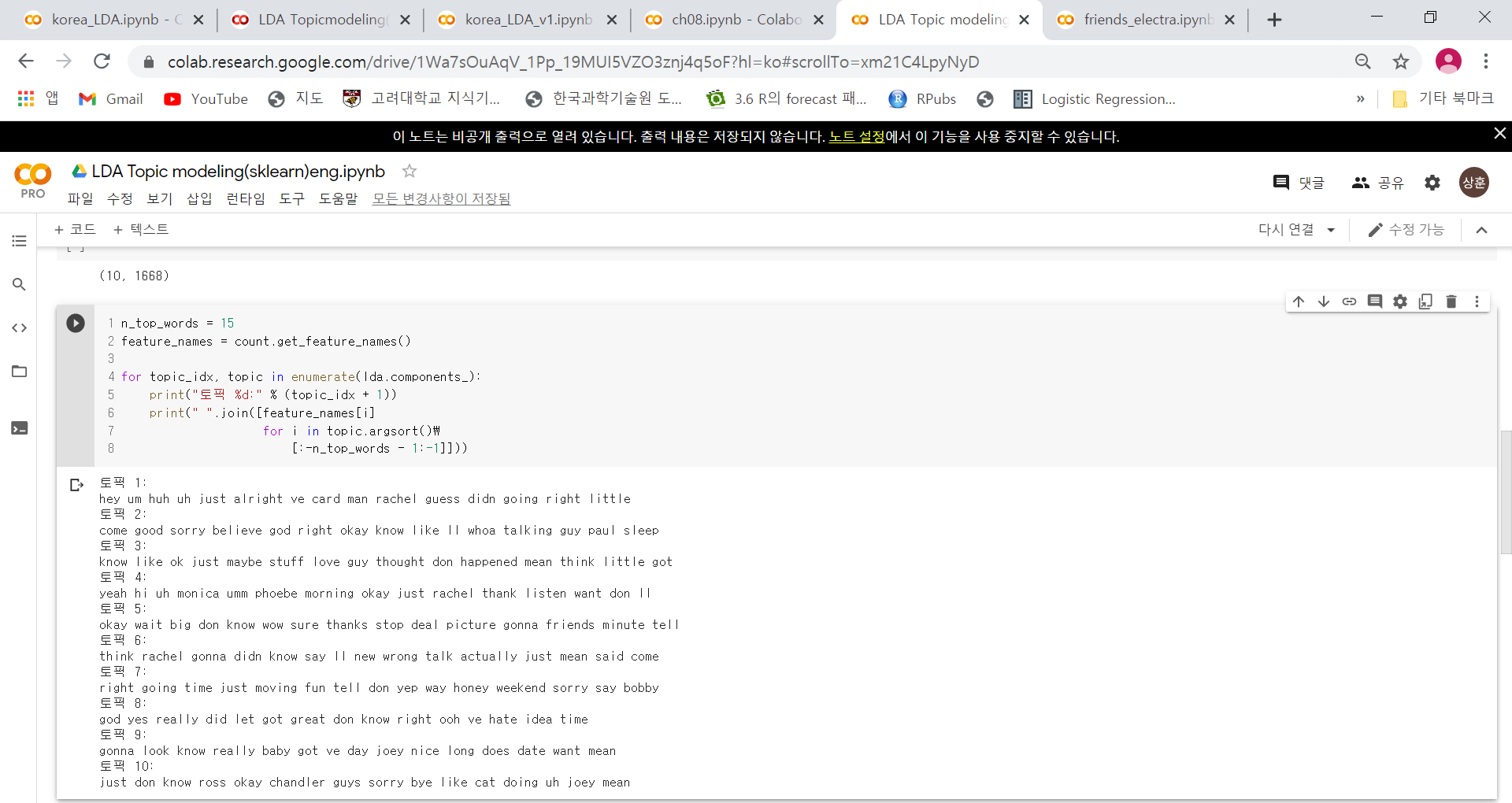
* Coherence는해당 모델이 얼마나 실제로 의미 있는 결과를 내는지 의미론적으로 일관성 있는지 확인하는 지표로 높을수록 의미론적 일관성 높은 것으로 판단함.
* **측정 내용 분석 결과 토픽 개수 13개일 때 가장 성능이 좋은 것으로 측정됨.**
* 보완점: 기존에는 언어모델 평가로 Coherence Model 만을 사용 후 원하는 토픽 개수의 Coherence 모델을 지속 학습시켜 토픽을 할당했으나 추후에는 두 가지 모델 함께 적용해보는 것과 좀 더 정밀한 사용이 필요함.

