텍스트 마이닝을 활용한 통화정책 분석

KOREAN NLP PYTHON PACKAGE FOR ECONOMIC ANALYSIS

정윤성 강동훈 김요한 이동재

역할 분배

정윤성

의사록 데이터 처리

강동훈

뉴스 데이터 처리

김요한

자료 정리 및 발표자료 준비

이동재

채권 데이터 처리

프로젝트 환경 구축

Github	프로젝트에서 진행한 코드를 공유 및 버전관리를 합니다
Google drive	프로젝트에 필요한 데이터를 공유하고 논문문서 파일을 공유합니다
Server	크롤링, 전처리 등 많은 연산작업을 필요로 하는 코드는 서버로 처리합니다
Slack	Gitgub, Google Drive, Trello의 업데이트를 Slack를 통해 확인합니다
Trello	프로세스, 업무 todo 확인합니다

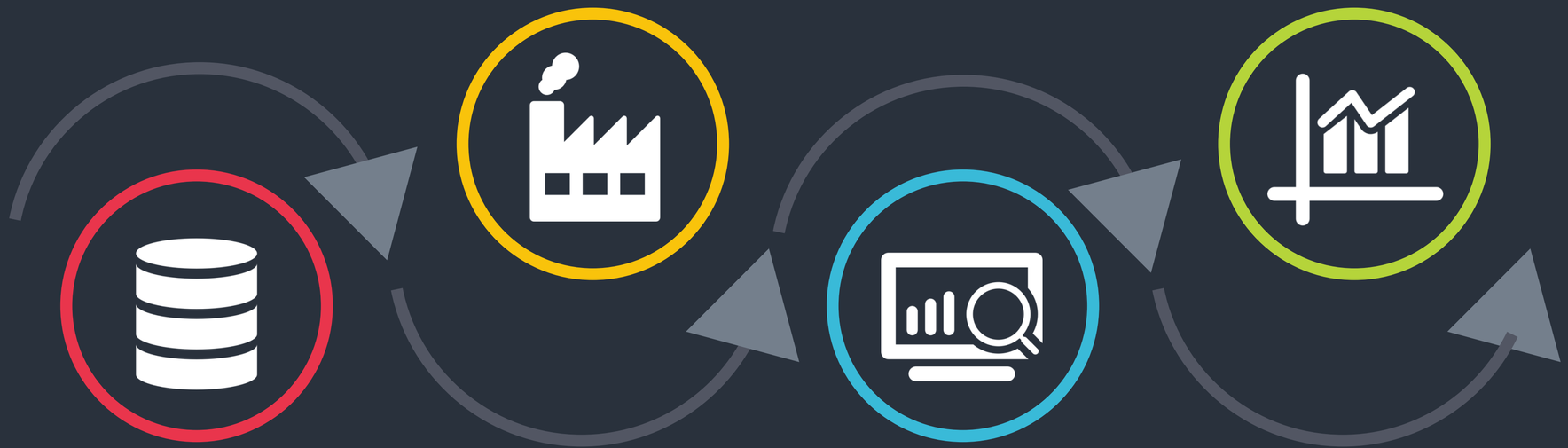
프로젝트 프로세스

콜금리, 토큰화. 품사 부착, 원형 복원,
불용어 처리, N-grame 처리

텍스트 전처리 단계

Polarity score를 통해
금통위의사록을 분석

감성분석 단계



데이터 수집 단계

뉴스기사, 금통위의사록, 채권분석보고서
(2005년~2017년)

극성분석 단계

콜 금리를 기준으로 라벨링,
Polarity score를 작성하여 pos, nag로 분류

• 데이터 수집 단계

1

뉴스 기사 크롤링

- 연합 뉴스, 연합 인포맥스, 이데일리 언론사 2005년~2017년 '금리' 단어 포함 기사를 크롤링
- 데이터를 통합

2

금통위의사록 크롤링

- 한국은행 홈페이지에서 2005년~2017년 금통위의사록 파일을 추출(PDF)
- PDF → TXT → CSV 파일로 변환

3

채권분석보고서 크롤링

- 네이버 금융 탭에서 2005년~2017년 채권분석보고서 파일을 추출(PDF)
- PDF → TXT → CSV 파일로 변환

1 뉴스 기사 크롤링

- 기사 30만 1223개
- 날짜, 매체명, 내용, 제목, url로 이루어진 csv 생성

```
1 import pandas as pd
2
3 file1_csv = pd.read_csv('news_final.csv')
4 file1_csv
```

