[xwMOOC R Meetup 9회, 세션5 라이트닝 Talk]

– 머신러닝 기반 뉴스기사 문맥분석 엔진 및 딥러닝 분석 Case –

서강대 머신러닝 Lab 황문기 교수

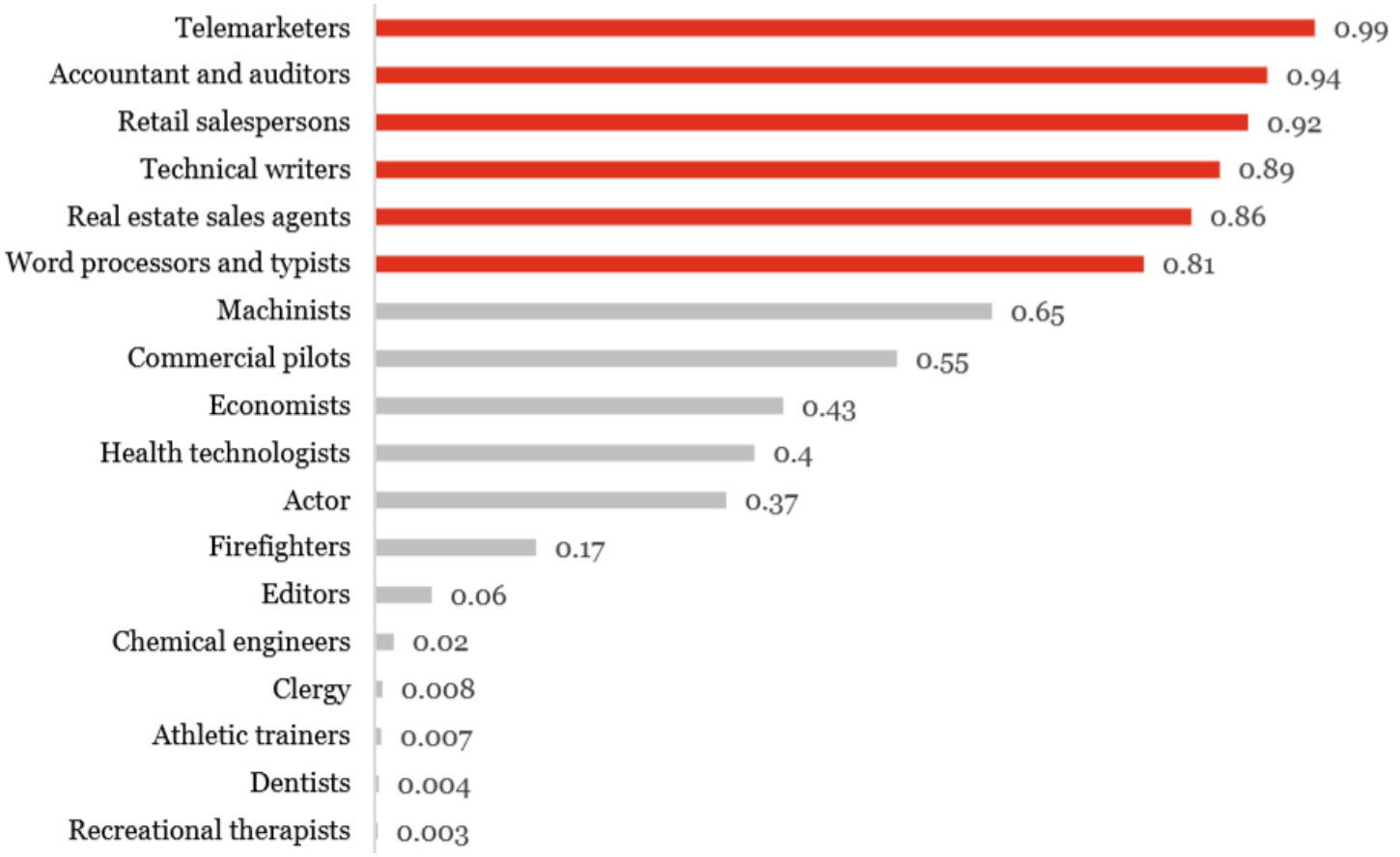
2018. 04. 25

Contents

1. Job vs. Robot

로봇이 대체할 직업들...

Fig.1) Probability Robots Will Take Over Jobs In The Next 20 Years (1=Certain)



Source: "The Economist, The future of employment: how susceptible are jobs to computerisation?"