**4.2 Friends LDA 토픽모델링**

**4.2.1 테스트 환경**

1. 데이터셋: Kaggle용testset: en\_data.csv:1,623개
2. 작동환경: Google Colab pro
3. sklearn LatentDirichletAllocation 패키지 이용
4. 전처리: sklearn CountVectorize

**4.2.2 토픽 모델링 결과**



**<그림 13> Sklearn LDA Topic modeling 결과**

* 각 토픽에서 가장 중요한 단어 다섯 개를 기반으로 LDA가 다음 토픽을 구별했다고 추측할 수 있습니다.
* 토픽 1: 일반적인 감탄사(hey,oh etc),레이첼

토픽 2: 안타까움, Paul, 잠

토픽 3: 긍정, 사랑

토픽 4: 감탄사, 모니카

토픽 5: 일상적, 친구

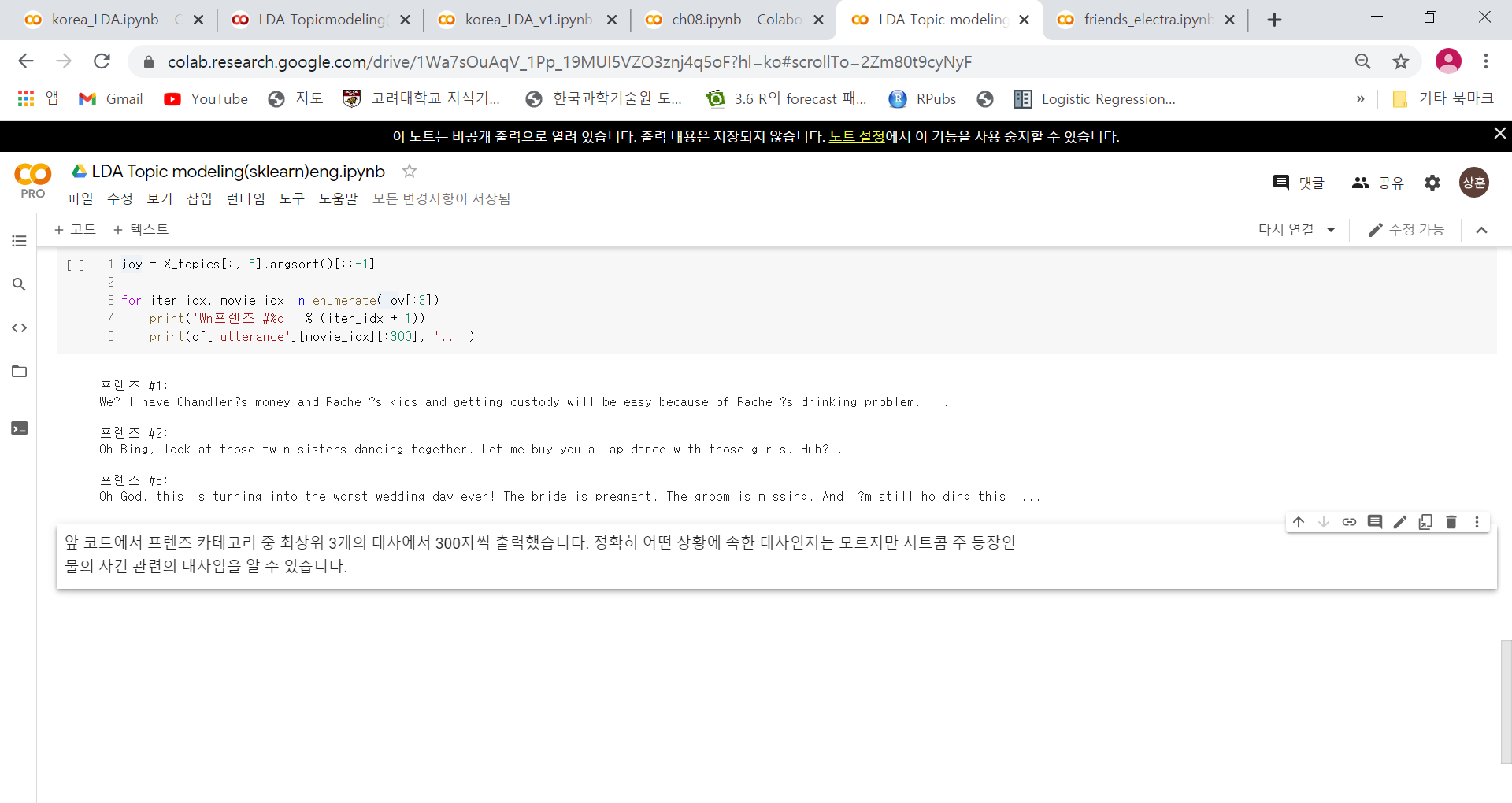
토픽 6: 답답함, 레이첼

토픽 7: 재미, 미안함, 바비

토픽 8: 놀람, 증오

토픽 9: 좋음, 조이

토픽 10: 미안함, 챈들러, 조이, 고양이

* 일부 경진대회용 데이터셋만 활용하여 실제 의미 있는 Topic이 추출되지는 않았음.
* 다만 상위 토픽 1,2,3에 나오는 레이첼, Paul, 잠, 사랑 등과 연관 지어 대사 내용을 추론하면 프렌즈 101번째 에피소드인 “The One Where Monica Gets A Roommate”에 관한 주요 키워드가 된다는 점에서 추후 토픽모델링을 통해 주인공들의 감정변화, 대사를 통한 해당 에피소드 추측 등 흥미로운 연구 방향이 있을 것이라 예측됨.

**<그림 14> 토픽모델링 상위 3개 대사 추출**

- 프렌즈 카테고리 중 최상위 3개의 대사에서 300자

씩 출력한 결과로 정확히 어떤 상황에 속한 대사인

지는 시트콤의 특성상 간접적인 비유 및 Joke의 포

함으로 자세하게 파악되지 않으나 시트콤 주 등장인

물의 중요 사건 관련 대사임을 알 수 있다.