301218	301218	2011.06.07	연합인포맥스	(뉴욕=연합인포맥스) 김홍규 특파원 = 미국 달러화는 7일 뉴욕 외환시장에서 중국발...	美달러, 中발 악재 불구 엔화에 강보합	http://news.einfomax.co.kr/news/articleView.ht...
301219	301219	2011.06.08	연합인포맥스	(서울=연합인포맥스) 신경원 기자 = 엔화는 유로존의 재정 우려와 벤 버냉키 연방준...	<유럽환시> 엔화, 안전선호심리로 강세	http://news.einfomax.co.kr/news/articleView.ht...
301220	301220	2011.06.09	연합인포맥스	(뉴욕=연합인포맥스) 김홍규 특파원 = 유로화는 8일 뉴욕 외환시장에서 안전통화 선...	유로화, 안전통화 매수로 하락 지속	http://news.einfomax.co.kr/news/articleView.ht...
301221	301221	2011.05.23	연합인포맥스	(서울=연합인포맥스) 정선미 기자= 매파적인 유럽중앙은행(ECB)이 유로화에 자산이...	"매파적 ECB는 유로화에 자산 아닌 부채"<다우존스>	http://news.einfomax.co.kr/news/articleView.ht...
301222	301222	2011.06.09	연합인포맥스	(뉴욕=연합인포맥스) 김홍규 특파원 = 유로화는 8일 뉴욕 외환시장에서 세계 경제 ...	유로화, ECB 금리인상 가능성 약화로 낙폭 확대	http://news.einfomax.co.kr/news/articleView.ht...

301223 rows × 6 columns

2 금통위의사록 크롤링

- 의사록 147개 PDF (2017년 까지)

B	C	D
0.hwp금융통화위원회 의사록2006년도 제20차 회의1. 일 자 2006년 9월 21일 (목) 2. 장 소 금융통화위원회 회의실3. 출석위	2006.9.21	
1.hwp- 1 -금융통화위원회 의사록2006년도 제21차 회의1. 일 자 2006년 10월 12일 (목)2. 장 소 금융통화위원회 회의실3. 출	2006.10.12	
2.hwp금융통화위원회 의사록2006년도 제22차 회의1. 일 자 2006년 10월 26일 (목) 2. 장 소 금융통화위원회 회의실3. 출석	2006.10.26	
3.hwp금융통화위원회 의사록2006년도 제23차 회의1. 일 자 2006년 11월 9일 (목) 2. 장 소 금융통화위원회 회의실3. 출석위	2006.11.9	
4.hwp금융통화위원회 의사록2006년도 제24차 회의1. 일 자 2006년 11월 23일 (목) 2. 장 소 금융통화위원회 회의실3. 출석	2006.11.23	
5.hwp금융통화위원회 의사록2006년도 제25차 회의1. 일 자 2006년 12월 7일 (목)2. 장 소 금융통화위원회 회의실3. 출석위	2006.12.7	
6.hwp금융통화위원회 의사록2006년도 제26차 회의1. 일 자 2006년 12월 21일 (목) 2. 장 소 금융통화위원회 회의실3. 출석	2006.12.21	
2C2F720B1DDC5EBC0A720C0C7BBE7B7CF2E687770>금융통화위원회 의사록2006년도 제2차 회의1. 일 자 2006년 1월 12일 (목)	2006.1.12	
.hwp금융통화위원회 의사록2006년도 제3차 회의1. 일 자 2006년 1월 26일 (목) 2. 장 소 금융통화위원회 회의실3. 출석위원	2006.1.26	
.hwp금융통화위원회 의사록2006년도 제4차 회의1. 일 자 2006년 2월 9일 (목) 2. 장 소 금융통화위원회 회의실3. 출석위원	2006.2.9	
.hwp금융통화위원회 의사록2006년도 제6차 회의1. 일 자 2006년 3월 9일 (목) 2. 장 소 금융통화위원회 회의실3. 출석위원	2006.3.9	
.hwp금융통화위원회 의사록2006년도 제7차 회의1. 일 자 2006년 3월 23일 (목) 2. 장 소 금융통화위원회 회의실3. 출석위원	2006.3.23	
.hwp금융통화위원회 의사록2006년도 제8차 회의1. 일 자 2006년 4월 7일 (금) 2. 장 소 금융통화위원회 회의실3. 출석위원	2006.4.7	
131C2F720B1DDC5EBC0A720C0C7BBE7B7CF2E687770>금융통화위원회 의사록2007년도 제11차 회의1. 일 시 2007년 5월 10일	2007.5.10	
위원회 의사록(2007년도 제13차)(2007.6.8).hwp금융통화위원회 의사록2007년도 제13차 회의1. 일 시 2007년 6월 8일 (금)2. 장 소 금	2007.6.8	
43[1].hwp- 1 -금융통화위원회 의사록2007년도 제14차 회의1. 일 자 2007년 6월 21일 (목)2. 장 소 금융통화위원회 회의실3	2007.6.21	
5.hwp금융통화위원회 의사록2007년도 제15차 회의1. 일 자 2007년 7월 12일 (목)2. 장 소 금융통화위원회 회의실3. 출석위	2007.7.12	
6.hwp금융통화위원회 의사록2007년도 제16차 회의1. 일 자 2007년 7월 26일 (목)2. 장 소 금융통화위원회 회의실3. 출석위	2007.7.26	
137C2F720B1DDC5EBC0A720C0C7BBE7B7CF2E687770>금융통화위원회 의사록2007년도 제17차 회의1. 일 자 2007년 8월 9일 (	2007.8.9	
9.hwp금융통화위원회 의사록2007년도 제19차 회의1. 일 자 2007년 9월 7일 (금)2. 장 소 금융통화위원회 회의실3. 출석위원	2007.9.7	
.hwp금융통화위원회 의사록2007년도 제1차 회의1. 일 자 2007년 1월 5일 (금)2. 장 소 금융통화위원회 회의실3. 출석위원	2007.1.5	
0.hwp금융통화위원회 의사록2007년도 제20차 회의1. 일 자 2007년 9월 20일 (목)2. 장 소 금융통화위원회 회의실3. 출석위	2007.9.20	
1.hwp금융통화위원회 의사록2007년도 제21차 회의1. 일 자 2007년 10월 11일 (목) 2. 장 소 금융통화위원회 회의실3. 출석	2007.10.11	
3.hwp금융통화위원회 의사록2007년도 제23차 회의1. 일 자 2007년 11월 8일 (목) 2. 장 소 금융통화위원회 회의실3. 출석위	2007.11.8	
4.hwp금융통화위원회 의사록2007년도 제24차 회의1. 일 자 2007년 11월 22일 (목)2. 장 소 금융통화위원회 회의실3. 출석위	2007.11.22	
5.hwp금융통화위원회 의사록2007년도 제25차 회의1. 일 자 2007년 12월 7일 (금)2. 장 소 금융통화위원회 회의실3. 출석위	2007.12.7	

