텍스트 마이닝을 활용한 금통위 의사록 분석

Deciphering Monetary Policy Board Minutes through Text Mining Approach : The Case of Korea

Team 3

김형석 송지혜 이진아 이하영 홍현택

CONTENTS

Ⅰ. Introduction

Ⅱ. Literature Review

Ⅲ. Data and Methodology

Ⅳ. Empirical Analysis

Ⅴ. Concluding Remarks

I . Introduction

DATA로서 TEXT ?

숫자 데이터에 비해 정량화, 해석의 어려움

해결책은 텍스트 마이닝

TEXT MINING 적용방법

통화정책

의사록

정량적인 지표로
변환하여 어조 분석

어휘기반의 지표가 현재와 미래 통화정책을 설명

분야별 사전(field-specific dictionary)과 한국어 원문 텍스트의 중요성

연구의 의의

1. 감정분석을 통화정책에 적용한 첫 번째 연구

2. 텍스트 마이닝 분야의 최신 방법론을 사용
   - N-gram
   - 시장 접근법과 어휘 접근법 활용

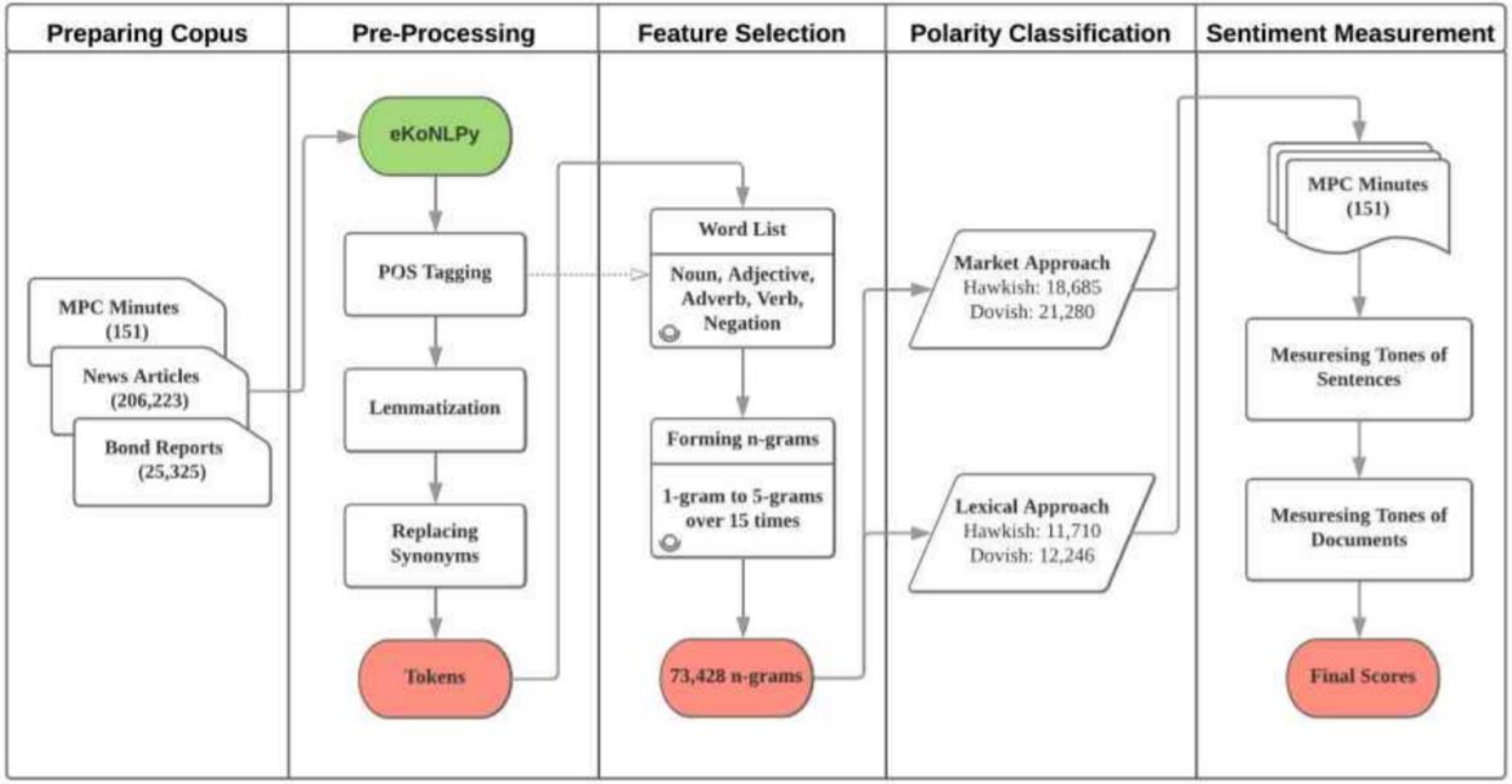
3. NLP Python Library인 eKoNLPy
   - 한국어의 언어적 특성으로 인한 문제를 해소
   - 특정분야(경제, 금융)의 텍스트마이닝 연구에 유용

Ⅱ. Literature Review

Ⅲ. Data and Methodology

Data and Methodology

Figure 1. Procedure of Sentiment Analysis



Note: This figure, along with table 1, summarizes our discussion in section 3.

1. Preparing the Corpus

MPB의사록 외에 많은 문서를 활용하여 필드 특정 어휘집(field-specific lexicons)을 만든다.

1. **금통위 의사록(MPB Minutes)**
   - MPB 의원들의 경제상황, 외환 및 국제 금융, 금융시장 및 통화정책에 관한 토론 요약
   - 통화 정책결정에 대한 의원들의 견해 기록
2. **뉴스 기사(News Article)**
   - 경제, 통화정책, 금융시장, 한은의 미래 통화정책 기조에 대한 대중의 인식 파악
3. **채권 애널리스트 분석 보고서(Bond Analysts' Reports)**
   - 통화정책과 채권시장에 대한 전문가의 견해
   - 뉴스기사에 비해 비정형화된 작문스타일을 사전에 포함

Document type	No. of docs	Average no. of sentences	Max no. of sentences
금통위 의사록	151	165	326
뉴스 기사	206,223	15	340
채권 애널리스트 보고서	25,325	49	2,515
Total	231,699	19	2,515

2.1 Typical Steps of Pre-processing

전처리에는 토큰화(tokenization)와 정규화(normalization)가 포함된다.

1. **토큰화(tokenization)**
   - 문서와 문장의 긴 문자열을 토큰으로 분할
   - 품사태깅(POS)을 진행

2. **정규화(normalization)**
   - 정규화는 텍스트를 단일 정규형으로 변환
   - 원형 복원(Stemming, Lemmatization), 불용어처리(Stopword) 진행

2.2 Korean NLP Python Library for Economic Analysis(eKoNLPy)

경제, 금융 분석에 특화된 eKoNLPy 라이브러리를 사용한다.

No	문제점
1	띄어쓰기 문제, 영어와 달리 후치사가 공백으로 구분되지 않으며 공백 규칙이 엄격하게 준수되지 않음.
	a week ago : ‘ago’는 후치사 / in a week : ‘in’은 전치사 동구 밖 과수원 길은 아름답다. : ‘밖’ 과 ‘은’ => 영어와 다르게 공백규칙이 없음
2	표준 외래어 표기법을 따르지 않는 외국어가 존재
	naive-- 바른표기 : 나이브하다 / 잘못된 표기 : 네이브하다
3	‘인플레이션’, ‘인플레’, ‘물가’ 처럼 같은 의미인 단어에 대해서 여러 가지 표기법이 존재
4	많은 동사와 형용사가 불규칙하게 결합
	벗다-벗고-벗으니-벗어서- [규칙 활용] 짓다-짓고-지으니-지어서- [불규칙 활용] (땅에)묻다-묻고-묻으니-묻어서- [규칙활용] (남에게)묻다-묻고-물으니-물어서- [불규칙 활용]

2.2 Korean NLP Python Library for Economic Analysis(eKoNLPy)

eKo(nomic)NLPy은 경제 분석을 위한 Korean NLP Python Library

1. KoNLPy의 Mecab tagger를 기반으로 경제관련 전문용어, 금융기관, 기업명 등을 하나의 명사로 분류하는 후처리 기능
2. add_dictionary를 통하여 str 혹은 list of str 형식의 단어를 사전에 추가
3. 통화정책(Monetary Policy)의 어조(Hawkish/Dovish)를 판단할 수 있는 Sentiment Analysis 기능
4. 경제의 불확실성(Uncertain/Stable)을 판단할 수 있는 Economic Uncertainty Analysis 기능
5. 경제 문서의 주제를 분류할 수 있는 Topic Analysis 기능
6. 분류 정확도가 낮은 문장을 neutral로 분류하는 강도를 설정(default: 1.3)

3. Feature Selection

금리와 관련된 어조를 포함한 단어(words)나 구(phrases)로 제한하는 작업이 필요

1. **단어의 벡터 차원을 줄여 처리 속도를 향상시킴**
2. **n-gram**

(1) 정확한 문맥 파악

ex) 느린 회복(회복 – 긍정 / 느린 회복 – 부정), 실업률 감소(실업률 – 부정 / 실업률 감소 – 긍정)

(2) overfitting과 curse of dimensionality(차원의 저주)

→ n=6이상일 경우 'overfitting'과 '차원의 저주' 문제가 존재

(3) explosion of dimension

→ 메모리 크기와 처리 속도로 인한 계산상 문제

→ 단어의 품사태깅(POS)을 명사(NNG), 형용사(VA, VAX), 부사(MAG), 동사(VA), 부정사로 제한

→15번 미만으로 나오는 n-gram은 제외

4. Polarity Classification

극성 사전이 따로 준비되지 않은 경우 Feature의 특징에 기반한 극성분류가 선행되야 함

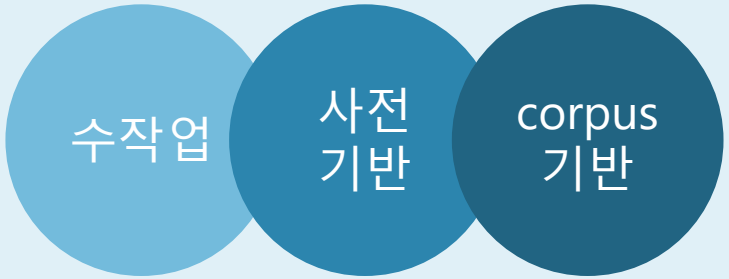
Machin learning-based

- 기계학습 기반의 접근
  - market approach
- Naïve-Bayes Classifier(NBC)

VS

Lexical-based

- 코퍼스 기반의 접근
  - lexical approach
- ngram2vec, SentProp



- 시드 단어(seed word)와 함께 발생하는 패턴을 통해 단어의 극성을 결
- 특정 도메인(경제, 금융)의 corpus를 활용해 분야 및 문맥 별 감정어의

4.1 Market Approach

기계학습(Machine-learning)중 하나인 단순확률 분류자, Naïve-Bayes Classifier(NBC)를 사용

- NBC는 모든 클래스의 Feature값(hawkish, dovish)이 독립적이라고 가정
- 학습을 통해 얻은 각 Feature의 조건부확률값을 극성 점수로 사용

$$polarity\,score = \frac{p(feature|hawkish)}{p(feature|dovish)} = \frac{p(feature\&hawkish)/p(hawkish)}{p(feature\&dovish)/p(dovish)} \quad (1)$$

(1) 콜금리의 1개월 변동이 발표된 날에 ▶ 콜금리가 상승할 경우 positive(hawkish)
▶ 콜금리가 하락할 경우 negative(dovish)으로 라벨링

(2) 라벨링된 문장(4백만 문장 이상)을 무작위(randomly)로 9 : 1의 비율로 train set과 test set으로 나눔

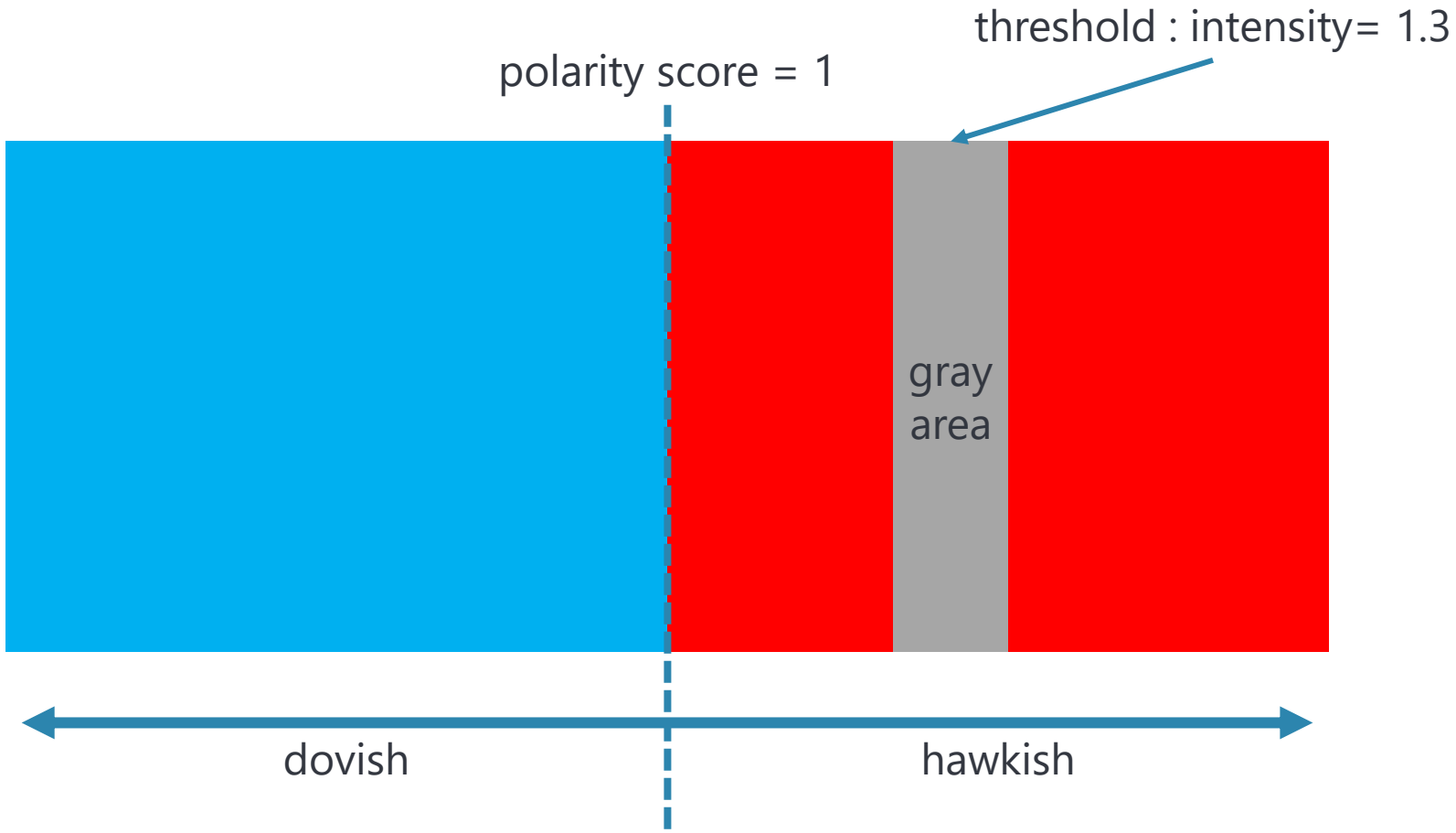
(3) 각 문장마다 5-gram(1-gram ~ 5-gram)을 feature로 사용하여 분류자를 훈련시키고 정확성을 검사

(4) 훈련된 NBC는 각 feature(hawkish, dovish)의 조건부 확률을 산출하며, 이 값을 feature의 극성 점수로 사용

(5) 이 절차를 30 번 반복하여 얻어진 극성 점수의 평균을 최종값으로 사용(bagging)

4.1 Market Approach

극성 분류



* threshold(임계값) : 가설 검정에서 기각역과 채택역의 경계가 되는 값

** gray area : 분류가 모호한 단어들의 모음

2.2 Korean NLP Python Library for Economic Analysis(eKoNLPy)

eKo(nomic)NLPy은 경제 분석을 위한 Korean NLP Python Library

1. KoNLPy의 Mecab tagger를 기반으로 경제관련 전문용어, 금융기관, 기업명 등을 하나의 명사로 분류하는 후처리 기능
2. add_dictionary를 통하여 str 혹은 list of str 형식의 단어를 사전에 추가
3. 통화정책(Monetary Policy)의 어조(Hawkish/Dovish)를 판단할 수 있는 Sentiment Analysis 기능
4. 경제의 불확실성(Uncertain/Stable)을 판단할 수 있는 Economic Uncertainty Analysis 기능
5. 경제 문서의 주제를 분류할 수 있는 Topic Analysis 기능
6. 분류 정확도가 낮은 문장을 neutral로 분류하는 강도를 설정(default: 1.3)

4.2 Lexical Approach

동일한 문맥에서 두 단어가 자주 함께 나타나면 같은 극성을 가질 가능성이 높음

다른 단어와의 공존 빈도를 계산하여 알 수 없는 단어의 극성을 결정

1. **동시발생에 근거하여 극성 판단할 경우 반의어를 인식하지 못하는 문제 발생**

   → ngram2vec을 사용

2. **결과가 시드단어(seed words)의 선택에 영향을 받는 문제 발생**

   → 도메인 별 감정 사전인 SentProp을 통해 시드단어(seed words)를 bootstrap

3. **학습**

   (1) 전체 코퍼스(corpus) 232,658건의 문서를 사용하여 ngram2vec을 학습

   (2) 학습에 사용된 parameter

   → center words(5-gram), context words(5-gram), window size = 5, negative sampling size = 5, vector representation = 300 dimension

3. Feature Selection

금리와 관련된 어조를 포함한 단어(words)나 구(phrases)로 제한하는 작업이 필요

1. **단어의 벡터 차원을 줄여 처리 속도를 향상시킴**
2. **n-gram**

(1) 정확한 문맥 파악

ex) 느린 회복(회복 – 긍정 / 느린 회복 – 부정), 실업률 감소(실업률 – 부정 / 실업률 감소 – 긍정)

(2) overfitting과 curse of dimensionality(차원의 저주)

→ n=6이상일 경우 'overfitting'과 '차원의 저주' 문제가 존재

(3) explosion of dimension

→ 메모리 크기와 처리 속도로 인한 계산상 문제

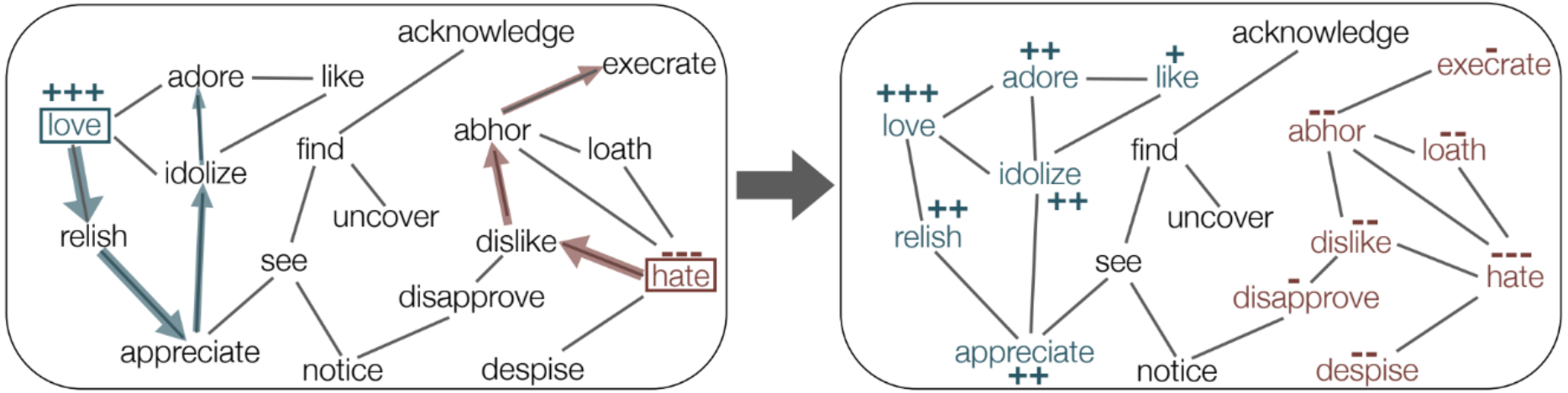
→ 단어의 품사태깅(POS)을 명사(NNG), 형용사(VA, VAX), 부사(MAG), 동사(VA), 부정사로 제한

→15번 미만으로 나오는 n-gram은 제외

4.2 Lexical Approach

극성 분류

1. 단어의 시드 세트(seed set of words)와 n-gram을 벡터공간에 배치하고 시드에 대한 n-gram의 근접성을 측정
2. n-gram의 극성은 해당 n-gram에 hitting하는 시드세트의 random-walk 확률에 비례
3. 각각의 feature는 매파, 비둘기파 두 가지 확률을 가짐
4. 최종 극성 점수는 방정식 (1)과 같이 상대적 비율



a. Run random walks from seed words. **b. Assign polarity scores based on frequency of random walk visits.**

Figure 3: Visual summary of the SENTPROP algorithm.