2. 비정형 빅데이터

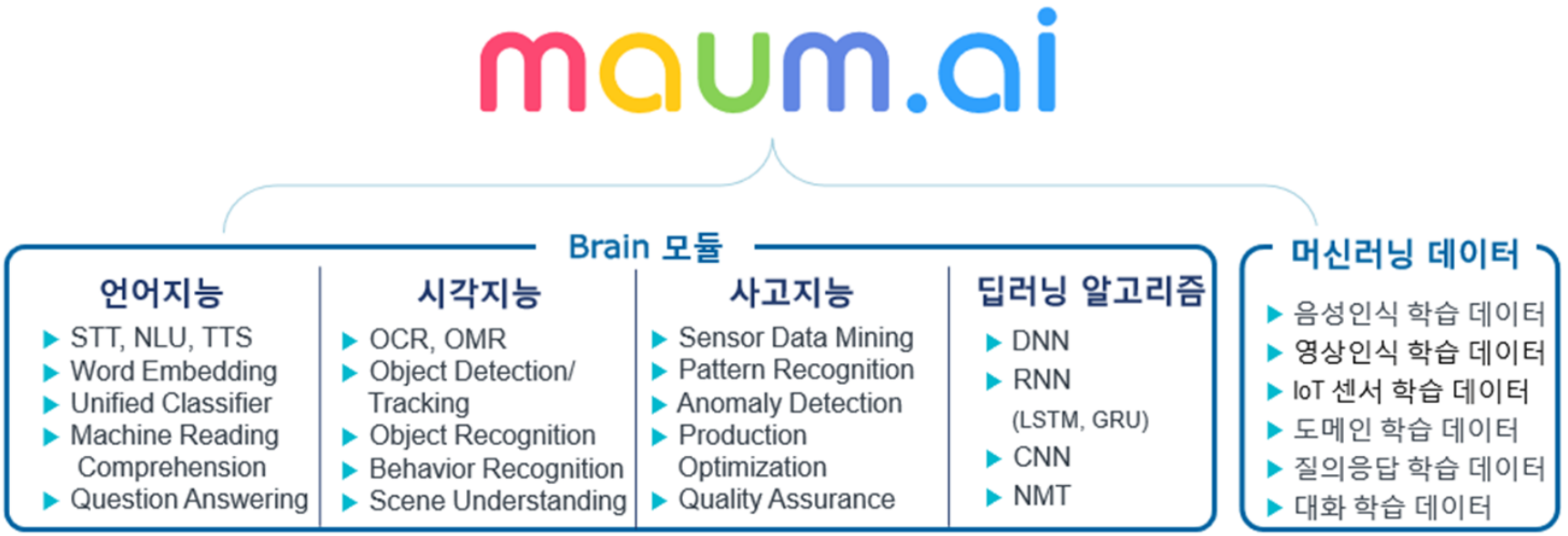
비정형 Text 빅데이터 (금융, 경제, 미디어, 전문가 분석)

국내 외 시장 데이터 (가격, 거래량, 수급)	**기업 데이터** (재무, 실적, 연혁, 사업, 자본금 변동, 주주 변동)	**뉴스** 등 문서 (뉴스,공시,IR,애널리포트)
경제 데이터 (국내외 주요 경제지표, 통계청, 한국은행 데이터)	**글로벌 지표 데이터** (곡식,에너지,귀금속, 환율,유가 등)	**전문가 분석** 정보 (알파)

전체 DBU 30년+ / 1,571,112,904 건+

3. AI 지능과 머신러닝 플랫폼

AI 플랫폼 서비스 사례



자료) 2018 PwC-MindsLab Business Forum 2018.03.14

4. 시장에서 각광받을 3가지 AI 기술

1. 고급 자연어 이해(Advanced Natural Language Understanding) & 가상 비서(Virtual Assistant), 2. 사고지능(Machine Intelligence), 3. 시각지능(Image Recognition)

자료) 2018 PwC-MindsLab Business Forum, Minds Lab 박동규 2018.03.14

5. 인공지능 산업적용

1. Smart Automation

제품의 혁신

2. Customer Care

고객과 연결

운영 최적화

직원 능력 강화

3. Robot Process Automation

4. Employee Training & Education

자료) 2018 PwC-MindsLab Business Forum 2018.03.14

6. 직원 능력 강화 - 채봇 서비스 Case

삼성전자 Nexshop Training Chatbot

https://youtu.be/TMll3oZtSrk

Contents

1. Demo 엔진

PageRank 알고리즘으로 키워드추출, 핵심문장 요약, 딥러닝 CNN으로 뉴스 카테고리 분석과 긍·부정 분석을 Clicker 리타겟팅 프로젝트를 수행