3

채권분석보고서 크롤링

- 채권 분석 리포트 3458개
- 날짜, 증권사, 내용으로 이루어진 csv 생성

3447	08.05.21	대우증권	Microsoft Word - Daewoo Bond Brief_0521.doc 1 ...
3448	08.05.20	대우증권	Microsoft Word - 0520_Daewoo Bond Brief.doc 1 ...
3449	08.05.19	대우증권	Microsoft Word - new Bond Brie_080519.doc 1 Da...
3450	08.05.13	대우증권	Microsoft Word - FixedIncome_0513_.doc 레인지의...
3451	08.05.04	대우증권	ㅎㅎㅎ2008_05 월간채권투자채 권 시 장 전 망금 융 시 장 차 트 북機會의 ...
3452	08.04.28	대우증권	Microsoft Word - FixedIncome_0428.doc FOMC ...
3453	08.04.21	대우증권	Microsoft Word - 920_0421_Fixed Income Weekly....
3454	08.04.14	대우증권	Microsoft Word - Fixedincome0414.doc 4월 금통위...
3455	08.04.07	대우증권	Microsoft Word - 0407.docFixed Income Weekly 2...
3456	08.04.07	대우증권	Microsoft Word - 0407.docFixed Income Weekly 2...
3457	08.04.01	대우증권	ㅎㅎㅎ2008_04 월간채권투자채 권 시 장 전 망금 융 시 장 차 트 북對應과 ...

3458 rows × 3 columns

이슈

▪ 의사록 PDF, HWP

의사록에 같은 파일임에도 PDF, HWP 두가지로 나뉘어서 올라가 있는 경우가 있음

- 텍스트 전처리 단계

1. eKoNLPy를 이용하여 n-gram 및 토큰화
2. 토큰의 불용어 처리
3. n-gram과 토큰 통합
4. 콜 금리 크롤링 후 전처리
5. 콜 금리 라벨링

1 n-gram 추출 및 라벨링

- date, n-gram, label 로 이루어진, 536457개의 행을 가진 csv 생성

```
In [41]: 1 tot_df
```

266517	2011-06-02	경제/NNG,지표/NNG,실망/NNG,나타나/VV,미/NNG,달러/NNG,대거/MA...	상승
266518	2011-06-02	유로/NNG,뉴욕/NNG,외환시장/NNG,경제/NNG,지표/NNG,약화/NNG,fe...	상승
266519	2011-06-03	달러/NNG,뉴욕/NNG,외환시장/NNG,고용/NNG,지표/NNG,약화/NNG,fe...	하락
266520	2011-03-30	엔/NNG,글로벌/NNG,경제/NNG,회복/NNG,기대/NNG,달러/NNG,유로/N...	상승
266521	2011-06-06	뉴욕/NNG,유로/NNG,외환시장/NNG,단기/NNG,급등/NNG,따르/VV,매물/...	하락
266522	2011-06-07	달러/NNG,외환/NNG,영향/NNG,받/VV,통화/NNG,일제히/MAG,약세/NN...	상승
266523	2011-06-07	못난이/NNG,형제/NNG,콘테스트/NNG,달러/NNG,파운드/NNG,유로/NNG,...	상승
266524	2011-05-28	달러/NNG,뉴욕/NNG,외환시장/NNG,미/NNG,소프트패치/NNG,가능성/NNG...	상승
266525	2011-06-07	달러/NNG,뉴욕/NNG,외환시장/NNG,중국발/NNG,악재/NNG,불구/NNG,엔...	상승
266526	2011-06-09	유로/NNG,뉴욕/NNG,외환시장/NNG,안전통화/NNG,선호/NNG,현상/NNG,...	하락
266527	2011-06-09	유로/NNG,뉴욕/NNG,외환시장/NNG,세계/NNG,경제/NNG,둔화/NNG,부채...	하락