4.2 Lexical Approach

학습 결과

Table 5: Seed Words for Polarity Induction

Posive		Negative	
높/VA	팽창/NNG	낮/VA	축소/NNG
인상/NNG	매파/NNG	인하/NNG	비둘기/NNG
성장/NNG	투기/NNG;억제/NNG	둔화/NNG	악화/NNG
상승/NNG	인플레이션/NNG;압력/NNG	하락/NNG	회복/NNG;못하/VX
증가/NNG	위험/NNG;선호/NNG	감소/NNG	위험/NNG;회피/NNG
상회/NNG	물가/NNG;상승/NNG	하회/NNG	물가/NNG;하락/NNG
과열/NNG	금리/NNG;상승/NNG	위축/NNG	금리/NNG;하락/NNG
확장/NNG	상방/NNG;압력/NNG	침체/NNG	하방/NNG;압력/NNG
긴축/NNG	변동성/NNG;감소/NNG	완화/NNG	변동성/NNG;확대/NNG
흑자/NNG	채권/NNG;가격/NNG;하락/NNG	적자/NNG	채권/NNG;가격/NNG;상승/NNG
견조/NNG	요금/NNG;인상/NNG	부진/NNG	요금/NNG;인하/NNG
낙관/NNG	부동산/NNG;가격/NNG;상승/NNG	비관/NNG	부동산/NNG;가격/NNG;하락/NNG
상향/NNG	(Total 25 seeds)	하향/NNG	(Total 25 seeds)

4.3 Evaluation

정확성(accuracy) 판단 기준 → 어조(sentiment)의 분류가 사람의 판단과 얼마나 일치하는가

1. **어휘집 구축에 사용되지 않은 문서(documents)를 사용하여 정확성 평가**

   (1) 한국은행(BOK) 총재의 기자간담회 문서 (2009년 5월 ~ 2018년 1월)

   (2) 수동으로 2,341개 문장을 [매파, 중립, 비둘기파]로 분류

   (3) 60%의 train set으로 분류해 NBC를 학습시키고 40%의 test set으로 정확성 테스트

2. **위의 작업을 bagging(30번)한 결과, 평균 정확도는 약 86%**

3. **극성분류 방법에 따른 정확도**

   (1) 시장 접근법(market approach)의 정확도는 68%

   (2) 어휘 접근법(lexical approach)의 정확도는 67%

5. Measuring Sentiments

개별 문장과 문서의 어조(tone) 측정

1. 개별 문장의 어조(tone) 측정

   비둘기파(dovish)와 매파(hawkish) feature의 개수를 기준으로 문장의 어조($tone_s$)를 계산

$$tone_s = \frac{No.\ of\ hawkish\ features - No.\ of\ dovish\ features}{No.\ of\ hawkish\ featuers + No.\ of\ dovish\ features}$$

$tone_{mkt}$: market approach 기반 사전을 사용하여 측정한 점수

$tone_{lex}$: lexical approach 기반 사전을 사용하여 측정한 점수

$tone_{nbc}$: Naïve-bayes classifier를 사용하여 측정한 점수

$tone_{ksa}$: 형태소 분석기 꼬꼬마(Kkma)를 사용하여 측정한 점수

5. Measuring Sentiments

개별 문장과 문서의 어조(tone) 측정

2. 문서의 어조(tone) 측정

문서를 구성하는 모든 문장의 어조 점수를 합산하여 문서의 어조($tone_i$)를 계산

비둘기파(dovish) : -1, 매파(hawkish) : +1 사이의 값을 가지는 연속확률변수

$$tone_i = \frac{No.\ of\ hawkish\ tone_{s,i} - No.\ of\ dovish\ tone_{s,i}}{No.\ of\ hawkish\ tone_{s,i} + No.\ of\ dovish\ tone_{s,i}}$$

IV. Empirical Analysis

QUESTION.

- ✓ 어휘 기반의 어조(tone)가 한국은행의 현재와 미래 통화정책 결정을 설명할 수 있나?
  - ✓ 특히 거시경제 변수로 설명할 수 없는 **추가적인 정보**를 가지고 있는가?

YES

IV. Empirical Analysis

(a) With macroeconomic variables

1. 통화정책의 tone의 측정

상관계수를 보고 두 변수의 상관관계 파악가능

-시장기반 Tone & 사전 기반 Tone

-시장기반 Tone & 기준금리

-시장기반 Tone & EPU

-시장기반 Tone & UI

-기준금리 & 인플레이션

-기준금리 & IP(산업총생산)

Ⅳ. Empirical Analysis

1. 통화정책 결정(MP Decisions)에 대한 설명

- 한국은행의 금리에 대한 결정과 의사록의 내용과의 관계를 알아보기 위해서 현재와 미래의 결정 에 대한 어휘적 기반의 tone의 설명력을 test
- Ordered Probit Model(종속변수가 범주형일 때 쓰는 모형)을 사용하여 거시경제적 변수, 다른 불확실성 지표, 어휘기반의 지표의 설명력을 비교
- 테일러 준칙에 따른 통화정책 기준금리의 기본 모델을 사용

2. 한국은행의 통화정책 결정에 대한 설명

$$\Delta \mathrm{M\ P}\ t + k = \rho \Delta \mathrm{M\ P} t + \gamma 1 \Delta(\pi t - \pi^*) + \gamma 2 \Delta(yt - yt^*) + \gamma 3 \Delta \pi te + \gamma 4 \Delta yte + \beta \mathrm{X} t + ut$$

K=1,2

- $(\pi_t - \pi^*)$: inflation gap, π^*: 목표 물가상승률 (2%)
- $(y_t - y^*)$: output gap, y^*: hpfilter로 뽑아낸 trend
- π_t^e: 기대 인플레이션 (한은조사)
- y_t^e: 경기종합지수 선행지수 순환변동치

- Xt는 잠재적으로 거시경제학 변수 이외에도 통화정책의 변화를 설명하는 데 도움을 줄 만한 변수
- 어휘기반의 지표와 경제적 불확실성 지수, 불확실성 지수
- 통화정책의 입장변화($\Delta \mathrm{MP}\ t$)를 대표하기 위해 금리의 변화($\Delta \mathrm{BOK}\ t$) or 통화정책결정($\mathrm{MPD}\ t$)

2. 한국은행의 통화정책 결정에 대한 설명

Table 7: Ordered Probit, Changes in BOK Policy Rate

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable: ΔBOK_t					
ΔBOK_{t-1}	1.893**	1.790**	-0.209	[illegible]	1.611*	1.296
	(0.622)	(0.632)	(0.736)	(0.797)	(0.642)	(0.725)
$\Delta(\pi_t - \pi^*)$	0.142	0.0163	-0.364	-0.490	-0.0690	0.0274
	(0.331)	(0.341)	(0.517)	(0.431)	(0.348)	(0.352)
$\Delta(y_t - y^*)$	7.068	5.614	6.025	8.351	5.696	4.803
	(4.362)	(4.634)	(5.298)	(5.160)	(4.660)	(4.764)
$\Delta\pi_t^e$		1.734	1.553	1.635	1.948*	1.883*
		(0.910)	(1.262)	(1.107)	(0.928)	(0.923)
Δy_t^e		0.322	0.153	0.0661	0.294	0.313
		(0.450)	(0.637)	(0.536)	(0.456)	(0.455)
$tone_t^{mkt}$			5.327***			
			(1.114)			
$tone_t^{iex}$				4.515***		
				(0.797)		
$EPU_t(Korea)$					-0.00374	
					(0.00191)	
$UI_t(Korea)$						-2.886
						(2.155)
N	143	143	[illegible]	143	143	133
pseudo R^2	0.076	0.095	0.446	0.364	0.116	0.107
	Dependent variable: [illegible]					

- 통화정책에 대한 입장의 변화를 금리의 변화(BOK)로 측정했을 때의 측정치 결과

- 시장기반분석의 어조와 어휘기반분석의 어조가 매우 중요하고 전반적인 경제적 불확실성 지표에 대한 측정치는 중요하지 않음

$tone^{mkt}$ or $tone^{iex}$

- 이러한 요소가 현저하게 결정계수 상승시켜 금리변화에 대한 설명력을 높힘

2. 한국은행의 통화정책 결정에 대한 설명

Table 8: Ordered Probit, Changes in MP Stance

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable: MPD_t					
MPD_{t-1}	0.759***	0.748***	-0.0467	-0.275	0.714**	0.583*
	(0.215)	(0.216)	(0.267)	(0.292)	(0.219)	(0.239)
$\Delta(\pi_t - \pi^*)$	0.109	0.0166	-0.336	-0.513	-0.0655	0.00875
	(0.331)	(0.340)	(0.512)	(0.433)	(0.346)	(0.352)
$\Delta(y_t - y^*)$	7.627	6.101	8.136	11.55*	6.247	5.913
	(4.634)	(4.917)	(5.603)	(5.705)	(4.961)	(5.210)
$\Delta\pi_t^f$		1.393	0.959	1.220	1.576	1.600
		(0.894)	(1.218)	(1.086)	(0.909)	(0.910)
Δy_t^f		0.355	-0.0585	-0.183	0.307	0.298
		(0.441)	(0.621)	(0.531)	(0.447)	(0.446)
$tone_t^{mkt}$			5.464***			
			(1.122)			
$tone_t^{lex}$				4.900***		
				(0.853)		
$EPU_t(Korea)$					-0.00364	
					(0.00189)	
$UI_t(Korea)$						-3.829
						(2.105)
N	143	143	143	143	143	133
pseudo R^2	0.095	0.109	0.461	0.397	0.128	0.130

- 종속변수를 비공식적인 한국은행의 정책(MPD)이라고 했을 때의 결과
  → 비슷한 결과 도출

- $tone^{mkt}$ and $tone^{lex}$ 가 결정계수를 상당히 높힘

- MPD_t

t시점 한국은행 기준금리 변경 크기에 따라 범주화

-1: 인하(<= -25*bp*,-0.25%)

0: 변경없음

+1: 인상(>= +25*bp*,+0.25%)

2. 한국은행의 통화정책 결정에 대한 설명

- 결과의 견고함을 확인하기 위해서

$$\Delta \hat{r}_{t+1} = 1.90\Delta r_t + 7.28\mathrm{IP}\ growth_t + 0.12\mathrm{CPI}_t,\ pseudo\ \mathrm{R}_2 = 0.08.$$

$$\Delta \hat{r}_{t+1} = -1.67\Delta r_t + 4.20\ tone_{tmkt} + 9.73\ \mathrm{IP}\ growth_t - 0.28\mathrm{CPI}_t,\ pseudo\ \mathrm{R}_2 = 0.37$$

- 만약 거시경제학적 변수가 다음 금리에 대한 관련있는 모든 정보를 포함한다면 x의 가중치)는 매우 작을 것
- 월별로 측정하기 위해 GDP성장률을 IP성장률로 대체한 후 ordered probit model을 적용

✓ 특정 분야의 사전을 사용하는 것이 중요한가?

✓ 한국어를 영어로 번역한 텍스트가 아닌 한국어 텍스트의 원문을 사용하는 것이 중요한가?

YES

2. 다른 텍스트 기반의 지표들과의 비교

1. $tone_{ksa}$

   : 서울대에서 발견한 데이터 시스템으로서 한국어 텍스트를 분석하기 위한 인기 있는 툴

   일반적인 목적의 사전으로 시장적 접근이나 어휘적 접근에서 사용했던 금융과 경제적 목적의 사전과 차이 존재

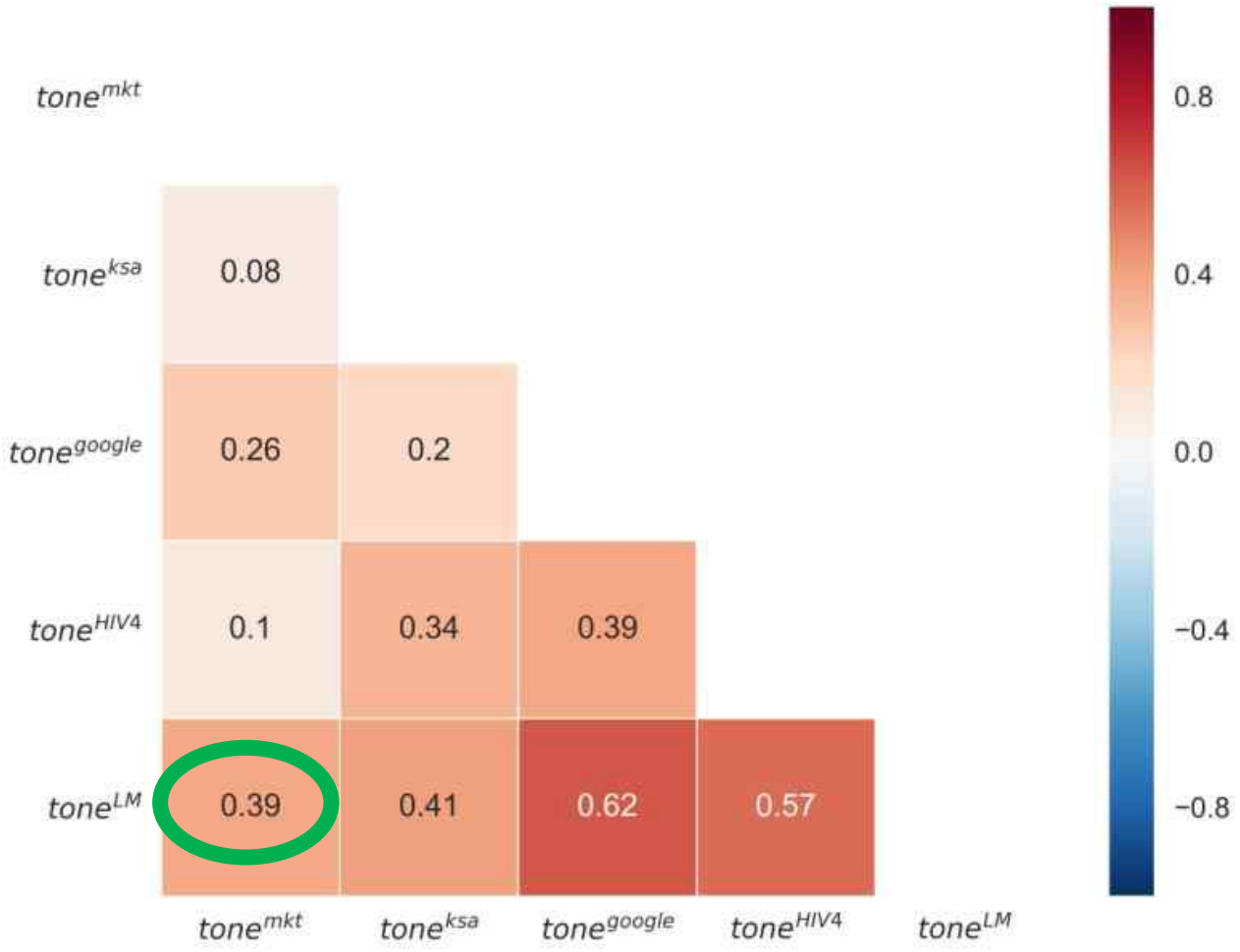
2. 영어기반의 텍스트 분석을 위해서 우리는 의사록을 구글 클라우드 번역기를 사용해 번역

   (1) $tone_{google}$: 감정분석 기능을 이용하여 어조를 측정

   (2) $tone_{HIV}$: 일반적인 목적의 하버드 사전에 기반

   (3) $tone_{LM}$: 특정 분야의 사전에 기반

2. 다른 텍스트 기반의 지표들과의 비교



(b) With other text-based indicators

IV. Empirical Analysis

2. 다른 텍스트 기반의 지표들과의 비교

결정계수를 비교해보면,

1. $tone^{mkt}$ 가 $tone^{ksa}$ 현재와 미래의 한국은행의 금리의 변화를 설명하는 데 있어 더 뛰어나다.
2. $tone^{LM}$ 가 $tone^{google}$ and $tone^{HIV4}$ 보다 뛰어나다.
3. $tone^{mkt}$ 가 $tone^{LM}$, $tone^{google}$, and $tone^{HIV4}$ 보다 뛰어나다.

1번과 2번은 특정 영역의 사전을 사용하는 것의 이점을 3번은 한국어 텍스트와 한국어를 영어로 번역한 텍스트로부터 나온 결과에 대한 것

V. Concluding Remarks

IV. Concluding remarks

1. 어떠한 종류의 정보가 한국은행의 금리와 거시경제학적인 변수와 비교되어야 하는지에 대해 알아보는 것은 중요하다
2. 우리의 측정은 중앙은행의 의사소통과 앞으로의 지도의 효율성을 위해 사용될 수 있다.
3. 우리의 방법론은 거시경제학적 불확실성과 미래의 통화정책 입장에 대한 대중의 기대, 주식 시장 감성 등을 측정하기 위한 다른 지표를 건설하는데 응용될 수 있다.