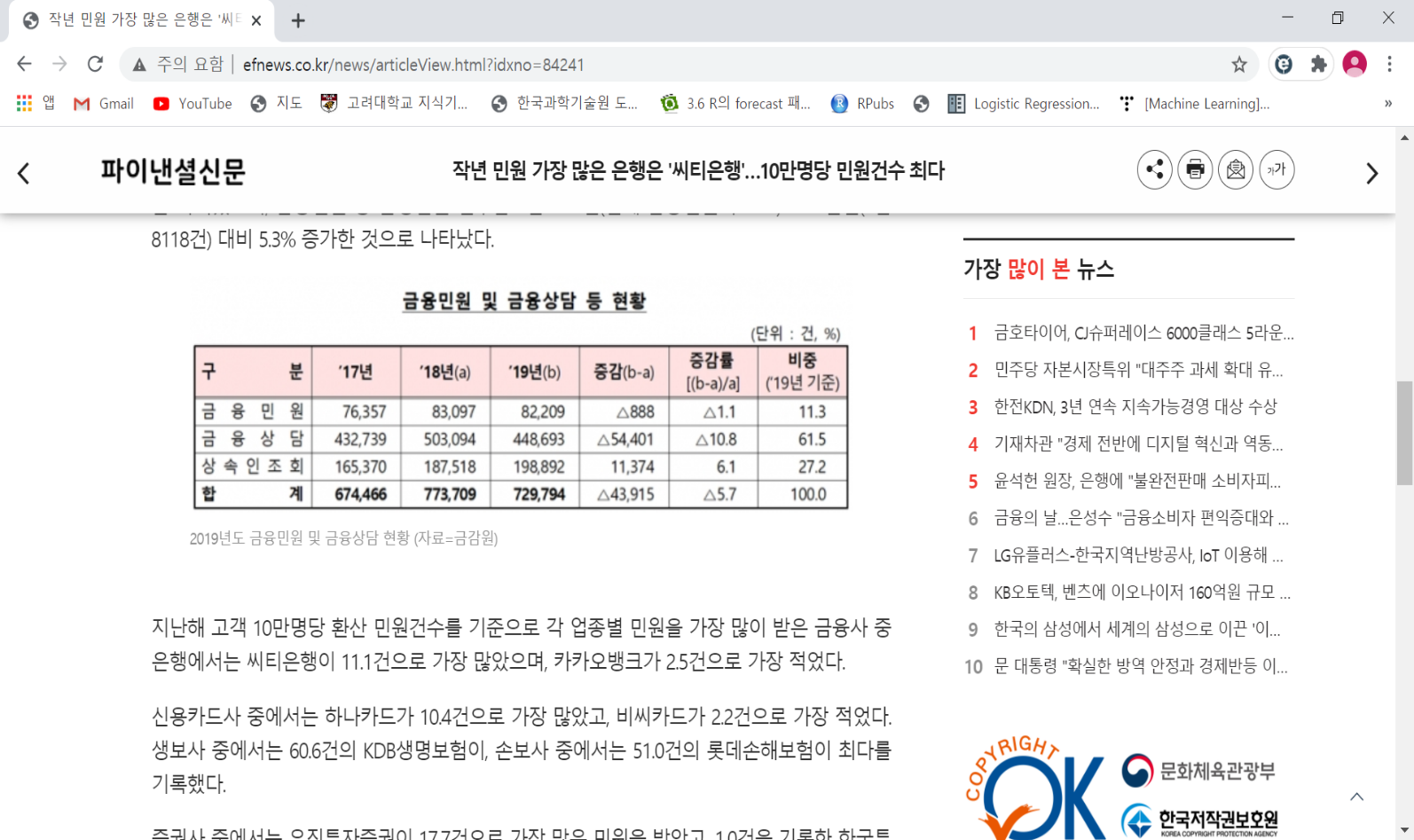
**5. 결론**

금융업 상담의 발화 주제는 일상대화에 비해 폭이 넓지 않다. 상품가입 관련, 대출진행 관련, 환율관련 등 전문분야에 대한 주제로 한정된다. 그러나 실제 상담시에 벌어지는 상담서비스의 질, 고객의 감정 상태 등 기타 상황에 따라 수많은 케이스의 민원이 발생하고 신속한 대응이 필요한 상황이다. 따라서 상담내용에 대한 감정분석을 통한 신속한 발화주제를 분석하여 최적의 상담서비스를 제공하여 최대한 고객 불만을 줄이는 것이 민원관련 해결의 기본 목표라고 생각한다.

현재 고객과의 상담 시 콜센터 응대비율이 늘어나고 있고 전부 녹취를 기본으로 하고 있다. 따라서 음성인식기술(STT: Speech to Text)이 도입되고 있으나 시초이고 많은 발전을 요하고 있다.

따라서 향후에는 두가지 정도 방향으로 발전되어야 한다고 생각한다.

1. 고객상담내용을 말뭉치로 분류하여 실시간으로 주제를 파악(Topic modeling) 후 관련 상품으로 안내 및 상담직원에 기본내용 전달하는 상담보조시스템
2. 민원내용의 요지를 파악하고 고객의 감정 상태를 실시간으로 구분(ELECTRA 등 최신 언어모델을 활용한 Sentiment Analysis)하여 상담직원 또는 AI상담사가 대응 실시

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 예측할 수 있다.

단어별로 토픽 모델링 결과를 잘 나타내고서 문서별로 다시 묶은 다음 각 주제에 해당하는 문서끼리 다시 토픽모델링 결과를 낸다면 하나의 주제에서 또 다르게 이야기하는 주제들을 끄집어 낼 수 있고 키워드 간의 관계를 밝힘으로써 의미론적 요소를 결합하면 감정유발 요인을 찾아 내어 문맥상 고객의 실제 감정을 반영하고 단순 감정 분류만이 아니라 실제 감정들이 어떤 요인으로 인해 유발되는지 보다 깊게 파악하여 민원이 유발될 수 있는 원인 분석에 다양한 용도로 활용될 것이라고 예상한다.

다만 향후에는 이러한 단순 토픽은 한국어의 명사, 형용사, 품사 등 만을 활용하기 때문에 감정 어휘와 감정유발요인 간의 명확한 인과관계의 규정이 추가적으로 연구되어야 조금 더 정확한 결과가 나올 것이고 이는 감정사전을 좀더 정교하게 구축하여 보완해야 할 것이다.