536457 rows × 3 columns

이슈

- tokenize할 때 ngram과 token을 동시에 가져왔어야 했는데 처음에 24시간 이상 투자해서 ngram만 가져옴. 나중이 되어서야 token의 필요성을 인지하고 또 약 18시간 정도 투자

- 극성 분석 단계

1 NBC로 극성 분류

2 사전 제작

2 극성 사전 제작

- 18233개의 N-gram으로 이루어진 업 사전, 17563개 N-gram으로 이루어진 다운 사전 생성

긍정

18218	18218	파산/NNG;늘/VV
18219	18219	악순환/NNG;차단/NNG
18220	18220	금리/NNG;하락/NNG;한계/NNG
18221	18221	지출/NNG;오르/VV
18222	18222	fed/NNG;양적완화/NNG;확대/NNG
18223	18223	금리/NNG;인상/NNG;금리/NNG;상승/NNG
18224	18224	맹수/NNG
18225	18225	패널조사/NNG
18226	18226	기업/NNG;이익/NNG;감소/NNG
18227	18227	막대/NNG;재정/NNG;적자/NNG
18228	18228	배재/NNG
18229	18229	계절적/VAX
18230	18230	도청/NNG
18231	18231	가져야/VV
18232	18232	게/NNG

18233 rows × 2 columns

부정

17549	17549	성장/NNG;둔화/NNG;부진/NNG
17550	17550	디플레이션/NNG;어렵/VA
17551	17551	집/NNG;마련/NNG;어렵/VA
17552	17552	가계/NNG;소득/NNG;오르/VV
17553	17553	미약/NNG;경기/NNG;개선/NNG
17554	17554	자산시장/NNG;개선/NNG
17555	17555	키프로스사태/NNG
17556	17556	저유가/NNG;긍정적/VAX
17557	17557	경제/NNG;고용/NNG;증가/NNG
17558	17558	제조업/NNG;고용/NNG;둔화/NNG
17559	17559	가계/NNG;부채/NNG;총량/NNG;늘/VV
17560	17560	구제금융법안/NNG
17561	17561	수출/NNG;부진/NNG;내수/NNG;둔화/NNG
17562	17562	도철/NNG

17563 rows × 2 columns

이슈

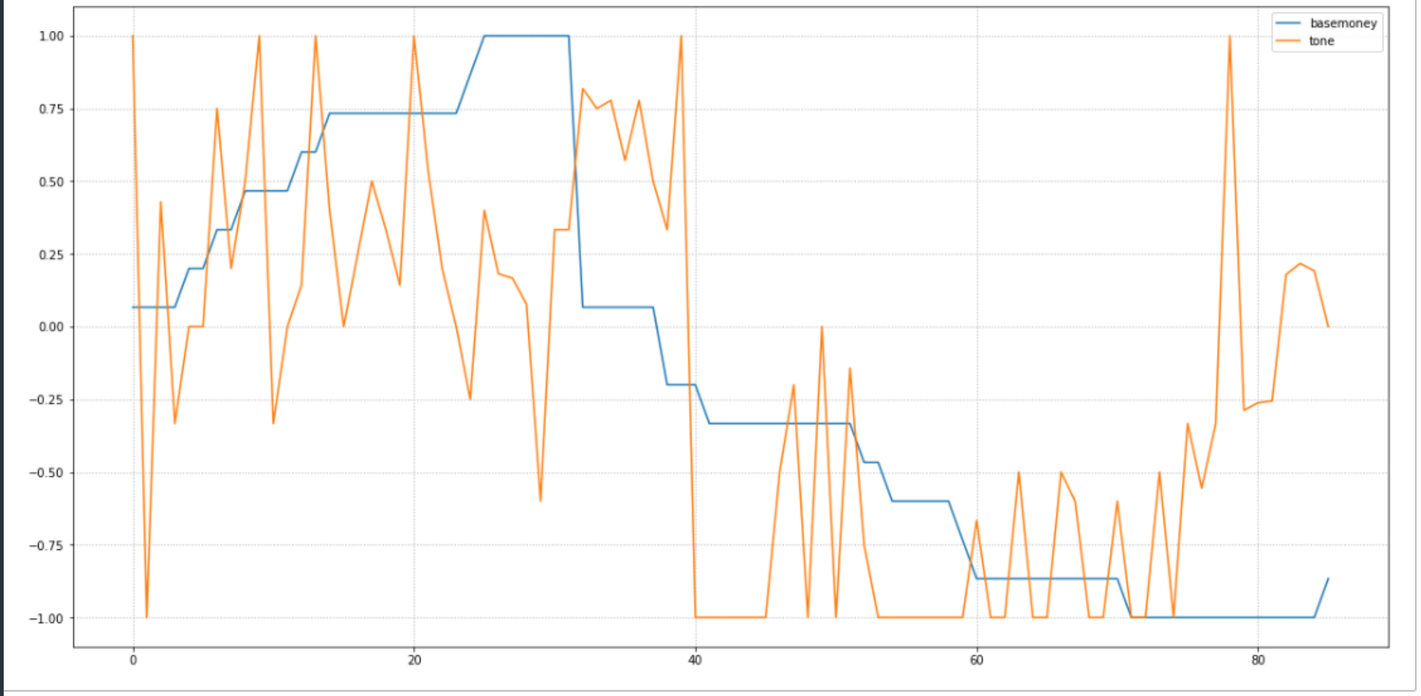
- 15개 이하 자르기에 관하여
  - 콜 라벨 중립인 값들까지 합친 값이 15개 미만일 때 잘라내는가?
  - 중립은 먼저 쳐내고, 상승/하락만 따져서 15개 미만일 때 잘라내는가?