THANK YOU

텍스트 마이닝을 활용한 금통위 의사록 분석

Deciphering Monetary Policy Board Minutes through Text Mining Approach : The Case of Korea

Team 3

김형석 송지혜 이진아 이하영 홍현택

CONTENTS

Ⅰ. 프로젝트 배경

Ⅱ. 프로젝트 과정

Ⅲ. 프로젝트 결과

Ⅳ. 프로젝트 시사점

Ⅴ. 프로젝트 정리

Ⅰ. 프로젝트 배경

< 텍스트 마이닝을 활용한 금통위 의사록 분석 >

텍스트 마이닝

기준금리 예측

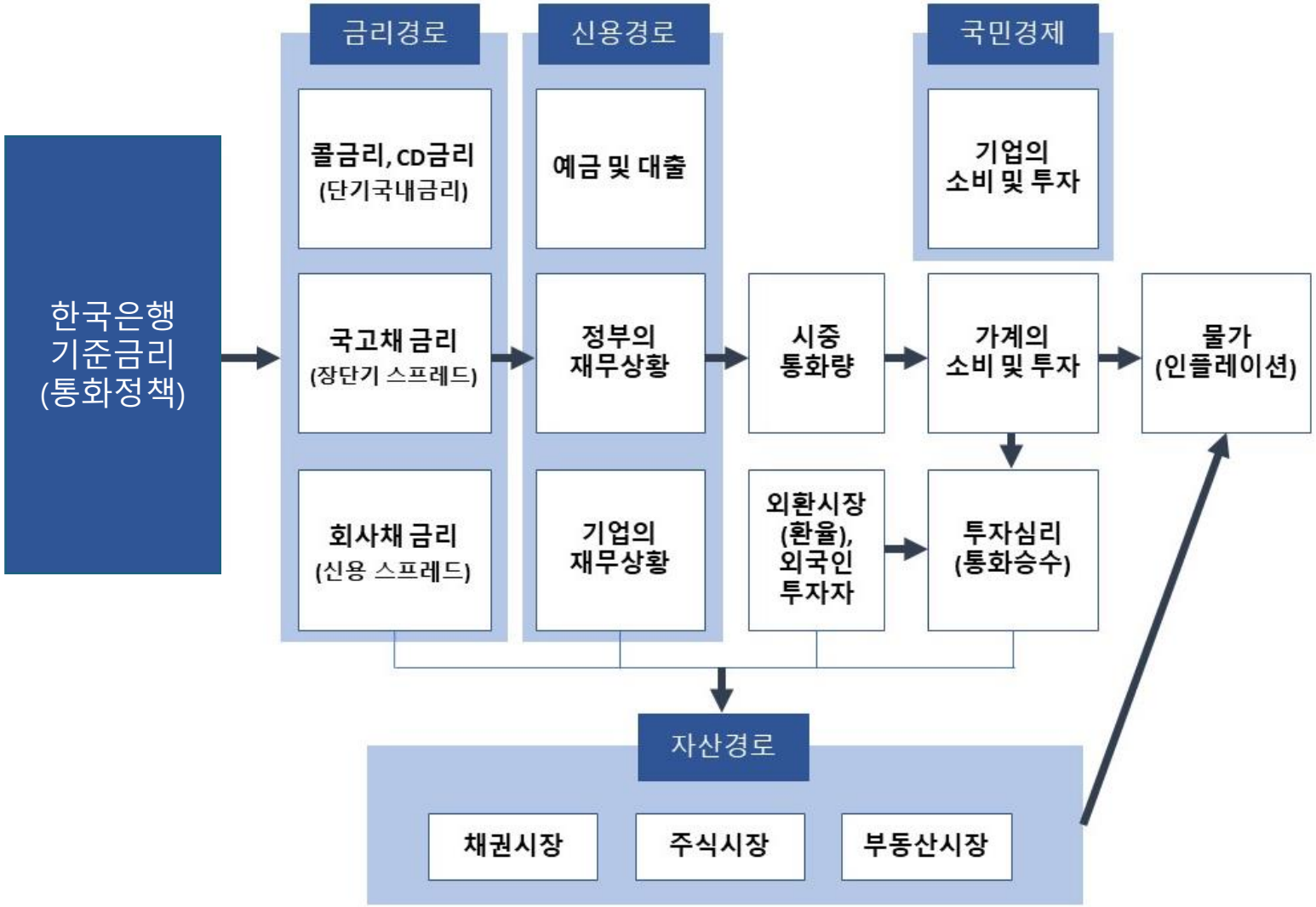
통화정책은 적시성과 신뢰성이 중요하다.

그러나 정책 결정자의 주관이 개입되기 때문에 언제나 불확실성을 내포하기도 한다

출처 : 세계일보

1.1 프로젝트 개요

통화정책 파급경로



< 프로젝트 목적 >

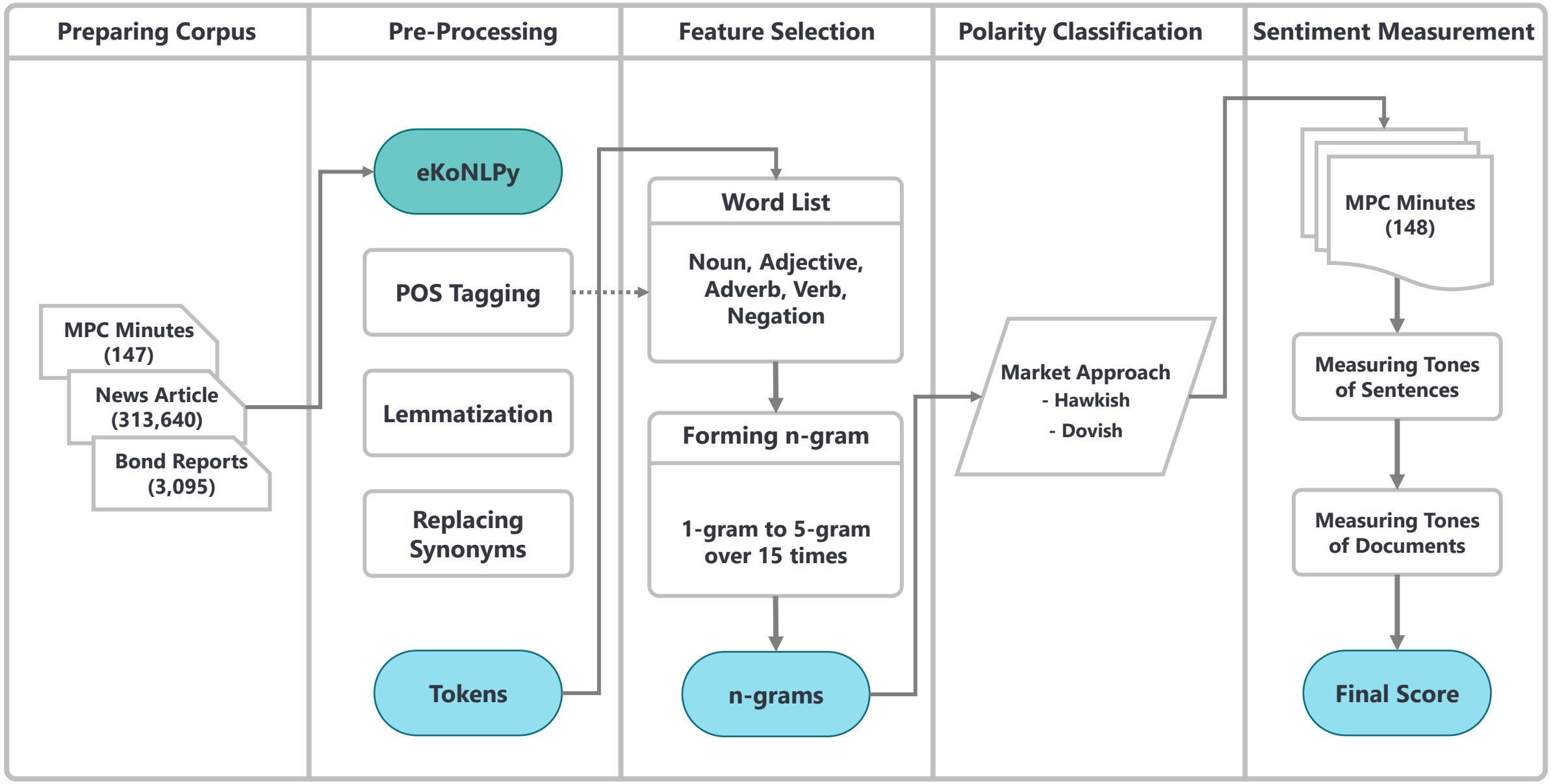
1. 데이터 처리부터 분석까지 논문을 충실히 따르며 직접 구현

2. 프로젝트 결과물의 활용 및 응용 가능성 검토

IDEAS
think?
good
CONNECT
click
People
Friends
Share
Inspiration
CHECK LIST:
cloud

II. 프로젝트 과정

2. Procedure of Sentiment Analysis



GitHub : https://github.com/nlpbokproject/jh

2.1 Preparing Corpus

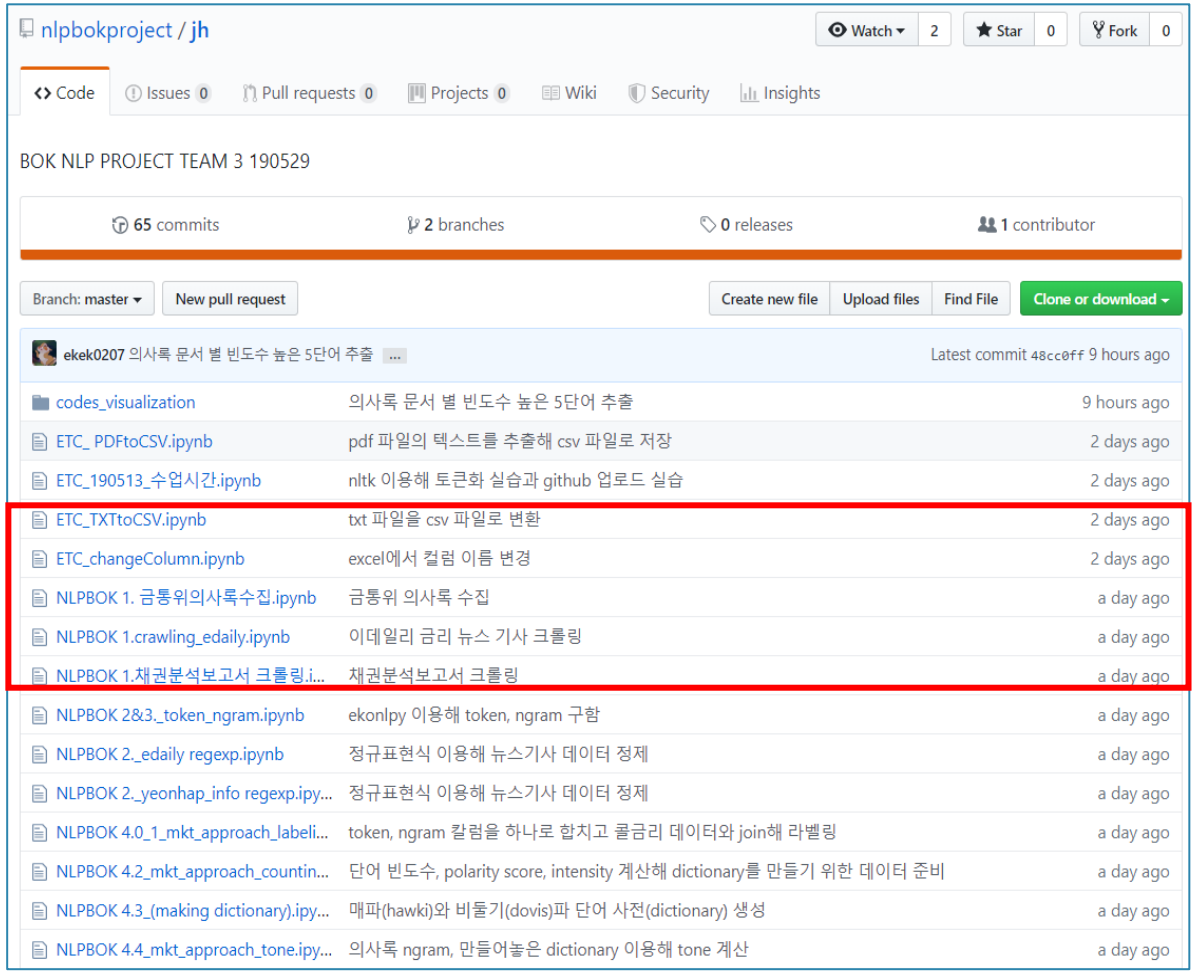
데이터 수집 과정

01 **데이터 수집**

Selenium, Beautiful Soup 활용

02 **데이터 저장**

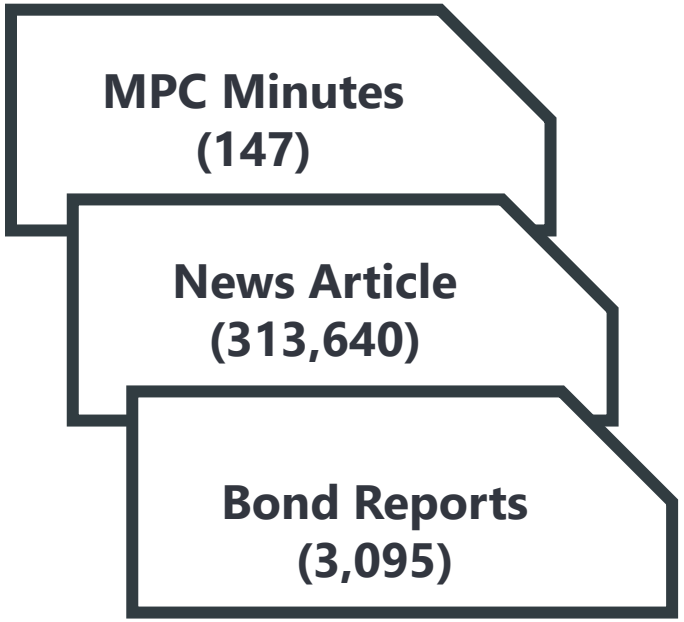
Data Frame 형태로 변환 후
corpus별 csv 형식으로 저장



2.1 Preparing Corpus

데이터 수집 과정

316,882건의 수집된 문서로 Filed-specific-lexicon 구성



MPC Minutes(147)

데이터스튜디오 : https://datastudio.google.com/open/16LHlInoEnt2JU5UBEm-NnJb5f__OEUrA

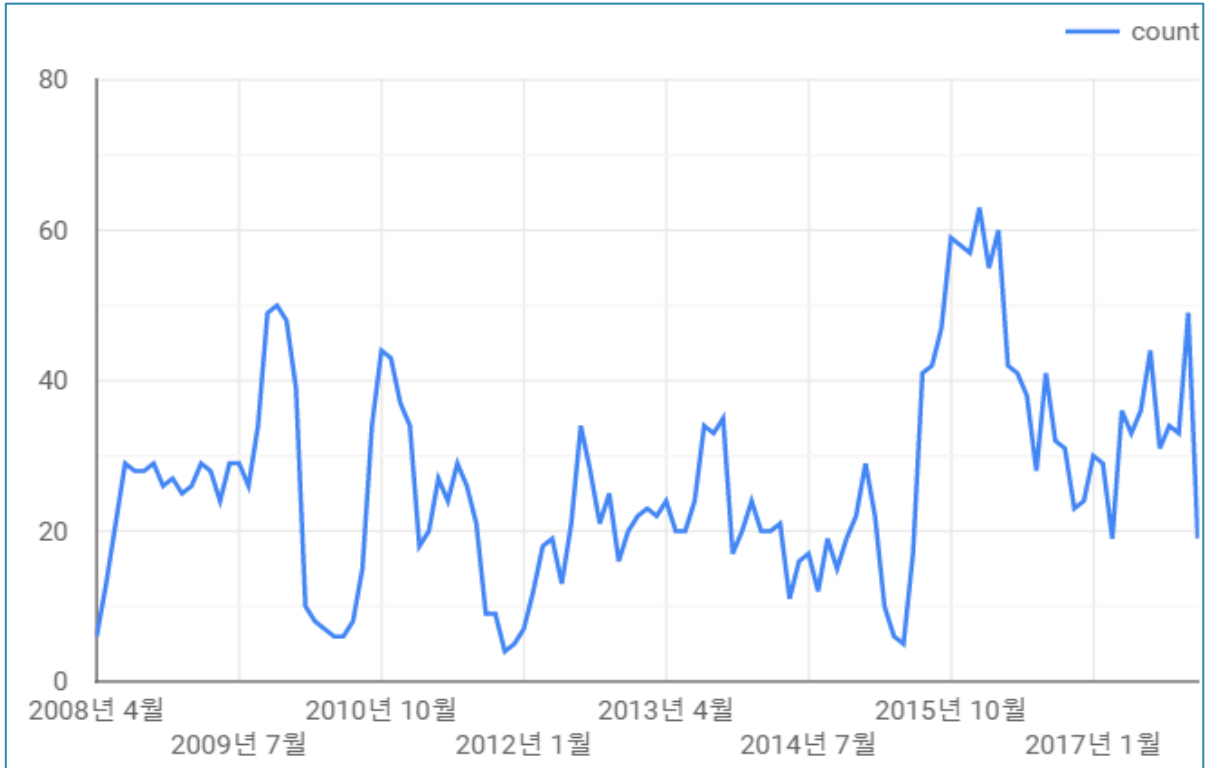
2.1 Preparing Corpus

금통위 의사록(MPB minutes)

활용 Data : 금융통화의원회 의사록

- 수집 출처 : 한국은행
- 기간 범위 : 2005년~2017년(13개년 데이터)
- 데이터 수 : MPB minutes 147건

	date	Economic Situation count	Foreign Currency count	Financial Markets count	Monetary Policy count	Participant Views count	Government View count	Foreign Currency ngram	Financial Markets ngram
0	20060810	0	0	0	0	78	0	NaN	NaN
1	20180412	74	69	0	0	400	0	fed/NNG;금리/NNG;인상/NNG;속도/NNG;가속/NNG,국제/NNG;금융시...	NaN
2	20070608	0	0	0	0	80	0	NaN	NaN
3	20070111	0	0	0	0	96	0	NaN	NaN
4	20051208	2	0	0	0	84	0	NaN	NaN
5	20170223	46	51	0	0	362	0	경제/NNG;성장률/NNG;잠재/NNG;성장률/NNG;상회/NNG,국제/NNG;금융...	NaN
6	20120209	3	0	0	0	194	0	NaN	NaN
7	20121109	25	12	15	0	318	0	외화차입/NNG;여건/NNG;양호/NNG,글로벌/NNG;유동성/NNG;증가/NNG,...	경기/NNG;하방/NNG;위험/NNG;증대/NNG,금융시장/NNG;변동성/NNG;완...
8	20180524	88	73	0	0	384	0	가계/NNG;대출/NNG;관리/NNG;강화/NNG,가계/NNG;대출/NNG;증가/N...	NaN
9	20060112	0	0	0	0	69	0	NaN	NaN
10	20181018	79	59	0	0	441	0	단기/NNG;금융시장/NNG;변동성/NNG;확대/NNG,fed/NNG;금리/NNG;...	NaN
11	20090910	0	0	0	0	10	0	NaN	NaN
12	20140410	43	12	22	0	362	0	유가/NNG,원자재/NNG;가격/NNG;하락/NNG,기대/NNG;시장/NNG;금리/	양적완화/NNG;축소/NNG;금리/NNG;인상/NNG,가계/NNG;부채/NNG;문