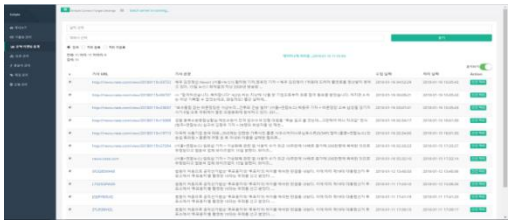
	1. 키워드 추출	2. 핵심문장 요약	3 카테고리 분류	4 긍·부정 분석
	▪ **휴리스틱 알고리즘 PageRank** - 랭킹 - 비중(weight) - 퍼센트 (%)	▪ **휴리스틱 알고리즘 PageRank** - 요약율 조정가능 (10~50%)	▪ **뉴스 카테고리 딥러닝 (CNN) 알고리즘** - 18개 - 24개 :	▪ **뉴스 댓글 딥러닝 (CNN) 알고리즘** - 18개 - 24개 : 82.7% (110만)
성능 지표	▪ 뉴스 컨텐츠 내의 핵심 키워드를 추출하여 비율로 표시 ▪ 사용자 키워드 매칭과 컨텐츠 카테고리 분류에 활용	▪ 장문의 뉴스 컨텐츠를 10% 이내의 분량으로 요약	▪ 사전학습된 데이터를 기반으로 새로운 뉴스 컨텐츠 생성 시, 자동 혹은 추천 분류 수행	▪ 뉴스 컨텐츠의 긍/부정 여부를 판단하여 지수로 제공 ▪ 뉴스댓글 분석을 통한 컨텐츠 만족도 측정
주요 결과	▪ Precision : 28 ▪ Recall : 38 ▪ F-Measure : 32	▪ Precision : 28 ▪ Recall : 38 ▪ F-Measure : 32	▪ Accuracy ▪ 82.7% (110만건)	▪ 극성(polarity) ▪ 90%(110만건)

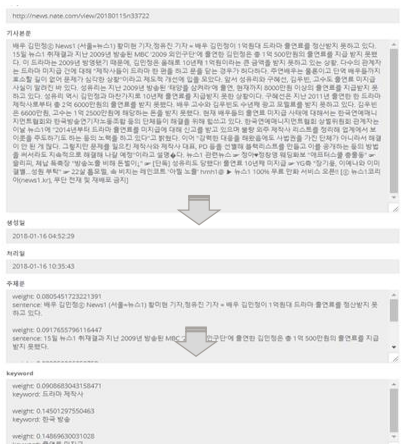
2. 시스템 구성

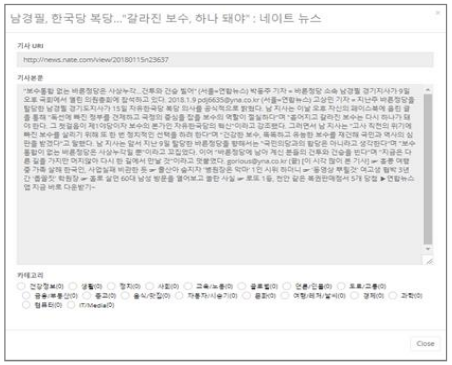
뉴스기사 크롤링 서버 ☞ 문맥 엔진 ☞ 머신러닝 분류기(학습/분석) ☞ 관리자 화면 구성

문서 서버 | 머신러닝 서버 (학습/분석) | 관리자 화면

1. 문서 서버
2. 문맥엔진 (키워드, 핵심문장)
3. 딥러닝 분류기 (카테고리 분류)
4. 분석 및 모니터링







카테고리 수	데이터 건수	모델링 방법	Layer 수	Train/Test 비율
6	107,306건	CNN 알고리즘	5	0.77 : 0.33
18	107,306건	CNN 알고리즘	5	0.77 : 0.33
18	269,044건	CNN 알고리즘	5	0.77 : 0.33

Train/Test 비율	학습시간 (최고 성능 도달 시간)	학습 환경	분석 시간(초)		예측 정확도
			CPU	GPU	
0.77 : 0.33	약 48분	CASE 1	9.37	4.68	78.8%
0.77 : 0.33	약 1시간	CASE 1	9.25	5.03	76.02%
0.77 : 0.33	약 2시간 45분	CASE 2	10.32	5.24	79%

정확도	정확도(Accuracy)	*성능평가 보고서*
	Precision	상동
	Recall	상동
	F1 Measure	상동
성능평가	모델 처리속도	상동
	모델 추론속도	상동
	모델 서비스속도	상동