- 감성 분석 단계

1. 의사록 특정 섹션 문장들의 ngram 생성, csv 만듦
2. 문장 별로 감성 분석,
   문장마다 tone score를 계산
3. 문장을 합쳐 total ton score를 계산
4. 그래프를 그리고 기준 금리 그래프와 비교, 상관관계 비교

1 그래프 그림과 정확도

```
plt.figure(figsize=(20, 10))
plt.plot(forgraph['basemoney'])
plt.plot(forgraph['tone'])
plt.legend(loc=0)
plt.grid(True, color='0.7', linestyle=':', linewidth=1)
```




```
corr = df.corr(method = 'pearson')
corr
```

	tone	basemoney
tone	1.00000	0.54348
basemoney	0.54348	1.00000

그 외의 이슈

- 함수화를 하지 않아 코드에서 틀린 부분을 찾지 못함

```
minute_df['tone'] = 0
count_p = 0
count_n = 0

for i in range(len(minute_df)):
    ngrams = minute_df.loc[i, 'n-gram'].split(',')

```

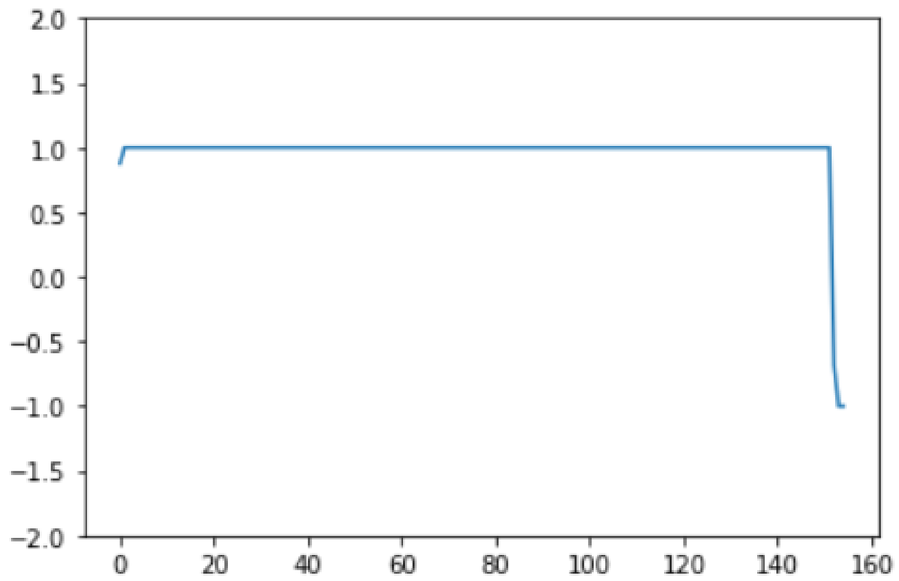
```
minute_df['tone'] = 0


for i in range(len(minute_df)):
    ngrams = minute_df.loc[i, 'n-gram'].split(',')

    count_p = 0
    count_n = 0

    for ngram in ngrams:
```