(a) Number of sentences in MPC minutes

데이터스튜디오 : https://datastudio.google.com/open/16LHlInoEnt2JU5UBEm-NnJb5f__OEUrA

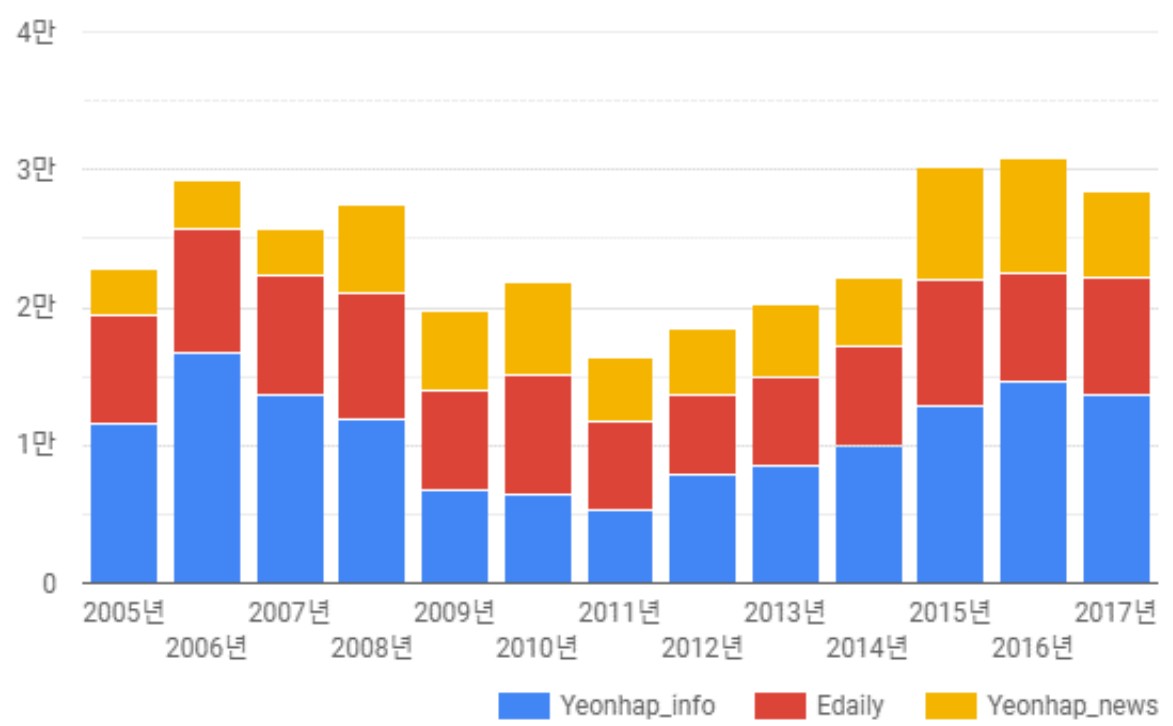
2.1 Preparing Corpus

뉴스 기사(News Article)

활용 Data : 뉴스 데이터 (키워드 : '금리')
- 수집 출처 : 이데일리, 연합뉴스, 연합인포맥스
- 기간 범위 : 2005년~2017년(13개년 데이터)
- 데이터 수 : 313,640건

종류	연합뉴스	연합인포	이데일리	총 합계
문서 수	72,149	139,974	101,517	313,640

	title	date	url	script
NaN	title	date	url	script
NaN	title	date	url	script
NaN	title	date	url	script
0.0	인도 모리 총리 신년사서 '화폐개혁' 정당성 주장	2017-01-01	http://www.edaily.co.kr/news/newspath.asp?news...	[이데일리 방성훈 기자] 인도의 모디 총리가 "부패, 검은 돈, 위조지폐는 인도의 ...
0.0	"초불확실성의 시대"…험난한 금리정책 예고	2017-01-01	http://www.edaily.co.kr/news/newspath.asp?news...	이주열 한국은행 총재가 지난달 15일 서울 중구 한은 본관에서 열린 금융통화위원회 ...
0.0	"초불확실성의 시대"…험난한 금리정책 예고	2017-01-01	http://www.edaily.co.kr/news/newspath.asp?news...	이주열 한국은행 총재가 지난달 15일 서울 중구 한은 본관에서 열린 금융통화위원회 ...
0.0	[새해 부동산 재테크]"금리 변수..부동산 투자는 보수적 접근 필요"	2017-01-01	http://www.edaily.co.kr/news/newspath.asp?news...	[이데일리 박태진 기자] 새해 부동산시장 전망은 그야말로 '안갯속'이다. 워낙 다양...
0.0	기업은행, '닭의 해 적금은 꼭이오 이벤트' 실시	2017-01-01	http://www.edaily.co.kr/news/newspath.asp?news...	[이데일리 김경은 기자] IBK기업은행은 새해를 맞아 오는 2일부터 2월말까지 영업...
0.0	[신년사]박정원 두산그룹 회장 "재무건전성 강화해 불확실성 극복"	2017-01-01	http://www.edaily.co.kr/news/newspath.asp?news...	박정원 두산그룹 회장[이데일리 성문재 기자] 박정원 두산(000150)그룹 회장이 ...
0.0	[신년사]정지원 증권금융 사장 "위기대응 위해 혁신·도전해야"	2017-01-01	http://www.edaily.co.kr/news/newspath.asp?news...	[이데일리 유재희 기자] 정지원 한국증권금융사장(사진)은 새해 불확실한 경영환경에 ...



(d) Number of news, by source

데이터스튜디오 : https://datastudio.google.com/open/16LHlInoEnt2JU5UBEm-NnJb5f__OEUrA

2.1 Preparing Corpus

채권 분석 보고서(Bond Analysts' Reports)

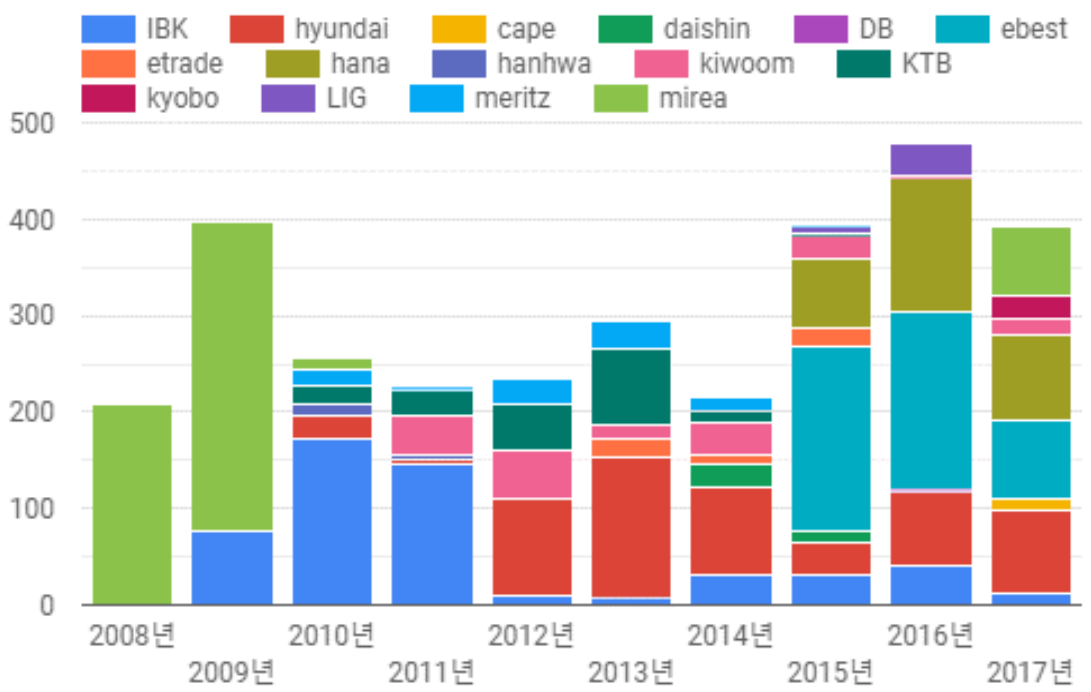
활용 Data : 채권 분석 보고서
- 수집 출처 : 네이버 금융 > 투자전략
- 기간 범위 : 2008년~2017년(10개년 데이터)
- 데이터 수 : 15개 금융사 3,095건

```
df.head()
```

	date	securities	title	report	tokens	ngram
0	13.10.17	KTB투자증권	10월 정치 불확실성 미완의 종결	10월 정치 불확실성, 미완의 종결 17 Oct. 2013 미 상원 부채한도 협상...	정치/NNG,불확실성/NNG,미완/NNG,종결/NNG,미/NNG,부채한도/NNG,협...	정치/NNG;불확실성/NNG;해소/NNG,채권시장/NNG;금리/NNG;상승/NNG,...
1	10.12.09	KTB투자증권	뉴트럴 엔딩 Neutral Ending	9 December 2010 2010년 12월 금융통화위원회 뉴트럴 엔딩 (Neut...	뉴트럴/NNG,엔딩/NNG,뉴트럴/NNG,통화정책/NNG,방향/NNG,연속선/NNG...	재정/NNG;문제/NNG;지정학/NNG;위험/NNG,금리/NNG;시장/NNG;예상/...
2	15.01.19	키움증권	저금리 장기화 주요 수출국과의 느슨해진 관계	Microsoft Word - 150119-WeeklyBondMarket.doc ...	저금리/NNG,장기/NNG,수출국/NNG,느슨/NNG,관계/NNG,저금리/NNG,장...	글로벌/NNG;안전자산/NNG;선호/NNG;강화/NNG,안전자산/NNG;선호/NNG...
3	15.12.30	이베스트투자증권	금통위 의사록 금리 인하보다 구조개혁 중요	Credit 손소현 02. 3779-0055 thecredit@ebestsec.co...	손/NNG,전병/NNG,하/XSV,종가/NNG,금리/NNG,스프레드/NNG,금리/N...	경제/NNG;성장률/NNG;전망/NNG;하향/NNG,성장/NNG;하방/NNG;위험/...
4	15.02.09	현대차증권	전화위복	<그림1> 주요 여전채 신용등급 변동 현황 01234567812.1 13.1 14....	여전채/NNG,신용등급/NNG,변동/NNG,현황/NNG,무림/NNG,자료/NNG,전...	외국인/NNG;국채선물/NNG;매도/NNG;금리/NNG;상승/NNG,우려/NNG;완...

```
df.head()['ngram'][0]
```

'정치/NNG;불확실성/NNG;해소/NNG,채권시장/NNG;금리/NNG;상승/NNG,디폴트/NNG;가능성/NNG;우려/NNG,채권시장/NNG;약세/NNG;되돌/VV,높/VA;정치/NNG;불확실성/NNG,기간스프레드/NNG;확대/NNG,안전자산/NNG;우호적/VAX,단기/NNG;불확실성/NNG'



(f) Number of bond analysts' reports, by section

GitHub : https://github.com/nlpbokproject/jh

2.2 Pre-Processing

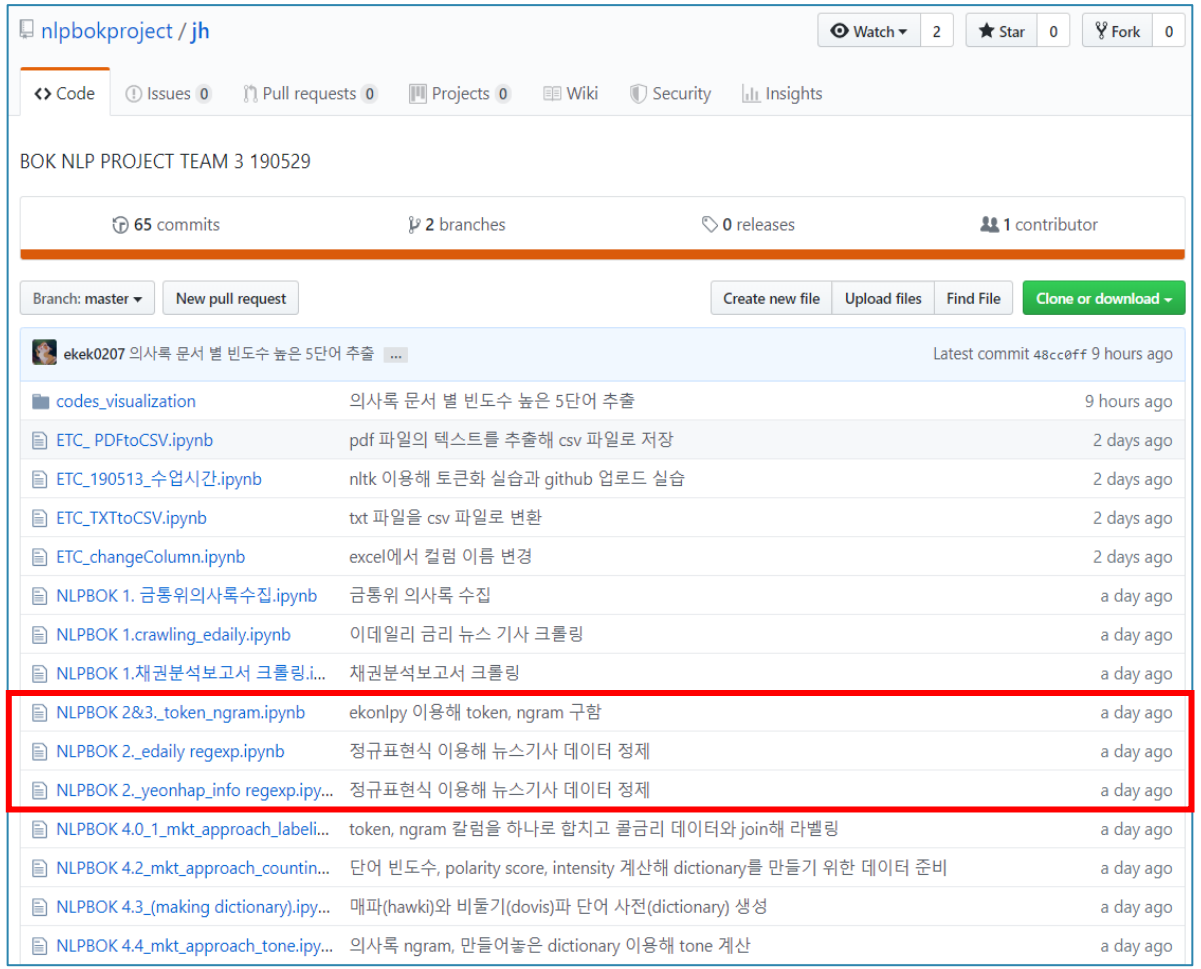
데이터 전처리 과정

01 Pre-processing

- Tokenization
- Normalization

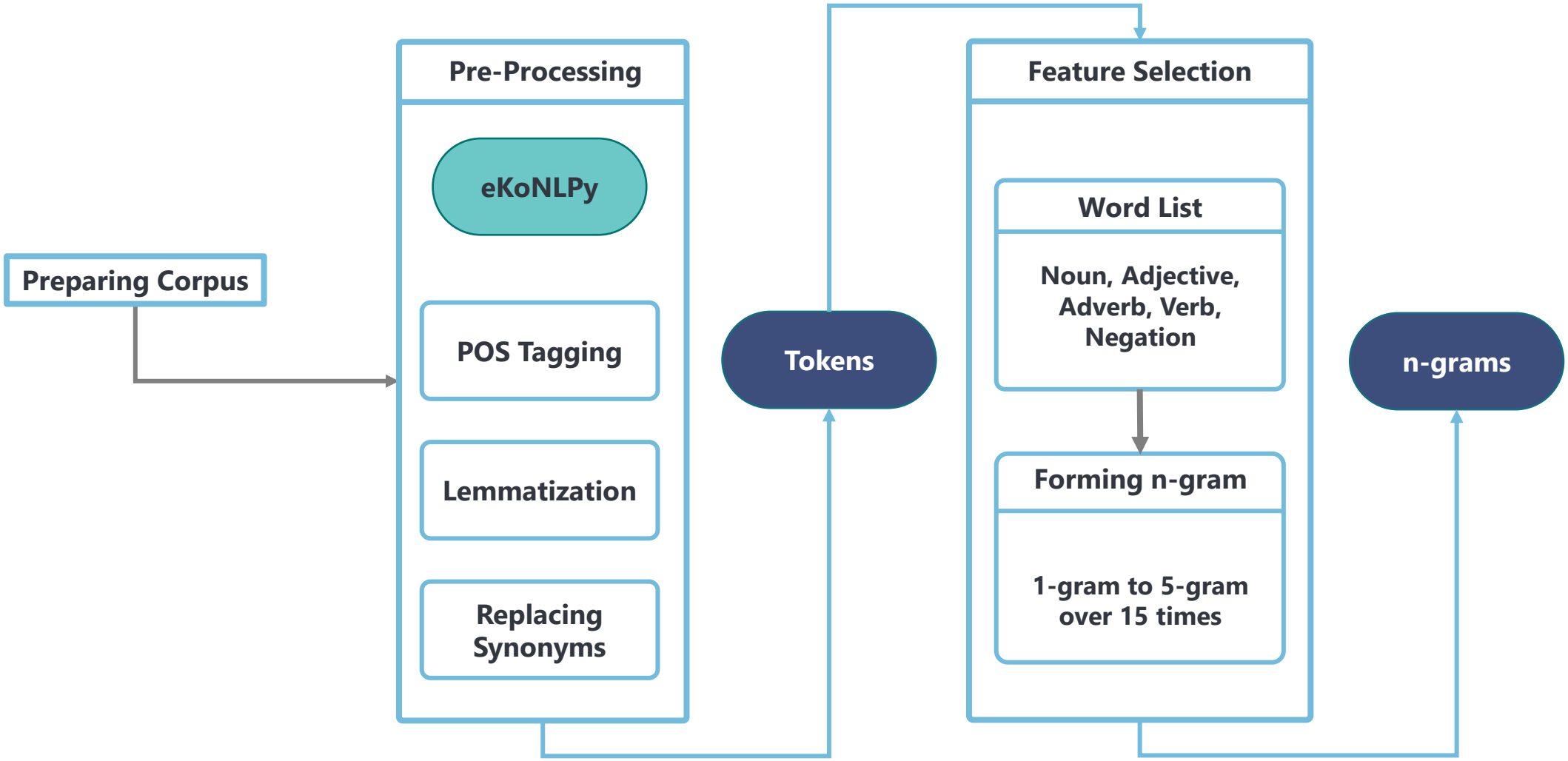
02 Feature Selection

- Token → n-gram



2.2 Pre-Processing

데이터 전처리 과정



2.2 Pre-Processing

Tokenization

```
news = []
mecab = Mecab()
for i in range(1,2): #1월부터 12월까지 돌림
    df = pd.read_csv("./news_201+str(i)+_preprocessed.csv")
    #print(news)

    #불러온 csv파일의 각 인덱스 script를 토큰화
    #for j in range(df.shape[0]):   #index 수만큼 range설정
    for j in range(0,df.shape[0]):   #index 수만큼 range설정
        #print(j)
        news = df['script']    #script만 토큰화하기 위해 script 선택

        t=mecab.pos(news[j]) #j번째 script의 tokenize
        t=str(t) #tokenize된 list 데이터를 string화
        #tstr = ''.join(t)

        df.loc[j, 'token'] = t #token 열 추가해서 j index에 해당하는 부분에 tokenize한 데이터 추가.

    df.to_csv("./news_"+str(i).zfill(2)+"-processingCompleted.csv", mode='w')
```

	com	date	title	url	script	token	ngram
0	Edaily	2017-01-01	"초불확실성의 시대"...험난한 금리정책 예고	http://www.edaily.co.kr/news/newspath.asp?news...	이주열 한국은행 총재가 지난달 15일 서울 중구 한은 본관에서 열린 금융통화위원회 ...	본관/NNG.열/VV.참석/NNG.생각/NNG.참겨/VV.상당기간/NNG.통화정책/...	변동성/NNG;확대/NNG;가능성/NNG;높/VA.외환시장/NNG;변동성/NNG;확...
1	Edaily	2017-01-01	[새해 부동산 재테크]"금리 변수..부동산 투자는 보수적 접근 필요"	http://www.edaily.co.kr/news/newspath.asp?news...	새해 부동산시장 전망은 그야말로 '안갯속'이다. 워낙 다양한 이슈와 변수들이 산재...	새해/NNG.부동산/NNG.전망/NNG.그야말로/MAG.안갯속/NNG.워낙/MAG....	주택/NNG;공급/NNG;과잉/NNG;우려/NNG.임대/NNG;수요/NNG;확보/N...
2	Edaily	2017-01-01	[신년사]박정원 두산그룹 회장 "재무건전성 강화해 불확실성 극복"	http://www.edaily.co.kr/news/newspath.asp?news...	박정원 두산그룹 회장 박정원 두산(000150)그룹 회장이 새해 재무건정성 강화에 ...	새해/NNG.재무/NNG.정성/NNG.강화/NNG.집중/NNG.하/XSV.의지/NN...	경영/NNG;환경/NNG;불확실성/NNG.금리/NNG;인상/NNG.정책/NNG;확대...