- **문서 데이터 서버**
  - 온라인 : STR 문서(hwp), XML
  - 문서변환기 : hwp → text
- **문맥분석 엔진**
  - **핵심문장** : 요약율 (10~50% 조정)
  - **키워드** : 가중치(weight) 빈도, 퍼센트(%)
  - **불용어**, 유의어, 동의어 等
- **딥러닝 분류(Classification)**
  - 딥러닝 : CNN 분석
  - 알고리즘 : 휴리스틱 알고리즘
- **문서 카테고리 분류**
  - 자동분류 성능평가
  - Recall
  - Precision
  - F1 Measure

3. Demo

‘국제자금세탁방지기구’ 관련기사 (한국경제 2018년 2월 26일자)

http://news.hankyung.com/article/201802260445Y

FATF, 아이슬란드 기준 미달로 제재 회부...中 의장국 때 韓도 평가 받아
국제자금세탁방지기구(FATF)가 가상화폐를 이용한 자금세탁 위험이 커졌다고 판단, 이를 예방하기 위한 국제 기준을 강화하기로 했다.
26일 금융위원회 산하 금융정보분석원(FIU)에 따르면 지난 18~23일 프랑스 파리에서 열린 FATF 총회에서 회원국들은 이같이 의견을 모았다.
회원국들은 소유자 신원 확인이 어려운 전자지갑, 무작위 거래를 일으키는 믹서(Mixer) 등으로 가상화폐 거래의 익명성과 자금세탁 위험성이 커졌다고 우려했다.
FATF는 2015년 마련된 '가상화폐 가이드라인'을 최근 상황에 맞춰 개정하고, 국제 기준을 강화하는 가상화폐 대응계획을 다음 달 열리는 G20(주요 20개국) 재무장관 회의에 보고하기로 했다.
이번 회의에서 우리나라는 최근 은행들을 상대로 가상화폐 관련 금융거래에 자금세탁 방지 의무를 엄격히 적용토록 한 가이드라인을 소개해 큰 관심을 받았다고 FIU는 전했다.
FATF는 활동 경과와 향후 계획 등을 담은 G20 재무장관 회의 보고서를 채택했다.
보고서에는 FATF의 주요 활동 중 하나로 연구·교육기관인 'TREIN(Training and Research Institute)'이 명시된다.
부산에 있는 TREIN은 올해 5월 FATF의 민·관 전문가 회의(JEM)를 개최한다.
이 자리에서 가상화폐 관련 논의가 구체화할 것으로 FIU는 전망했다.
FATF는 유엔 협약 및 안전보장이사회 결의와 관련한 금융 조치 이행을 위해 만든 기구다.
우리나라 등 경제협력개발기구(OECD) 회원국을 비롯한 35개 국가와 2개 국제기구가 정회원이다.
이번 총회에선 아이슬란드가 국제기준 이행이 미흡한 것으로 나타나 고위험·비협조 국가 제재를 담당하는 FATF 산하 국제협력점검그룹(ICRG) 제재 절차에 회부됐다.
국제 기준이 2012년 강화되고 나서 FATF 회원국 상호평가에서 정회원국이 제재 대상에 오른 것은 이번이 처음이다.
아이슬란드는 1년 유예기간에 미흡한 부분을 개선하지 않을 경우 '자금세탁 위험국가'로 공표된다.
각국은 아이슬란드 기업·개인과 금융거래가 제약되거나 금지된다.
이주열 "물가상승압력 크지 않아...통화완화 조정 신중히 판단"
FIU는 "이 경우 아이슬란드는 국가신인도 하락, 금융시장 불안, 경제적 부담 등이 발생할 것"이라며 "우리나라도 내년 상반기부터 시작될 상호평가에 잘 대비해야 한다"고 강조했다.
FATF의 31기(2019년 7월~2020년 6월) 의장국으로는 중국이 선출됐다.
중국은 일본(20기), 홍콩(23기), 우리나라(27기)에 이어 아시아 국가 중 4번째 FATF 의장국이 된다.
FIU는 "중국이 의장국으로서 FATF를 주도하는 시기에 우리나라에 대한 평가 결과가 총회에 회부될 가능성이 크다"며 "한·중·일이 유사한 시기에 평가받는 만큼, 상호 교류와 협력을 강화해야 한다"고 말했다.

4. Demo(계속)

1 키워드 추출(keyword extraction) 결과

키워드 추출 Demo ver.