그 외의 이슈

```
1  import matplotlib.pyplot as plt
2  %matplotlib inline
```

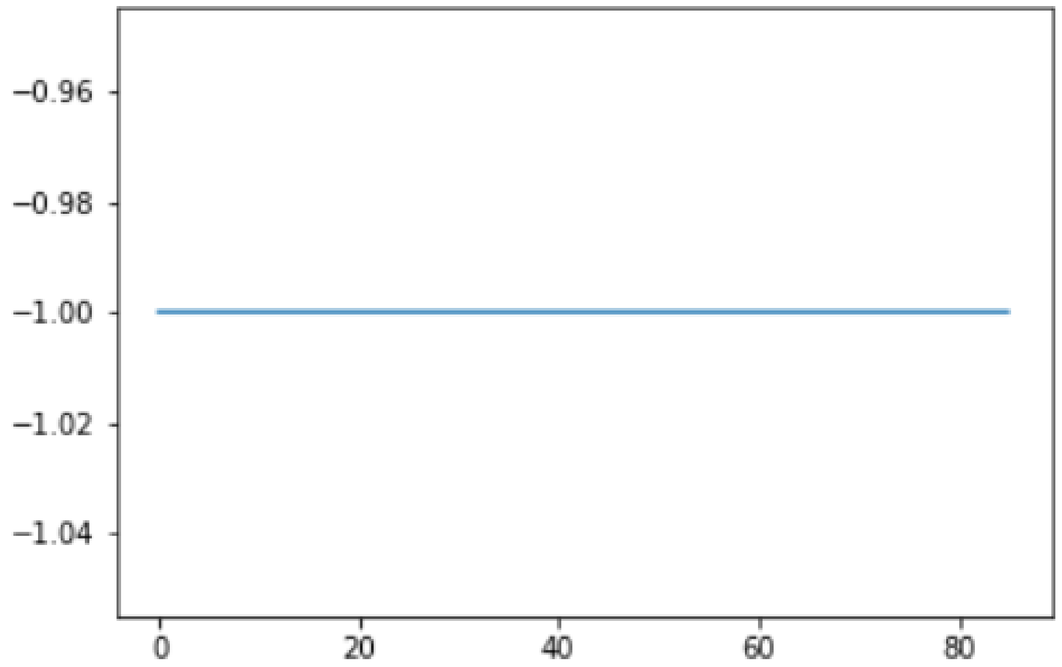
```
1  plt.plot(tone_list)
2  plt.ylim((-2.0,2.0))
```

```
(-2.0, 2.0)
```



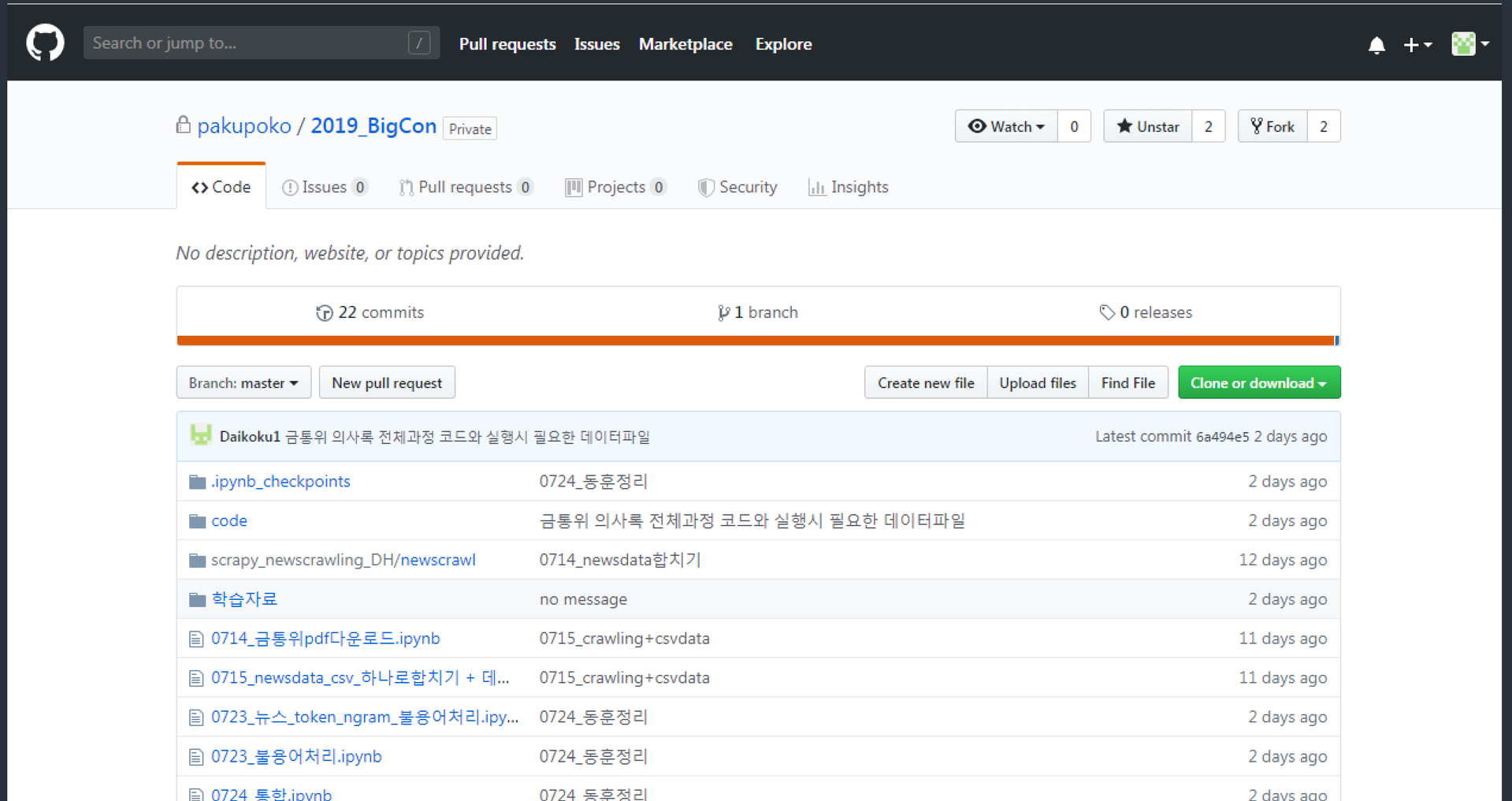

```
1  plt.plot(tone_list)
2  # plt.ylim((-2.0,2.0))
```

```
[<matplotlib.lines.Line2D at 0x1eba8e8efd0>]
```