Normalization

```
#pattern = r'[a-zA-Z0-9._%+-]+@[a-zA-Z0-9._%+-]+' #이메일지우기
#pattern = '&lt[a-zA-Z-._%+-]+&gt;'
#pattern = r'[0-9]+/[0-9]+/[0-9]+' #날짜 지우기
#pattern = r'[0-9]+:[0-9]+' #시간 지우기
#pattern = '송고'
#pattern = '[[가-힣]+ [가-힣]+]'
#pattern = '[[가-힣a-zA-Z0-9._%+-]+]'
pattern = '▶ 관련기사 ◀.+' # ▶ 관련기사 ◀.와 뒤의 관련기사 제목 제거
pattern = '사진=[가-힣]+[₩s]제공' # 사진=○○○ 제공 제거
pattern = '₩[이데일리[₩w₩s가-힣]+₩]'#[이데일리[]] 제거
pattern = '₩[사진=[가-힣]+[₩s]기자₩]' #[사진=○○○ 기자] 제거
pattern = '₩[이데일리[₩s][가-힣]+[₩s][가-힣]+₩]' # [이데일리 ~~~] 제거
pattern = '₩[이데일리[₩s][가-힣]+[₩s]기자₩]' # '[이데일리 ○○○ 기자] 제거
pattern = '₩[[^a-zA-Z0-9_]+₩]' # [] 안의 기타 문구 제거
pattern = '[가-힣]+[₩s]+기자'
pattern = '[가-힣]+ 특파원'
pattern = r'[a-zA-Z0-9._%+-]+@[a-zA-Z0-9._%+-]+' #이메일지우기

# ₩[[^a-zA-Z0-9_]+₩]
error_ls=[]

for file in files:
    df = pd.DataFrame(columns=['com', 'date', 'title','url','script'])
    df = pd.read_csv(file, names=['com', 'date', 'title','url','script'])

    script_ls = df['script']

    new_script_ls = []
    for i in script_ls:
        mail_ls = []
        try:
            mail_ls = re.findall(pattern, i)
            print(mail_ls)
        except:
            print(file)
        if mail_ls != []:
            for j in mail_ls:
                new_script = i.replace(j, '')
            new_script_ls.append(new_script)
        else:
            new_script_ls.append(i)
```

2.2 Pre-Processing

Feature Selection

Out[0]:

	date	rate	difference	label	media	title	url	news	tokens
0	2012.01.01	3.29	0.03	1	연합뉴스	2012 예산 확정...경기 위축 대비 조기집행	https://news.naver.com/main/read.nhn?mode=LSD&...	재정적자 14조 전망... 2013 균형재정 '순항' 31일 확정된 2012년 예...	재정/NNG, 적자/NNG, 전망/NNG, 균형재정/NNG, 순항/NNG, 확정/NNG, 되/...
1	2012.01.01	3.29	0.03	1	연합뉴스	2012 예산 확정...복지 늘리고 FTA 지원 강화	https://news.naver.com/main/read.nhn?mode=LSD&...	일자리 3천800억↑, 한미FTA 지원 3천억 증액 복지지출 비중 28.5%... 정부안보...	일자리/NNG, 한미fta/NNG, 지원/NNG, 증액/NNG, 복지/NNG, 지출비중/NN...

In [0]: row['ngram']

Out[0]: '유가/NNG;상승/NNG;국제/NNG;금융시장/NNG;불안/NNG.원자재/NNG;가격/NNG;상승/NNG;물가/NNG;상승/NNG.금리/NNG;인하/NNG;유동성/NNG;공급/NNG;확대/NNG.금리/NNG;은행/NNG;여수신/NNG;금리/NNG;상승/NNG.소비자/NNG;물가/NNG;국제/NNG;유가/NNG;상승/NNG.둔화/NNG;우려/NNG;fed/NNG;금리/NNG;인하/NNG.물가/NNG;상승/NNG;압력/NNG;물가/NNG;상승/NNG.국제/NNG;유가/NNG;곡물/NNG;가격/NNG;상승/NNG.가격/NNG;상승/NNG;물가/NNG;상승/NNG;압력/NNG.하향/NNG;위험/NNG;물가/NNG;상향/NNG;위험/NNG.고용/NNG;지표/NNG;부진/NNG;경기/NNG;둔화/NNG.지표/NNG;부진/NNG;경기/NNG;둔화/NNG;우려/NNG.fed/NNG;금리/NNG;인하/NNG;기대/NNG;높/VA.가계/NNG;부채/NNG;상환/NNG;부담/NNG;크/VV.글로벌/NNG;신용경색/NNG;우려/NNG;완화/NNG.수요측면/NNG;물가/NNG;상승/NNG;압력/NNG.국제/NNG;원자재/NNG;가격/NNG;상승/NNG.cd/NNG;은행채/NNG;발행/NNG;확대/NNG.cd/NNG;은행채/NNG;발행/NNG;증가/NNG.경기/NNG;침체/NNG;가능성/NNG;증대/NNG.부동산/NNG;가격/NNG;오름/NNG;제한/NNG.fed/NNG;목표/NNG;금리/NNG;인하/NNG.국제/NNG;곡물/NNG;가격/NNG;급등/NNG.물가/NNG;상승/NNG;압력/NNG;완화/NNG.금리/NNG;대출/NNG;금리/NNG;상승/NNG.수요/NNG;물가/NNG;상승/NNG;압력/NNG.물가/NNG;상승/NNG;압력/NNG;증대/NNG.유가/NNG;급등/NNG;경제/NNG;성장/NNG.경제/NNG;성장/NNG;둔화/NNG;우려/NNG.대외/NNG;여건/NNG;불확실성/NNG;높/VA.국제/NNG;곡물/NNG;가격/NNG;상승/NNG.은행/NNG;가계/NNG;대출/NNG;감소/NNG.해외/NNG;차입/NNG;여건/NNG;악화/NNG.경기/NNG;둔화/NNG;우려/NNG;확산/NNG.금리/NNG;인하/NNG;기대/NNG;약화/NNG.민간/NNG;신용/NNG;증가/NNG;둔화/NNG.세계/NNG;경제/NNG;성장률/NNG;낮/VA.물가/NNG;소비자/NNG;물가/NNG;높/VA.미/NNG;경기/NNG;둔화/NNG;우려/NNG.가계/NNG;대출/NNG;연체율/NNG;낮/VA.단기/NNG;외채/NNG;비중/NNG;줄/VV.경기/NNG;하향/NNG;위험/NNG;크/VV.경기/NNG;침체/NNG;우려/NNG;높/VA.신용/NNG;대출/NNG;금리/NNG;낮/VA.금융시장/NNG;신용경색/NNG;완화/NNG.위축/NNG;가능성/NNG;작/VA;않/VX.국고채/NNG;유통수익률/NNG;하락/NNG.인플레이션/NNG;압력/NNG;증대/NNG.외국인/NNG;채권투자/NNG;확대/NNG.가격변수/NNG;변동성/NNG;확대/NNG.근원/NNG;인플레이션/NNG;안정/NNG.기대/NNG;인플레이션/NNG;안정/NNG.소비자/NNG;물가/NNG;일시적/VAX.실물/NNG;경제/NNG;불확실성/NNG.금융시장/NNG;금리/NNG;인하/NNG.중소기업/NNG;대출/NNG;위축/NNG.은행/NNG;자금사정/NNG;악화/NNG.gdp/NNG;성장률/NNG;상향/NNG.민간/NNG;소비/NNG;부정적/VAX.인플레이션/NNG;압력/NNG;높/VA.성장/NNG;잠재력/NNG;확충/NNG.실물/NNG;경제/NNG;부정적/VAX.원리금/NNG;상환/NNG;부담/NNG.경기/NNG;상승/NNG;모멘텀/NNG.글로벌/NNG;금융/NNG;불안/NNG.고유가/NNG;영향/NNG;상승/NNG.광공업/NNG;생산/NNG;감소/NNG.자본재/NNG;수주/NNG;감소/NNG.글로벌/NNG;신용/NNG;불안/NNG.시장성/NNG;수신/NNG;감소/NNG.비관/NNG;전망/NNG;확산/NNG.하향/NNG;위험/NNG;확대/NNG.설비/NNG;투자/NNG;부진/NNG.물가/NNG;오름/NNG;확대/NNG.민간/NNG;소비/NNG;위축/NNG.소비/NNG;심리/NNG;악화/NNG.차입/NNG;여건/NNG;개선/NNG.은행/NNG;자금/NNG;이탈/NNG.자금/NNG;쏠림/NNG;완화/NNG.가계/NNG;소비/NNG;위축/NNG.경제/NNG;안정/NNG;성장/NNG.설비/NNG;투자/NNG;회복/NNG.수출/NNG;경기/NNG;침체/NNG.은행/NNG;대출/NNG;감소/NNG.물가/NNG;불안/NNG;확산/NNG.경제/NNG;하향/NNG;위험/NNG.은행/NNG;수지/NNG;악화/NNG.상환/NNG;부담/NNG;증대/NNG.성장/NNG;하향/NNG;위험/NNG.선행/NNG;지수/NNG;상승/NNG.국제/NNG;유가/NNG;급등/NNG.둔화/NNG;우려/NNG;증폭/NNG.투자/NNG;완만/NNG;개선/NNG.경제/NNG;예상/NNG;성장/NNG.경기/NNG;지표/NNG;악화/NNG.경제/NNG;잠재/NNG;성장/NNG.민간/NNG;소비/NNG;신장/NNG.수출/NNG;견조/NNG;증가/NNG.기업/NNG;심리/NNG;악화/NNG.가격/NNG;상승/NNG;확대/NNG.소비/NNG;심리/NNG;위축/NNG.주택/NNG;경기/NNG;부진/NNG.수출/NNG;증가/NNG;둔화/NNG.산업/NNG;생산/NNG;증가/NNG.주택/NNG;건설/NNG;부진/NNG.주택/NNG;판매/NNG;감소/NNG.금리/NNG;소폭/MAG;하락/NNG.공급/NNG;증가/NNG;하락/NNG.물가/NNG;상승/NNG;우려/NNG.금리/NNG;비슷/NNG;상승/NNG.대출/NNG;비중/NNG;확대/NNG.cd/NNG;발행/NNG;증가/NNG.신용/NNG;불안/NNG;우려/NNG.불안/NNG;우려/NNG;완화/NNG.은행/NNG;금리/NNG;인상/NNG.부채/NNG;구조/NNG;개선/NNG.집단/NNG;대출/NNG;증가/NNG.투자/NNG;회복/NNG;더디/VA.당분간/MAG;높/VA;상승/NNG.투자/NNG;증가율/NNG;낮/VA.소폭/MAG;증가/NNG;그치/VV.단기/NNG;외채/NNG;늘/VV.ted스프레드/NNG;축소/NNG.소비/NNG;영향/NNG;줄/VV.수출/NNG;높/VA;증가/NNG.나쁘/VA;가능성/NNG;높/VA.수출/NNG;높/VA;신장/NNG.불안/NNG;우려/NNG;높/VA.금융시스템/NNG;부정적/VAX.스왑베이시스/NNG;확대/NNG.경제/NNG;작/VV;않/VX.서브프라임/NNG;부실/NNG.우려/NNG;작/VA;않/VX.금융시스템/NNG;불안/NNG.인플레이션/NNG;우려/NNG.통신요금/NNG;인하/NNG.생산능력/NNG;확대/NNG.금융안정/NNG;저해/NNG.통화정책/NNG;완화/NNG.불확실성/NNG;점증/NNG.대외/NNG;불확실성/NNG.금융시장/NNG;안정/NNG.불확실성/NNG;증대/NNG.서비스업/NNG;성장/NNG.주택시장/NNG;안정/NNG.금융시장/NNG;위험/NNG.실효환율/NNG;절하/NNG.불안정성/NNG;증대/NNG.투자심리/NNG;악화/NNG.서비스업/NNG;증가/NNG.주택가격/NNG;하락/NNG.채권시장/NNG;불안/NNG.인하/NNG;불확실/NNG.투자/NNG;환위험/NNG.수요자/NNG;부담/NNG.부동산/NNG;안정/NNG.실업률/NNG;상승/NNG.금리갭/NNG;확대/NNG.변동성/NNG;증대/NNG.부실/NNG;우려/NNG.물가/NNG;안정/NNG.경제/NNG;비관/NNG.경제/NNG;둔화/NNG.생산/NNG;확대/NNG.안정/NNG;효과/NNG.여력/NNG;저하/NNG.대출/NNG;축소/NNG.자산/NNG;확대/NNG.대출/NNG;확대/NNG.임금/NNG;인상/NNG.소비/NNG;회복/NNG.경제/NNG;지행/NNG.건설/NNG;성장/NNG.소비/NNG;증가/NNG.수출/NNG;호조/NNG.비용/NNG;부담/NNG.수요/NNG;압력/NNG.위험/NNG;증대/NNG.물가/NNG;성장/NNG.위험/NNG;하향/NNG.수요/NNG;위축/NNG.불확실성/NNG;크/VV.수요/NNG;호조/NNG.회복/NNG;약화/NNG.가격/NNG;강세/NNG.판매/NNG;부진/NNG.재고/NNG;감소/NNG.가격/NNG;안정/NNG.위축/NNG;증가/NNG.수출/NNG;감소/NNG.수입/NNG;증가/NNG.수요/NNG;감소/NNG.불안/NNG;완화/NNG.위축/NNG;상승/NNG.수요/NNG;상승/NNG.하락/NNG;멈추/VV.연체율/NNG;높/VA.금리/NNG;오르/VV.체감도/NNG;높/VA.예상/NNG;나쁘/VA.변동성/NNG;크/VV.안정/NNG;많/VA.가격/NNG;높/VA.투자/NNG;낮/VA.낮/VA;증가/NNG.생산/NNG;높/VA.높/VA;성장/NNG.쉽/VA;해소/NNG.수주/NNG;높/VA.수입/NNG;높/VA.적자/NNG;늘/VV'

2.3 Polarity Classification

극성 분류

01 n-gram 라벨링

- 콜금리 데이터를 통한 라벨링
- 날짜 / n-gram / 라벨링(dovish, hawkish)