**참고문헌**

1. ELECTRA: Pre-training Text Encoders

as Discriminators Rather Than Generators, Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning ,2020

1. 텍스트 마이닝을 이용한 감정 유발 요인 Emotion Trigger에 관한 연구, Juyoung An, Junghwan Bae, Namgi Han, Min Song,2015
2. Latent Dirichlet Allocation, David M. Blei, Andrew Y.Ng, and Michel I. Jordan, Journal of Machine Learning Research 3page : 993-1022, Jan 2003
3. Efficient Estimation of Word Representations in Vector Space, T.Mikolov, K.Chen, G.Corrado, and J.Dean, arXiv preprint arXiv:1301,3781, 2013
4. Topic Modeling, LDA(고려대학교 강필성 교수님 강의 정리) ratsgo’s Blog for textmining, 2017
5. https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/
6. Leveraging Latent Dirichlet Allocation in processing free-text personal goals among patients undergoing bladder cancer surgery,2019
7. 특허 문서를 위한 형태소 분석기 비교평가. 한국정보기술학회 종합학술발표논문집, 이유진, 김세빈, 홍현석, 김장원 ,2019
8. <https://www.sentv.co.kr/news/view/579490> “코로나에 은행 민원 30% 급증…“상환 유예” 서울경제,2020.9
9. 금융민원 주요 통계 “금융감독원” www.fss.or.kr

* 기타 소스 코드 및 인터넷 참고자료 관련 모음: <https://github.com/csh301/NLP-FINAL-REPORT>