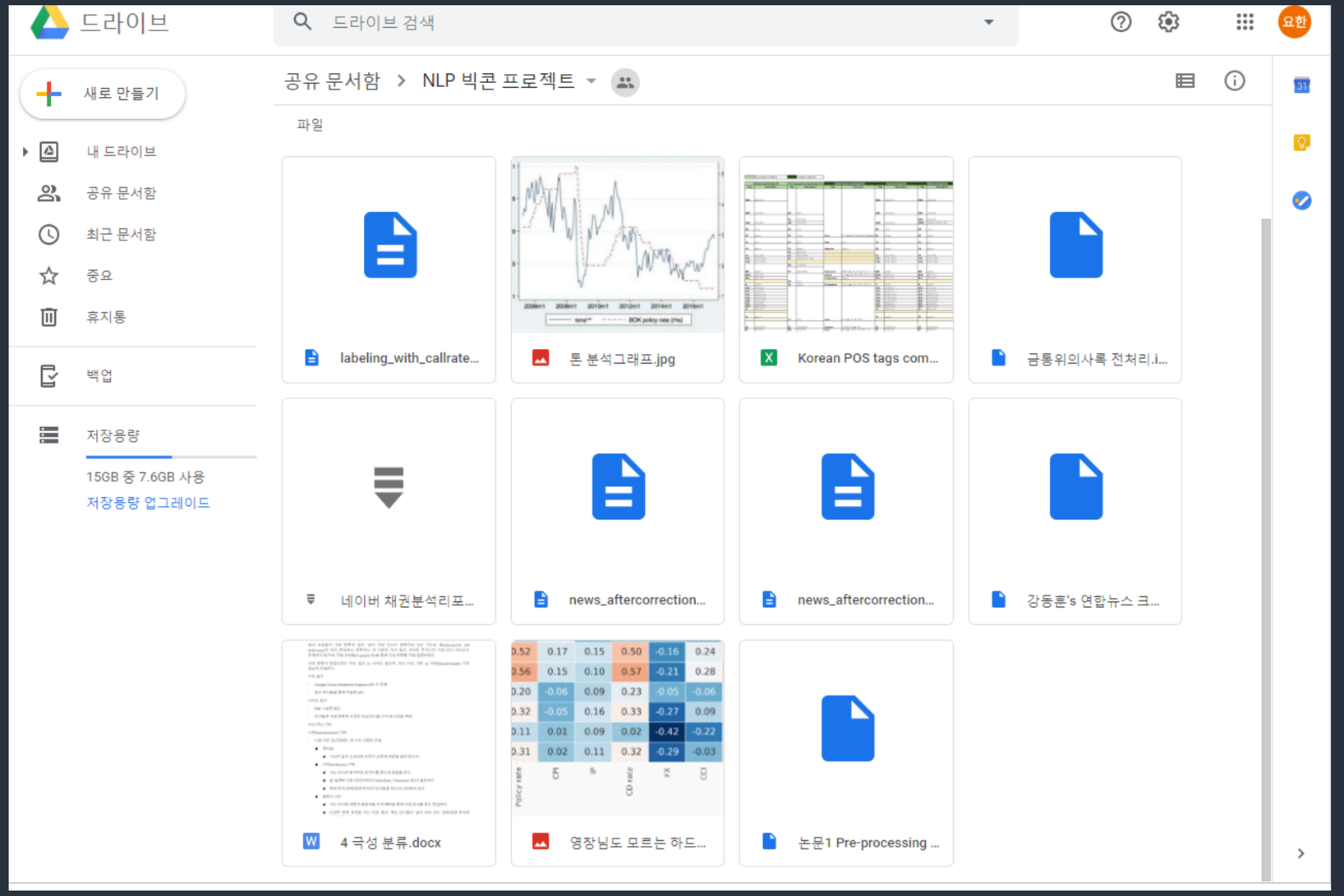
그 외의 이슈

- 환경 활용 미흡 - 깃허브



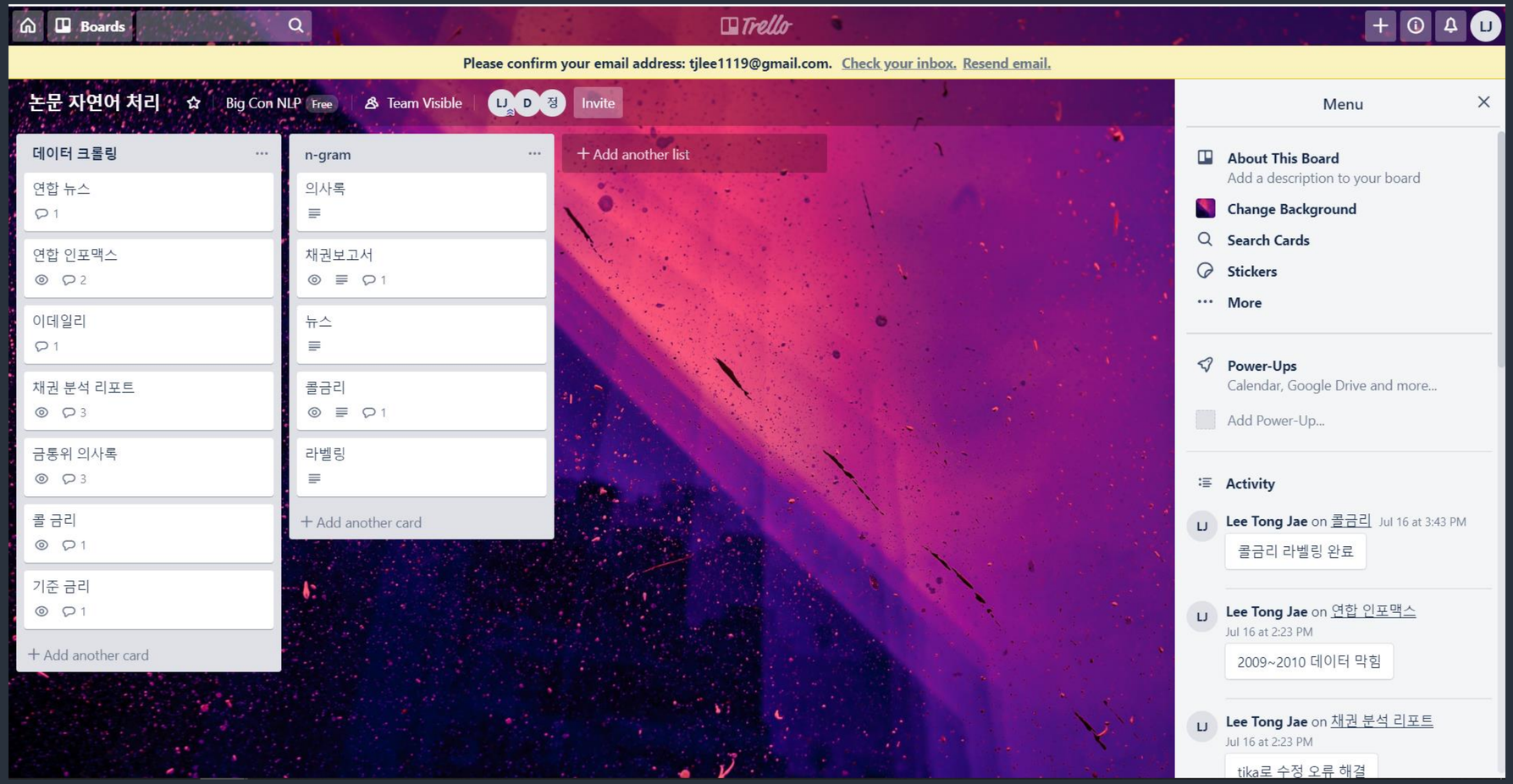
그 외의 이슈

- **환경 활용 미흡**
  - 구글 드라이브



그 외의 이슈

- 환경 활용 미흡
  - 트렐로



감사합니다