02 Naïve Bayes Classifier 모델링

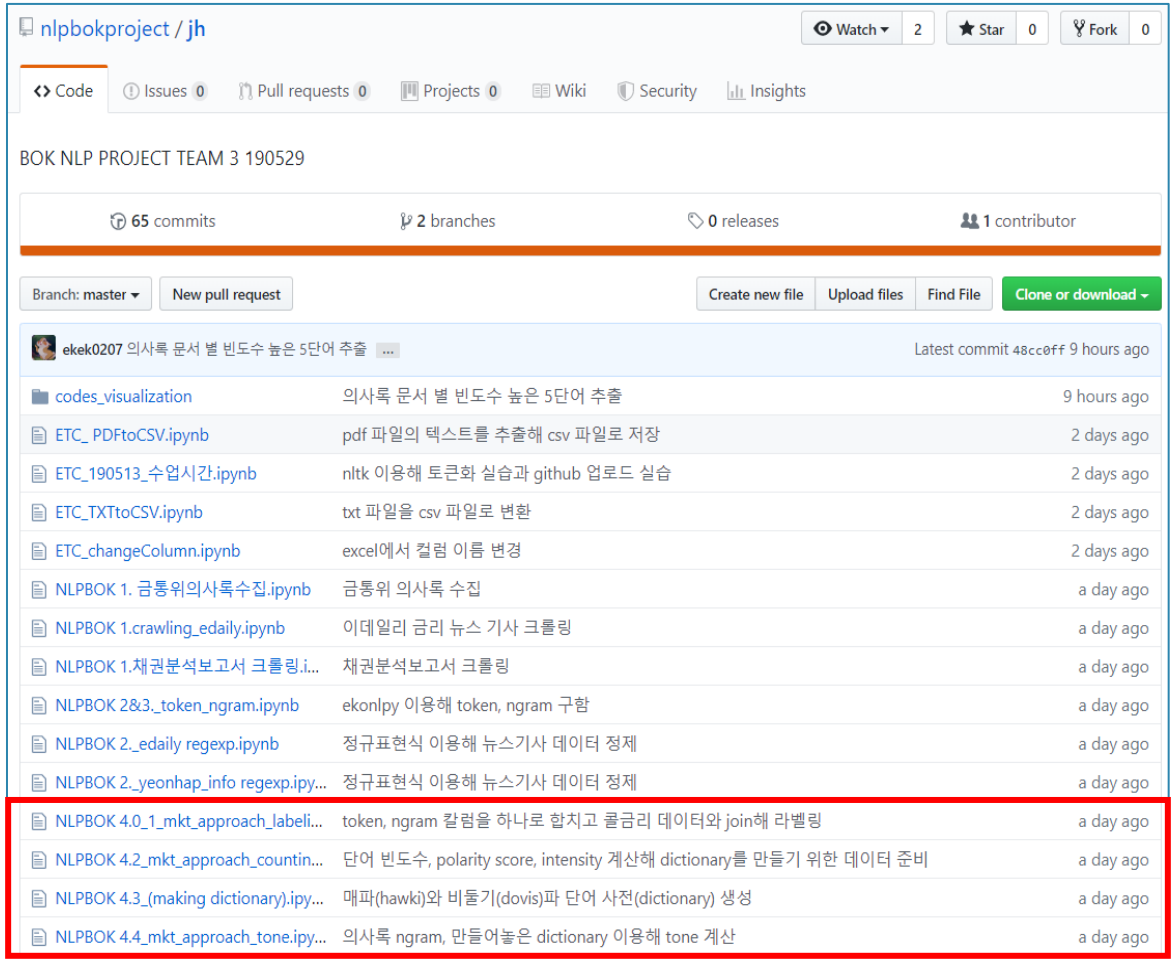
- NBC모델을 기반으로 조건부확률 도출
- 극성점수 도출

03 Dictionary

- Dovish / Hawkish n-gram 사전

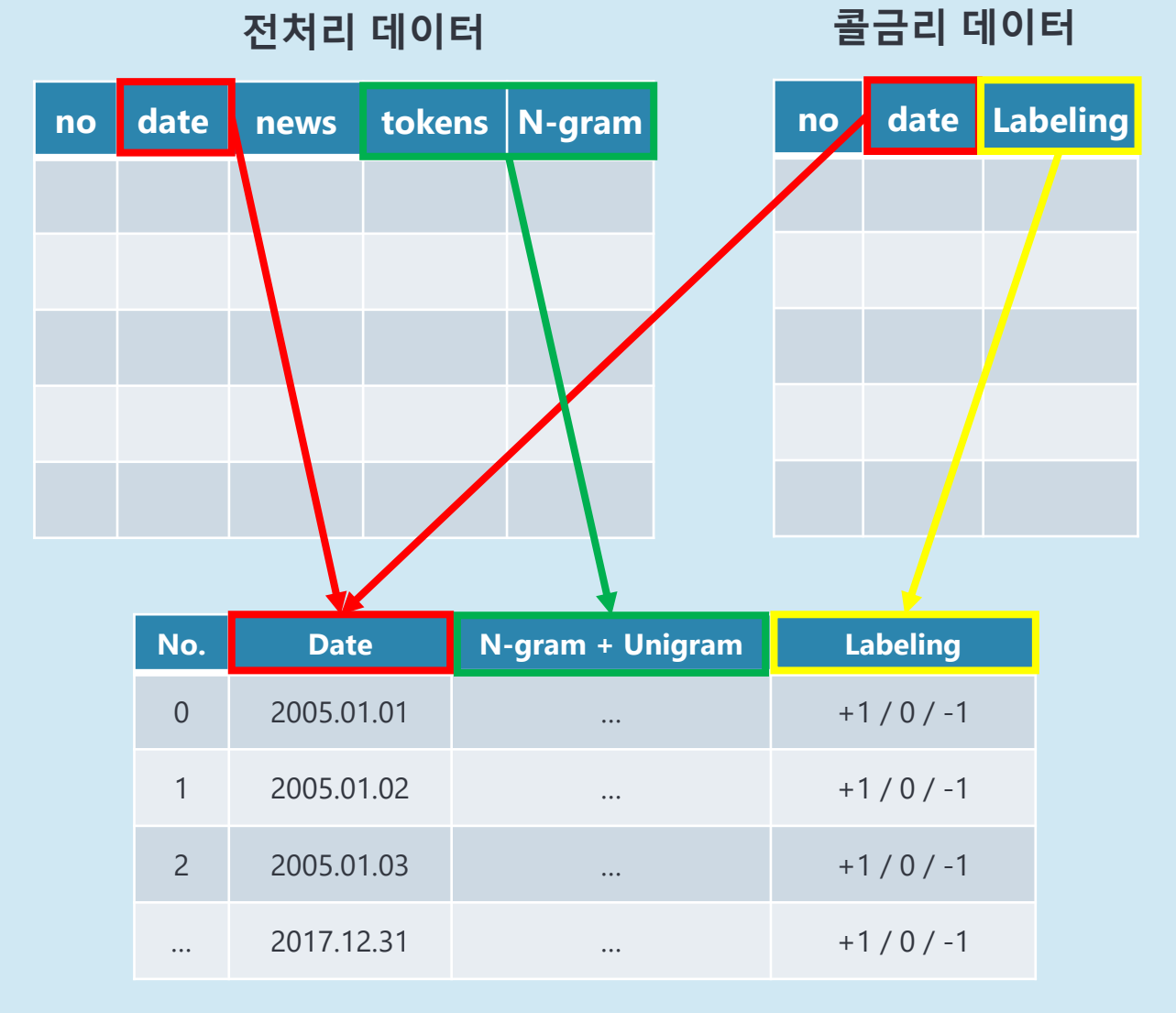
04 Measuring Sentiments

- Dovish / Hawkish 사전



3.1 Polarity Classification

Step 1. n-gram 라벨링



추출한 token과 n-gram을 콜금리 데이터로 라벨링

- ▶ **키 값은 'date'**
- ▶ 콜금리가 상승할 경우 positive(hawkish) : +1
- ▶ 콜금리가 하락할 경우 negative(dovish) : -1

같은 날짜 = 같은 라벨

2.3 Polarity Classification

Step 2. Naïve Bayes Classifier 모델링

기계학습(Machine-learning)중 하나인 단순확률 분류자, Naïve-Bayes Classifier(NBC)를 사용

$$polarity\ score = \frac{p(feature\ |hawkish)}{p(feature\ |hawkish)} = \frac{p(feature\ |hawkish)/p(hawkish)}{p(feature\ |hawkish)/p(dovish)}$$

Hawkish Documents

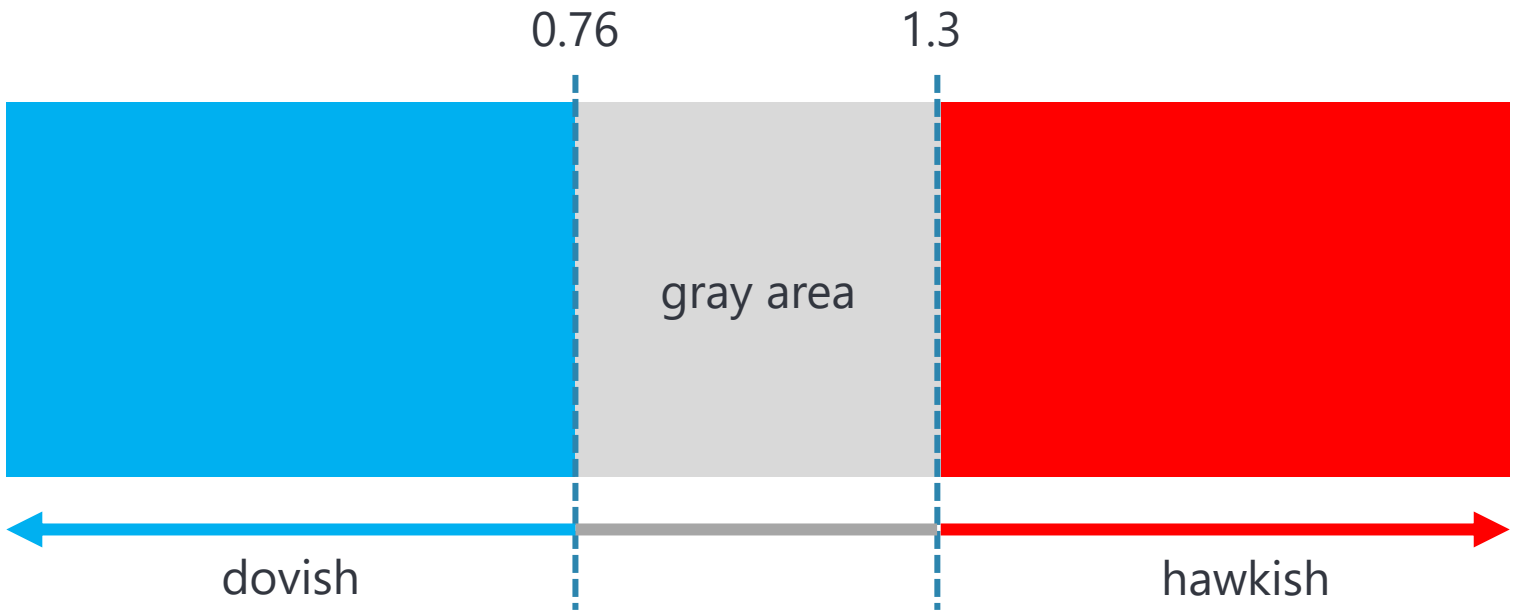
Dovish Documents

극성점수 도출

N-gram + unigram	Counting		Polarity Score	Intensity	Label
	Hawkish	Dovish			
인상 /NNG	153,428	116,129	1.374	1.374	+1
인하 /NNG	50,452	86,461	0.606	1.647	-1
…	…	…	…	…	…

Step 2. Naïve Bayes Classifier 모델링

1< Intensity < 1.3 인 gray area 값들은 제외



2.3 Polarity Classification

Step 3. Dictionary

Hawkish	Dovish
국고채/NNG;금리/NNG;뛰/VV	ecb/NNG;금리/NNG;인상/NNG;금리/NNG;인상
지수/NNG;고용/NNG;지수/NNG;상승/NNG	/NNG 수요/NNG;기대/NNG;어렵/VA 수요/NNG;
오래/MAG;금리/NNG;인상/NNG 안정/NNG;심리/NNG;회복/NNG	한계/NNG
실현지수/NNG 든게/VV 시끌벅적/MAG	안정/NNG;기대/NNG;확산/NNG
높/VA;국제/NNG;유가/NNG;상승/NNG 높/VA;주가/NNG;오르/VV	국채선물시장/NNG;인플레이션/NNG;우려/NNG
콜/NNG;금리/NNG;인상/NNG;통화정책/NNG;불확실성/NNG	인플레이션/NNG;유가/NNG;하락/NNG
통화정책/NNG;불확실성/NNG;부담/NNG 국제채/NNG	우려/NNG;심리/NNG;위축/NNG 전미소매업협회/NNG
인상/NNG;가능성/NNG;낮/VA;금리/NNG;인상/NNG	비지표채권/NNG
경제/NNG;성장률/NNG;호전/NNG 주택가격/NNG;내림/NNG	채권시장/NNG;단기/NNG;금리/NNG;하락/NNG
서비스업/NNG;지표/NNG;부진/NNG	봇물/NNG;터지/VV;쏟/VV
판매/NNG;재고/NNG;증가/NNG	달러엔/NNG;환율/NNG;하락/NNG;압력/NNG
미/NNG;국채/NNG;가격/NNG;떨어/VV	주택/NNG;판매/NNG;예상/NNG;하회/NNG
과잉/NNG;유동성/NNG;경기/NNG;회복/NNG	수출/NNG;출하/NNG;증가/NNG 총자본투자효율/NNG
불안요인/NNG;증가/NNG	하락/NNG;불구/NNG;높/VA 고용증가율/NNG;
주택/NNG;공급/NNG;과잉/NNG;우려/NNG	둔화/NNG
…	…

2.3 Polarity Classification

Step 4. Measuring Sentiments

n-gram의 극성점수(polarity score)를 이용해 문장(sentences)와 문서(documents)의 톤 측정

i . 개별 문장(sentence)의 어조($tone_s$) 측정

$$tone_s = \frac{No.of\ hawkish\ features - No.of\ dovish\ features}{No.of\ hawkish\ features + No.of\ dovish\ features}$$

ii. 개별 문서(document)의 어조($tone_i$) 측정

$$tone_i = \frac{No.of\ hawkishtone_{s,i} - No.of\ dovish\ tone_{s,i}}{No.of\ hawkish\ tone_{s,i} + No.of\ dovish\ tone_{s,i}}$$

2.3 Polarity Classification

Step 4. Measuring Sentiments

n-gram의 극성점수(polarity score)를 이용해 문장(sentences)와 문서(documents)의 톤 측정

No.	Date	BOK minutes (Foreign Currency + Financial Markets)	$tone_i$
1	20050609	일부 위원은 우리나라 경제가 일본경제 에 비해 더 나아질 특별한 요인이 없음에도 ...	0.172414
2	20050707	일부 위원은 6월 들어 국제유가가 크게 오르고 원화와 엔화간 동조화 현상이 뚜렷...	0.142857
3	20050811	일부 위원은 위안화 절상 이후 아시아 통화중 우리나라 원화의 절상폭이 가장 컸다...	0.380282
4	20050908	일부 위원은 최근 인도네시아 금융불안 사태가 발생한 가운데 국제 환투기세력이 인...	0.241379
5	20051011	일부 위원은 미국과 우리나라의 정책금 리 격차가 확대되고 있는 가운데 금년 두 차...	0.473684
...	...	...	...

Ⅲ.프로젝트 결과

3. Project Outcome

프로젝트 결과

01 **정확도(Accuracy) 측정**

- Intensity에 따른 모델 정확도 측정

02 **MP Sentiments ($tone_{mkt}$) ↔ MP Decision, BOK Policy rate, EPU**

- BOK Policy rate / MP Decision
- EPU

03 **MP Sentiments ($tone_{mkt}$) ↔ Macroeconomic Variables**

- CPI
- Stock Market Price / 장단기 스프레드

3. Project Outcome

정확도(Accuracy) 측정

```
predict(test_set, testdf, testdf["counting_h"].sum(), testdf["counting_d"].sum(), 1.3
```

Intensity = 1.3 | Accuracy = 69.800%

```
predict(test_set, testdf, testdf["counting_h"].sum(), testdf["counting_d"].sum(), 1.5
```

Intensity = 1.5 | Accuracy = 71.396%

```
predict(test_set, testdf, testdf["counting_h"].sum(), testdf["counting_d"].sum(), 1.3
```

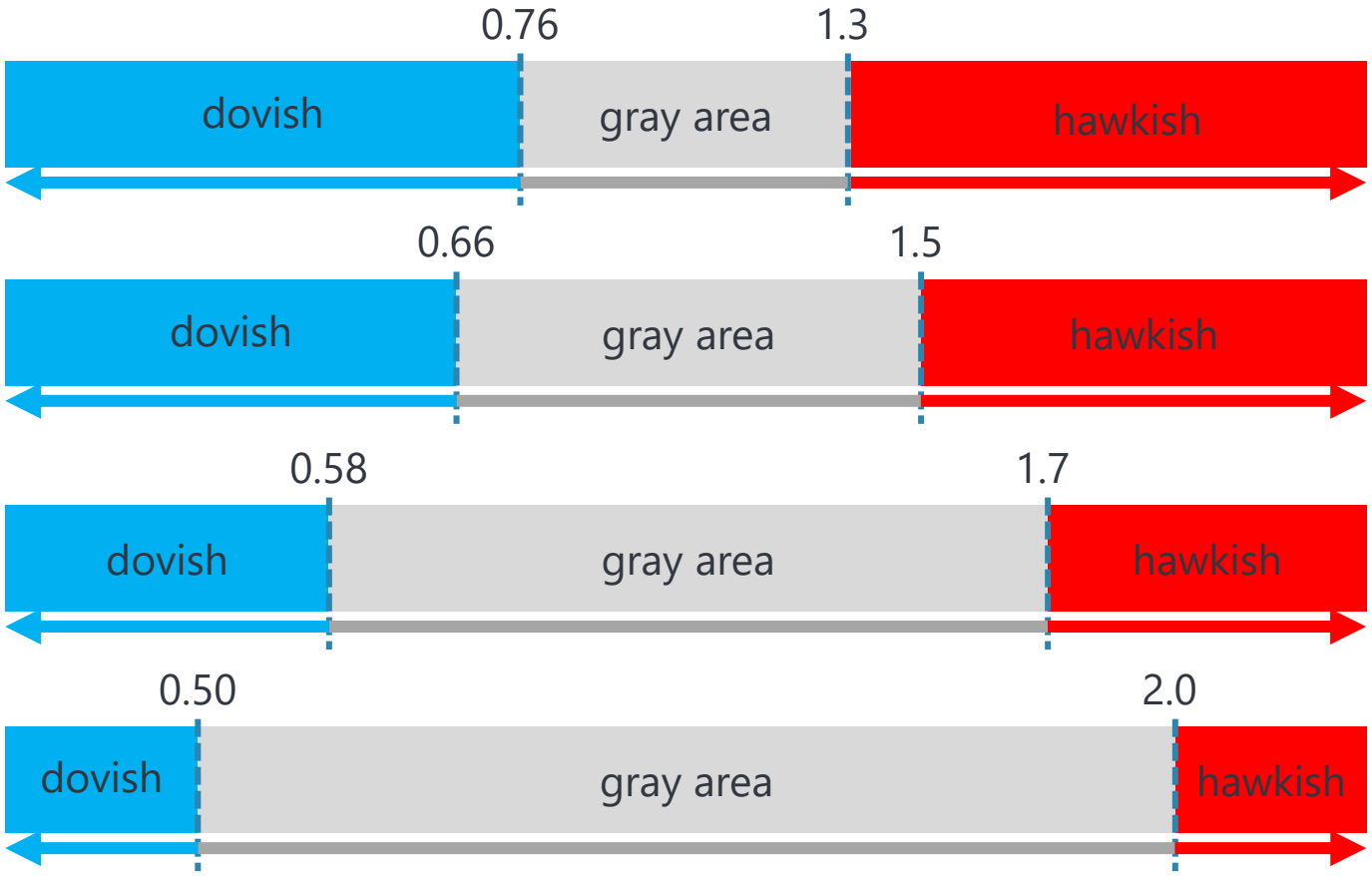
Intensity = 1.7 | Accuracy = 72.059%

```
predict(test_set, testdf, testdf["counting_h"].sum(), testdf["counting_d"].sum(), 1.3
```

Intensity = 2.0 | Accuracy = 71.181%

3. Project Outcome

정확도(Accuracy) 측정



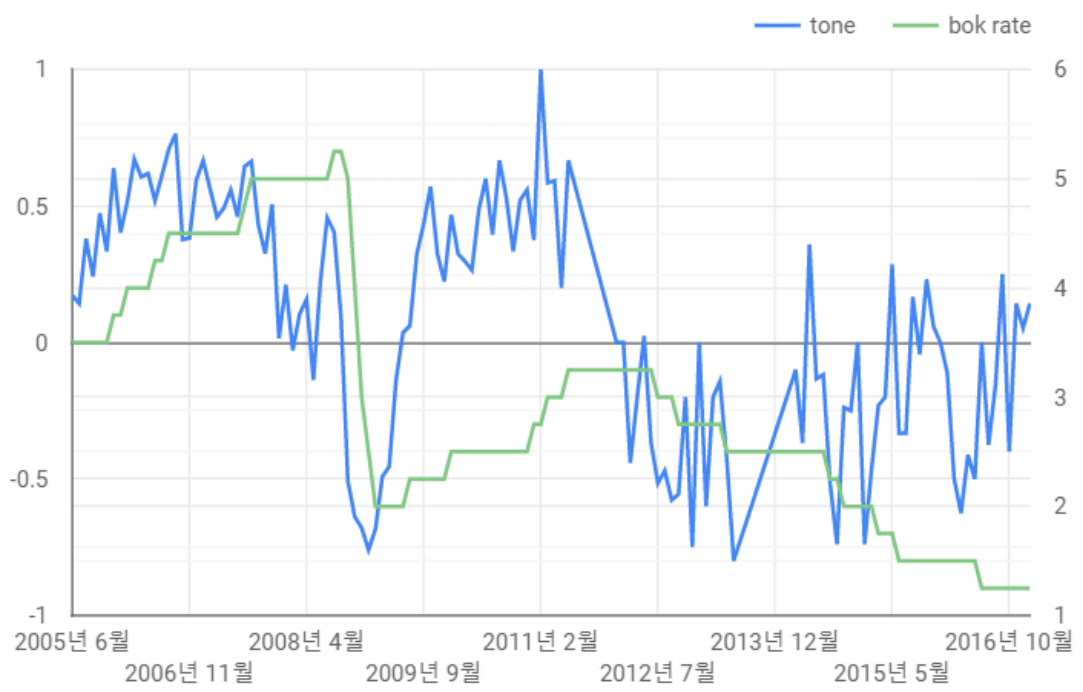
Intensity가 커질수록 Overfitting발생 가능성 ↑

데이터스튜디오 : https://datastudio.google.com/open/1tNwx2xuv0IZsjc2OoTE8IYKYPh5IbV9R

3. Project Outcome

프로젝트 결과 비교

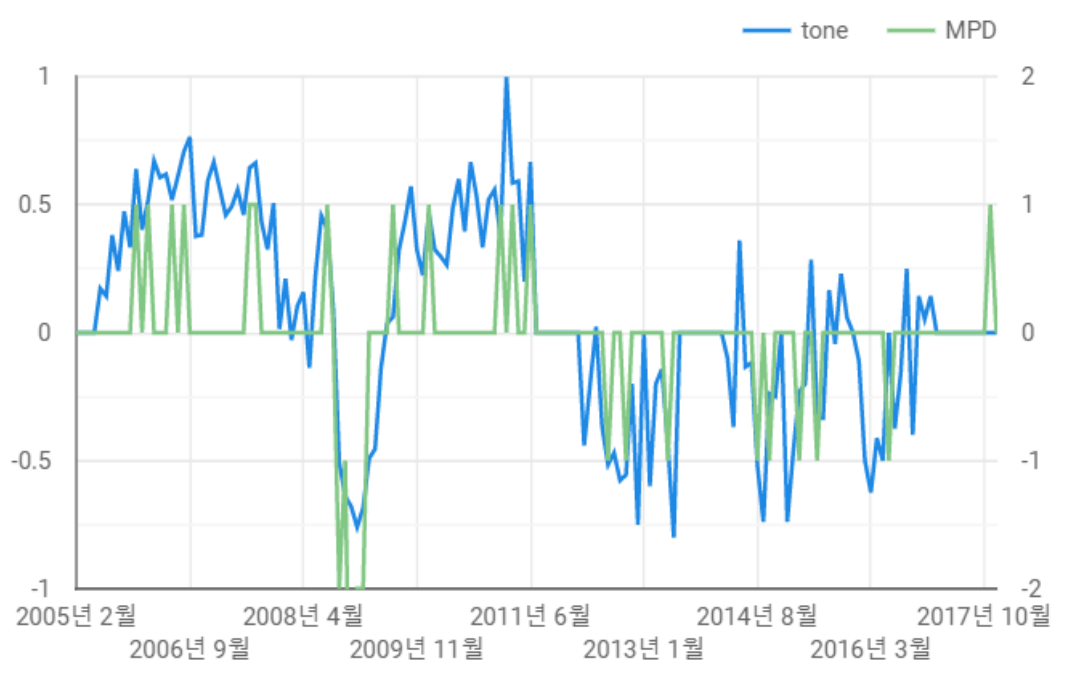
(e) $tone_{mkt}$ and BOK policy rate



(e) tone(mkt) and BOK policy rate

출처 : 한국은행

(a) $tone_{mkt}$ and MP decision



(a) tone(mkt) and MP decisions

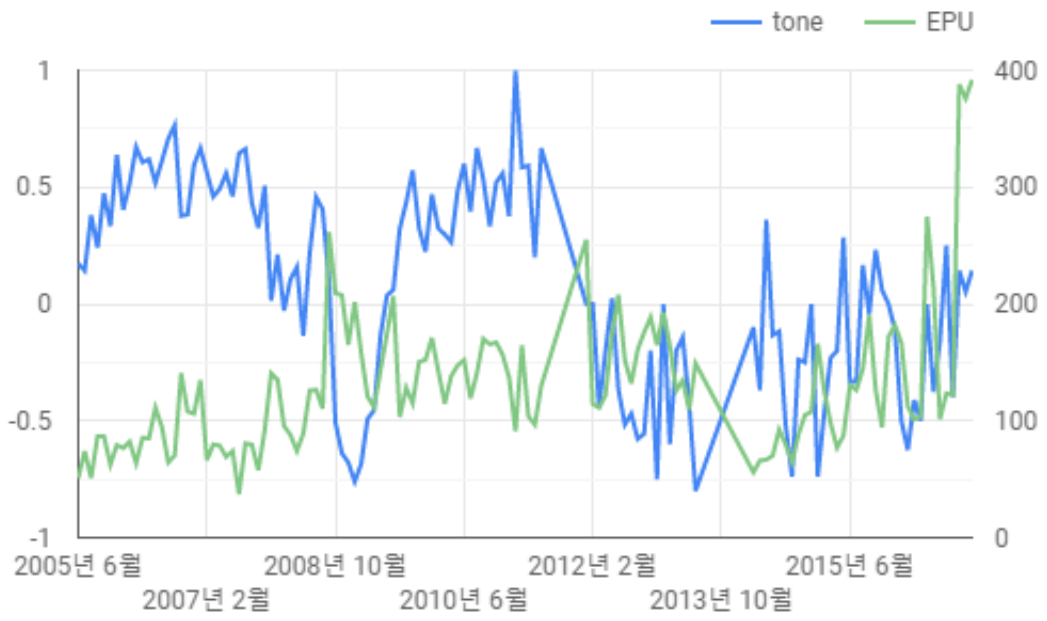
출처 : 한국은행

데이터스튜디오 : https://datastudio.google.com/open/1tNwx2xuv0IZsjc2OoTE8IYKYPh5IbV9R

3. Project Outcome

프로젝트 결과 비교

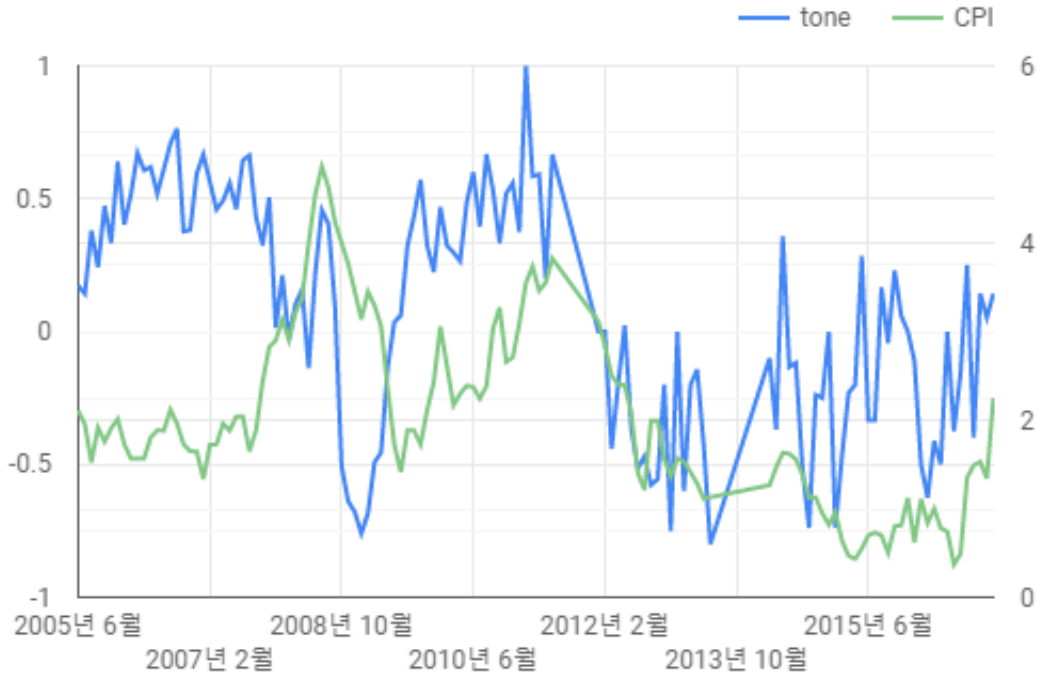
(d) $tone_{mkt}$ and EPU



(d) tone(mkt) and EPU

출처 : http://www.policyuncertainty.com/korea_monthly.html

(b) $tone_{mkt}$ and CPI



(b) tone(mkt) and CPI

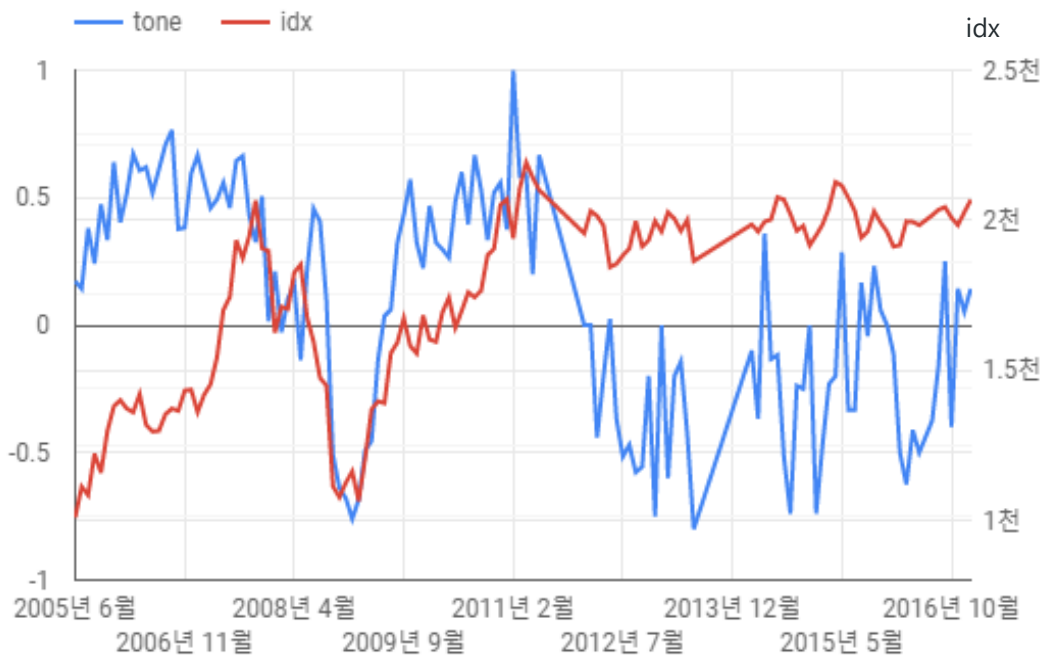
출처 : 통계청

데이터스튜디오 : https://datastudio.google.com/open/1tNwx2xuv0IZsjc2OoTE8IYKYPh5IbV9R

3. Project Outcome

프로젝트 결과 비교

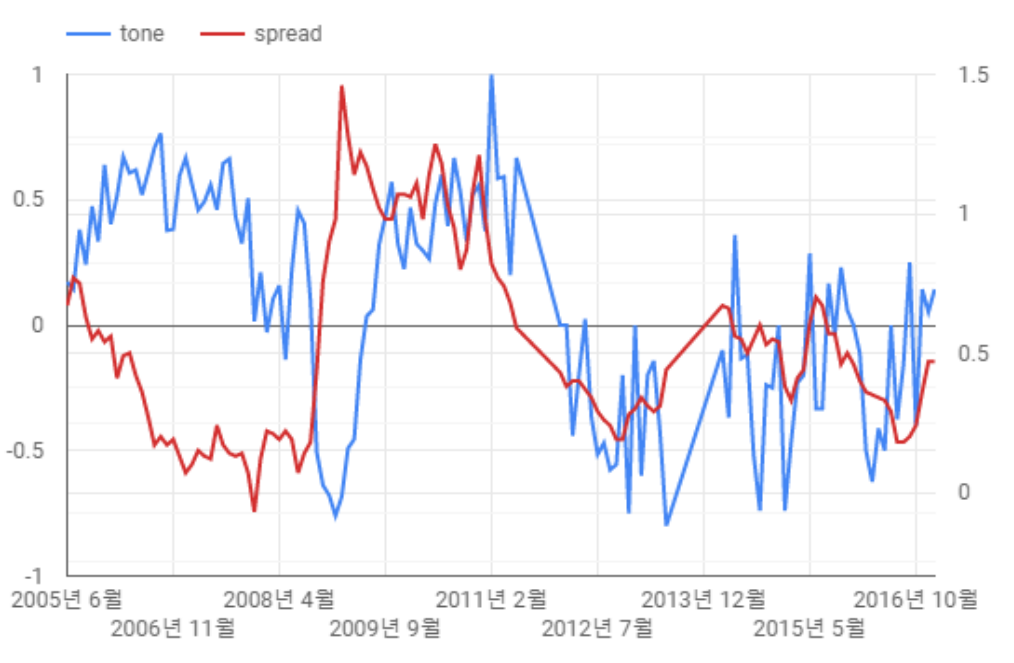
(c) $tone_{mkt}$ and stock market index



(c) tone(mkt) and stock market index

출처 : http://www.krx.co.kr/main/main.jsp

(f) $tone_{mkt}$ and 장단기 스프레드



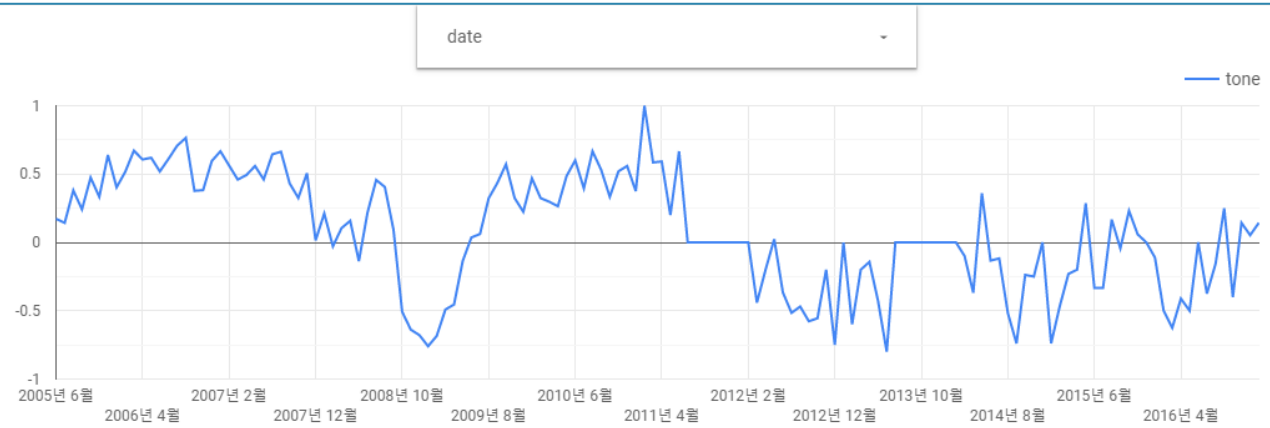
(f) tone(mkt) and 10-year treasury constant maturity minus 3-year treasury constant maturity

출처 : e-나라지표

데이터스튜디오 : https://datastudio.google.com/open/1Bt38fDrqhlyXoGSpnmyU1-6DSqMylAA6

3. Project Outcome

프로젝트 결과 비교



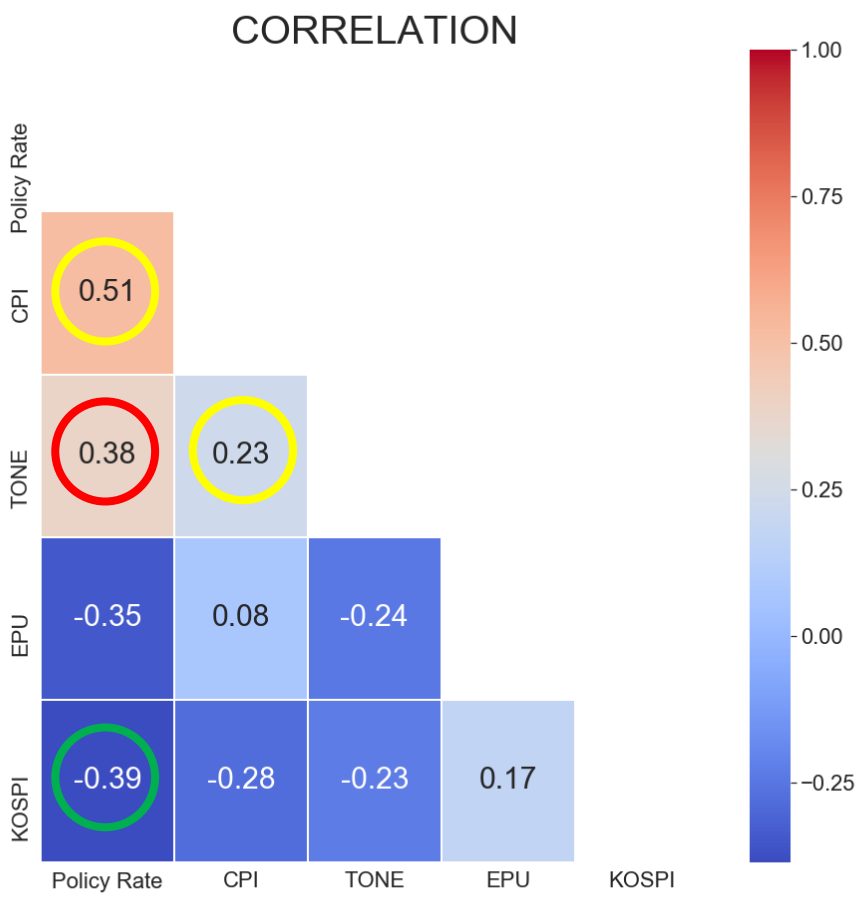
	date	1	2	3	4	5
1.	2005년 6월	('상승/NNG', 79)	('금리/NNG', 75)	('증가/NNG', 64)	('가격/NNG', 51)	('부동산/NNG', 44)
2.	2005년 7월	('금리/NNG', 89)	('증가/NNG', 77)	('상승/NNG', 69)	('생각/NNG', 44)	('지속/NNG', 44)
3.	2005년 8월	('금리/NNG', 87)	('상승/NNG', 57)	('증가/NNG', 53)	('부동산/NNG', 40)	('경제/NNG', 39)
4.	2005년 9월	('금리/NNG', 85)	('증가/NNG', 63)	('상승/NNG', 62)	('달러/NNG', 38)	('물가/NNG', 37)
5.	2005년 10월	('금리/NNG', 96)	('상승/NNG', 68)	('증가/NNG', 52)	('물가/NNG', 49)	('인상/NNG', 49)
6.	2005년 11월	('금리/NNG', 108)	('증가/NNG', 76)	('지속/NNG', 59)	('상승/NNG', 55)	('회복/NNG', 51)
7.	2005년 12월	('금리/NNG', 80)	('증가/NNG', 61)	('달러/NNG', 55)	('인상/NNG', 47)	('상승/NNG', 42)
8.	2006년 1월	('금리/NNG', 82)	('상승/NNG', 50)	('달러/NNG', 47)	('증가/NNG', 42)	('인상/NNG', 41)
9.	2006년 2월	('금리/NNG', 68)	('증가/NNG', 65)	('상승/NNG', 51)	('달러/NNG', 41)	('하락/NNG', 39)
10.	2006년 3월	('증가/NNG', 66)	('상승/NNG', 59)	('금리/NNG', 56)	('달러/NNG', 45)	('지속/NNG', 37)
11.	2006년 4월	('증가/NNG', 59)	('금리/NNG', 54)	('상승/NNG', 51)	('하락/NNG', 48)	('달러/NNG', 44)
12.	2006년 5월	('상승/NNG', 62)	('증가/NNG', 57)	('하락/NNG', 53)	('금리/NNG', 52)	('경기/NNG', 51)

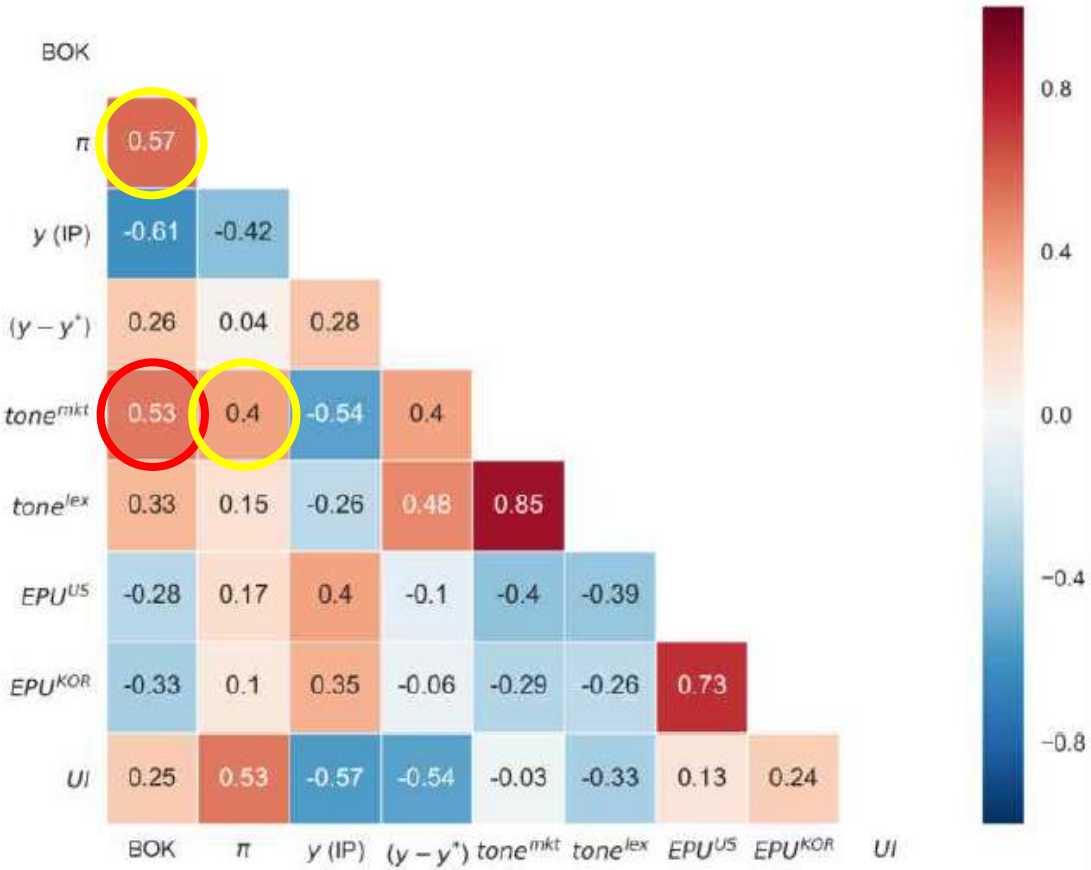
1 - 100 / 126

데이터스튜디오 : shorturl.at/ovCQ3

3. Project Outcome

Correlation Coefficients





(a) With macroeconomic variables

3. Project Outcome

Wordclouds

IV. 프로젝트 시사점

< 프로젝트 목적 >

1. 데이터 처리부터 분석까지 논문을 충실히 따르며 직접 구현

2. 프로젝트 결과물의 활용 및 응용 가능성 검토

4. Project Implications

논문에서 언급한 연구의 시사점

텍스트 마이닝을 이용한 금통위 의사록 분석

① 통화 정책 효과성 분석 모델에 적용 가능

② 장래의 통화 정책 결정 및 한은의 의사 소통의 효과성 평가

③ 거시경제 불확실성, 미래의 통화 정책 기조에 대한 대중의 기대, 주식 시장 정서 등을 측정에 활용

→ 자산가격이나 실제 변수에 미치는 영향 조사

4. Project Implications

프로젝트 후속 연구 방향

텍스트 마이닝을 이용한 기준금리 예측

~~① 통화 정책 효과성 분석 모델에 적용 가능~~

~~② 장래의 통화 정책 결정 및 한은의 의사 소통의 효과성 평가~~

③ 거시경제 불확실성, 미래의 통화 정책 기조에 대한 대중의 기대, 주식 시장 정서 등을 측정에 활용

→ 자산가격이나 실제 변수에 미치는 영향 조사

4. Project Implications

프로젝트 후속 연구 방향

③ 거시경제 불확실성, 미래의 통화 정책 기조에 대한 대중의 기대, 주식 시장 정서 등을 측정에 활용
→ 자산가격이나 실제 변수에 어떻게 영향을 미치는지 조사 가능

주식시장

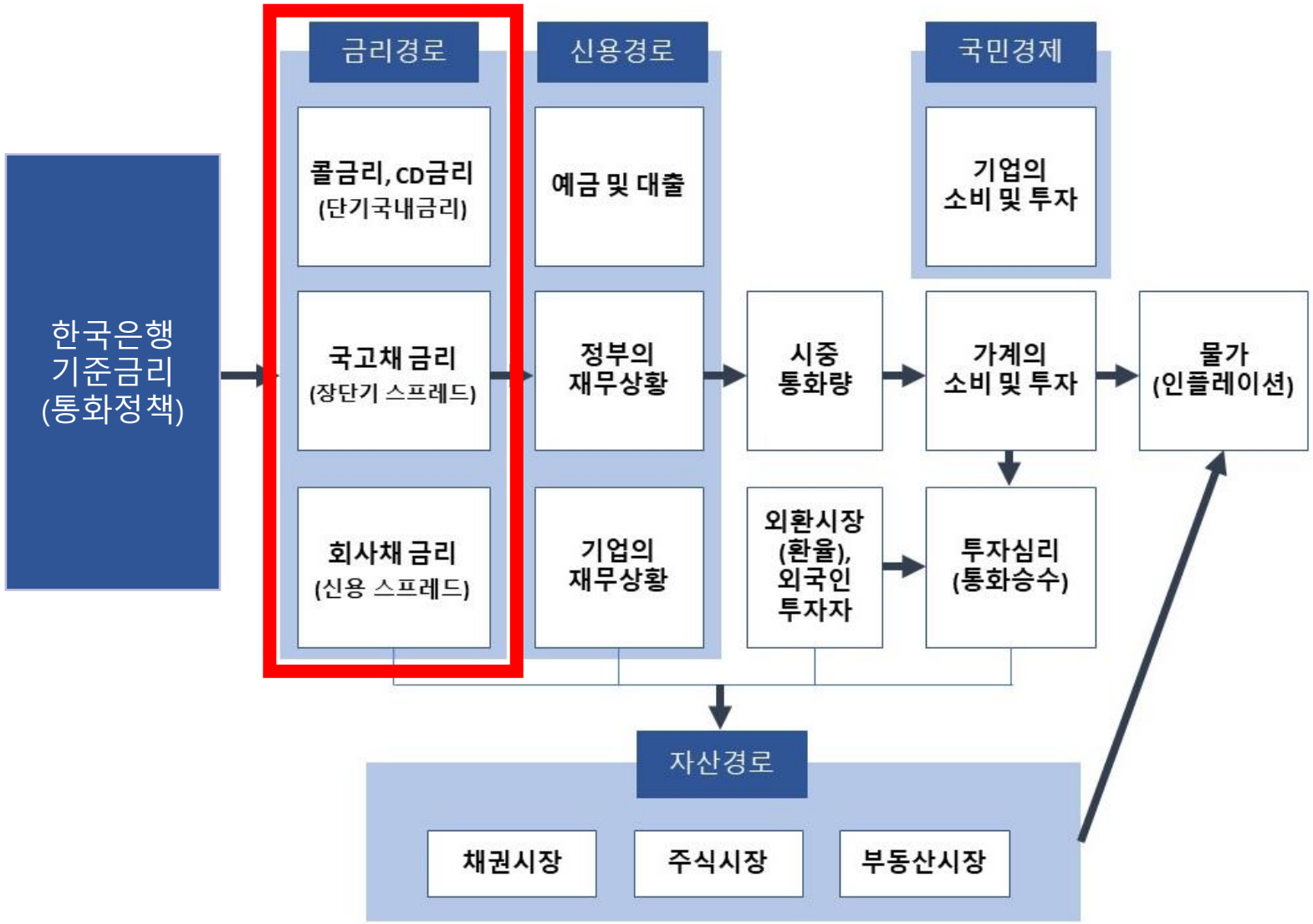
- 소규모개방경제인 국내 주식시장에 부적합
- 기준금리 이외에 수많은 변수가 존재

채권시장

- 기준금리 인상 시 직접적인 영향을 받음
- 논문 근거로 한 예측의 의미가 높음
- 채권투자, 금리나 채권을 기초자산으로 하는 파생상품 투자에 활용

4. Project Implications

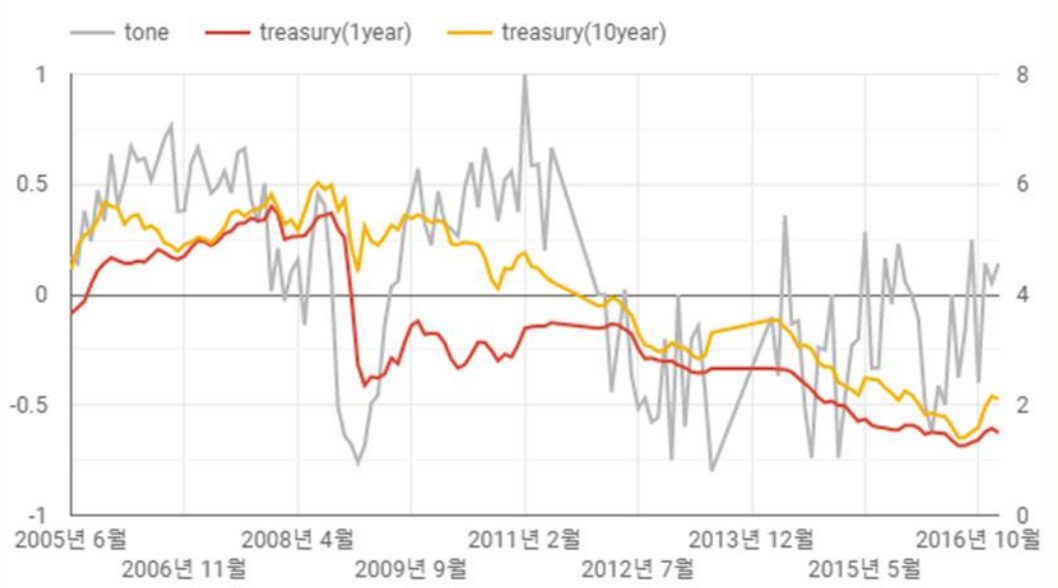
프로젝트 후속 연구 방향



4. Project Implications

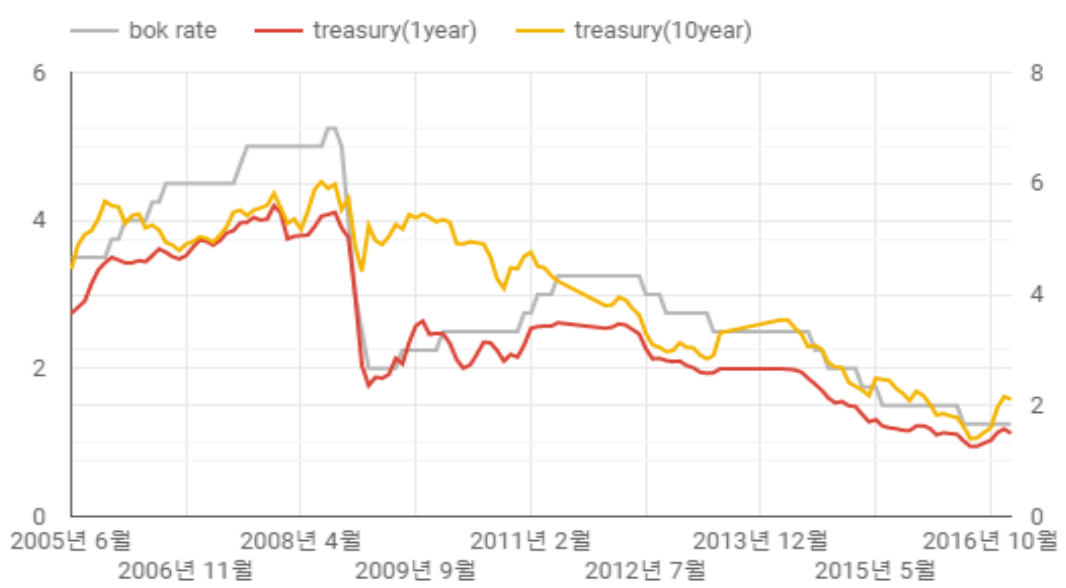
프로젝트 후속 연구 방향

(a) tone(mkt) and treasury bond rate(1year, 10year)



(g) tone(mkt) and treasury bond rate(1year, 10year)

(b) BOK policy rate and treasury bond rate(1year, 10year)



(h) BOK(rate) and treasury bond rate(1year, 10year)

4. Project Implications

프로젝트 후속 연구 방향

매일경제 2019년 06월 01일

시장 즉각반응...치솟은 채권값

"새로운 인하 사이클"
20년·30년물 금리도
기준금리 아래로 `뚝`

한국은행 금융통화위원회에서 기준금리 인하 소수 의견이 나왔다는 소식에 채권금리가 요동쳤다.
그동안 시장 일각에서 연내 기준금리 인하 가능성을 꾸준히 제기하긴 했지만 이번 금통위에서 소수 의견이 등장하자 금리 인하 전망이 더욱 힘을 얻는 분위기다. 시장에서는 하반기 기준금리가 하향 조정된다면 새로운 금리 인하 사이클의 시작으로 봐야 한다는 의견도 나온다.

10년물 금리 역시 전 거래일 대비 5.9bp 떨어진 1.682%로 마감했으며, 3년물 금리는 1.587%로 거래를 마쳐 1.6%를 하회했다.이날 이주열 한은 총재는 소수 의견과 금통위 의견은 다르다며 '선 긋기'에 나섰지만 채권시장에는 소수 의견 영향이 더욱 강하게 작용했다.
문홍철 DB금융투자 연구원은 "기존에도 채권시장에는 기준금리 인하 기대감이 녹아 있었는데, 소수 의견으로 더욱 강해진 것으로 보인다"며 "미·중 갈등이 장기화하며 불확실성이 더욱 커지는 상황에서 보험성 성격을 지닌 채권이 더욱 강세를 보이고 있다"고 설명했다.

V. 프로젝트 정리

4 프로젝트 정리

목 표

- Market Approach를 중심으로 논문을 충실히 따르며 직접 구현
- 비정형 텍스트 분석을 코드로 구현

잘한 점

- 목표 설정을 통해 짧은 기간 프로젝트 완수
- 적절한 업무분담으로 진행속도 향상

부족한 점

- Naïve Bayes Classifier 모델의 정확도 차이(69.8% ↔ 86%)
- 채권 분석 보고서 문서 개수 부족(3,095 ↔ 25,325)
- 시각화 및 포트폴리오 자료용 데이터 재처리로 일정 지체
- '왜 경제 신문사 기사가 포함되지 않았는가' 라는 의문을 해소하지 못함

eKoNLPy 전처리		
data		name
뉴스	이데일리	이진아
	연합인포	송지혜
	연합뉴스	김형석
MPB 보고서		이하영
채권보고서		홍현택

완료			
	to-do		
지혜		-이데일리, 연합인포 텍스트 전처리 코드	
형석		-연합뉴스 텍스트 전처리 코드	***↓ 코드 정리 완료 : 19.05.29 깃허브 링크 ↓***
현택		-채권보고서 텍스트 전처리	https://github.com/nlpbokproject/jh/tree/ekek0207-NLPBOK-version-1
진아		-이데일리, 연합인포 텍스트 전처리 코드	
하영		-의사록 텍스트 전처리 코드	
다같이			
no	**TO DO**		**데이터(code, csv 결과물) 위치 정리 완료된 코드 위치 : 52.141.32.158:8888/tree/3팀/codes_final**
1	**corpus 정제 (dataframe 형태로)**	뉴스기사 csv 파일로 변환	**COMPLETE 1.crawling-eadily.ipynb**
2		MPB 의사록 section 2,3만 뽑기	**COMPLETE 0_금통위의사록수집.ipynb - def function**
3		뉴스 기사 정제 작업	**./COMPLETE regexp-jh/190521 yeonhap_info regexp.ipynb**
4		채권보고서 정제 작업	csv : /3팀/DATA/NBC/bond_pre.csv
5		금리 데이터 수집(콜금리, 기준금리)	csv : /3팀/DATA/rate.csv
6	**Pre-processing**	corpus 정제(2005-2017) 데이터 확인	csv : /3팀/DATA/NBC/NBC_final/news_processed(2005~2017)_step1.csv
7		tokenizing	**COMPLETE 2&3_token_ngram.ipynb** csv : 3팀/DATA/NBC-merged.csv
8		pos 품사태깅	
9		lemmatization	
10		불용어 처리	
11	**Feature selection**	word list 작성	
12		n-gram화 (5-gram, 15번 미만 나온 n-gram 삭제)	
13	**market approach (머신러닝)**	콜금리 데이터 준비(등락률에 따른 라벨링)	**COMPLETE 4.0-1_mkt.approach_labeling+ngramunigram.ipynb** csv : 3팀/DATA/NBC/NBC_step1/*.csv
14		step1. tokens + n-gram 컬럼 합치기	
15		step2. counting / 조건부확률 계산	**COMPLETE 4.2_mkt.approach_counting.ipynb** csv : /3팀/DATA/NBC_step2-190529.csv
16		step3. 매파/비둘기파 dictionary	**COMPLETE NBC-step3(making dictionary).ipynb** 서버/3팀/DATA/NBC_step3(dictionary)_final-190529.csv
17		step4. 문장, 문서 별 tone 분류	**COMPLETE 4.4_mkt_approach_tone.ipynb** csv : /3팀/DATA/tone-simple-190529(final).csv

<프로젝트 단계별 목표 설정 스프레드 시트>

1.3 프로젝트 일정

Schedule

구분	일 정				
	5/5~5/11	5/12~5/18	5/19~5/25	5/26~6/1	6/2~6/3
논문분석					
데이터 수집					
데이터 전처리					
모델링					
시각화					
발표자료					

4 프로젝트 정리

분석 도구

- 언어 : Python
- 개발환경 : Jupyter Notebook
- 라이브러리 : Pandas, Numpy, Sklearn, mecab_eKoNLPy,

결과물

- Git Hub : **https://github.com/nlpbokproject/jh**
- Google Data Studio
  : **Figure 1. Text Data**
  : **Figure 2. MP sentiment and Macroeconomic Variables**
  : **Figure 3. tone(mkt) and MPB Minutes Top5 keywords**

참고 논문

- Deciphering Monetary Policy Board Minutes through Text Mining Approach
- Ngram2vec- Learning Improved Word Representations from Ngram Co-occurrence Statistics
- Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora

참고사이트

- 한국은행 http://www.bok.or.kr/
- e-나라지표 http://www.index.go.kr/main.do?cate=6
- 픽사베이 https://pixabay.com/ko/
- 한국자금중개 http://www.kmbco.com/bond/
- 한국거래소 http://www.krx.co.kr/
- Economic Policy Uncertainty http://www.policyuncertainty.com/

THANK YOU