FATF, 아이슬란드 기준 미달로 제재 회부...中 의장국 때 韓도 평가받아
국제자금세탁방지기구(FATF)가 가상화폐를 이용한 자금세탁 위험이 커졌다고 판단, 이를 예방하기 위한 국제 기준을 강화하기로 했다.
26일 금융위원회 산하 금융정보분석원(FIU)에 따르면 지난 18~23일 프랑스 파리에서 열린 FATF 총회에서 회원국들은 이같이 의견을 모았다.
회원국들은 소유자 신원 확인이 어려운 전자지갑, 무작위 거래를 일으키는 믹서(Mixer) 등으로 가상화폐 거래의 익명성과 자금세탁 위험성이 커졌다고 우려했다.
FATF는 2015년 마련된 '가상화폐 가이드라인'을 최근 상황에 맞춰 개정하고, 국제 기준을 강화하는 가상화폐 대응계획을 다음 달 열리는 G20(주요 20개국) 재무장관 회의에 보고하기로 했다.
이번 회의에서 우리나라는 최근 은행들을 상대로 가상화폐 관련 금융거래에 자금세탁 방지 의무를 엄격히 적용토록 한 가이드라인을 소개해 큰 관심을 받았다고 FIU는 전했다.
FATF는 활동 경과와 향후 계획 등을 담은 G20 재무장관 회의 보고서를 채택했다.

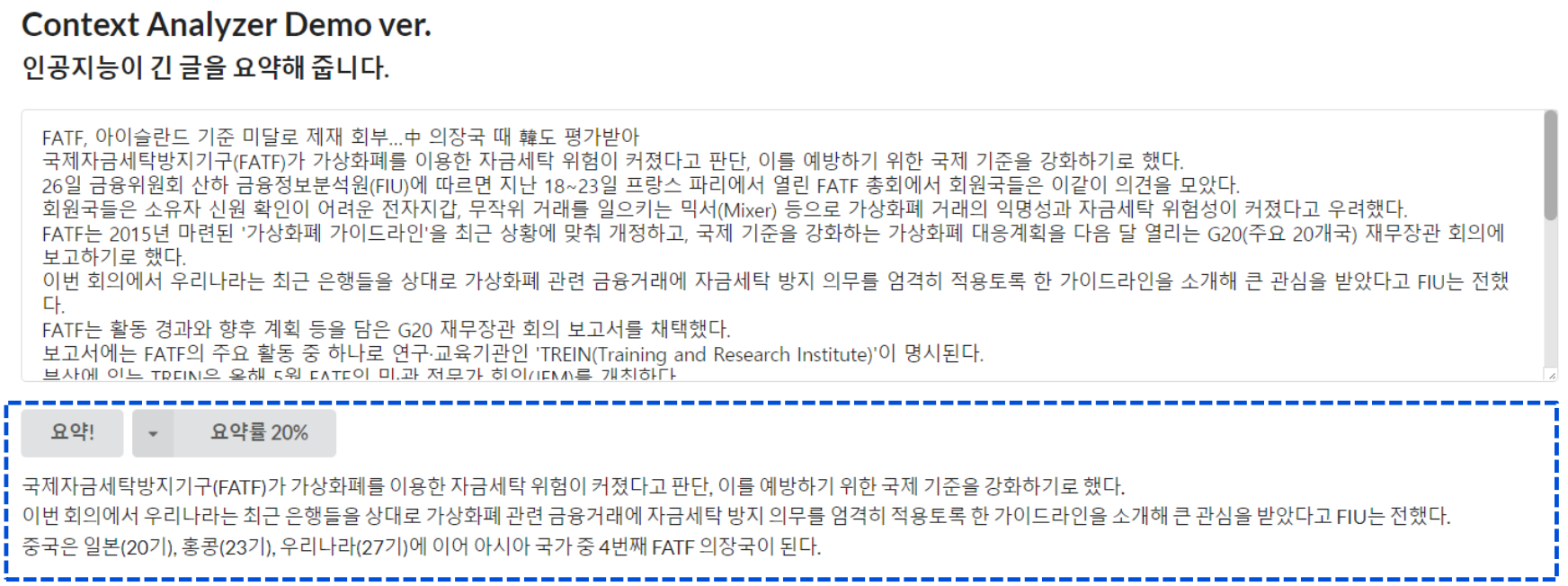
submit

가상 화폐 0.197545
자금 세탁 0.169375
금융 거래 0.163832
위험 국가 0.103528
회의 0.069778
국제 0.068338
평가 0.067486
관련 0.064329
나라 0.063130
아이슬란드 0.059847

Demo 서버) http://sgcslab.com:4500/

5. Demo(계속)

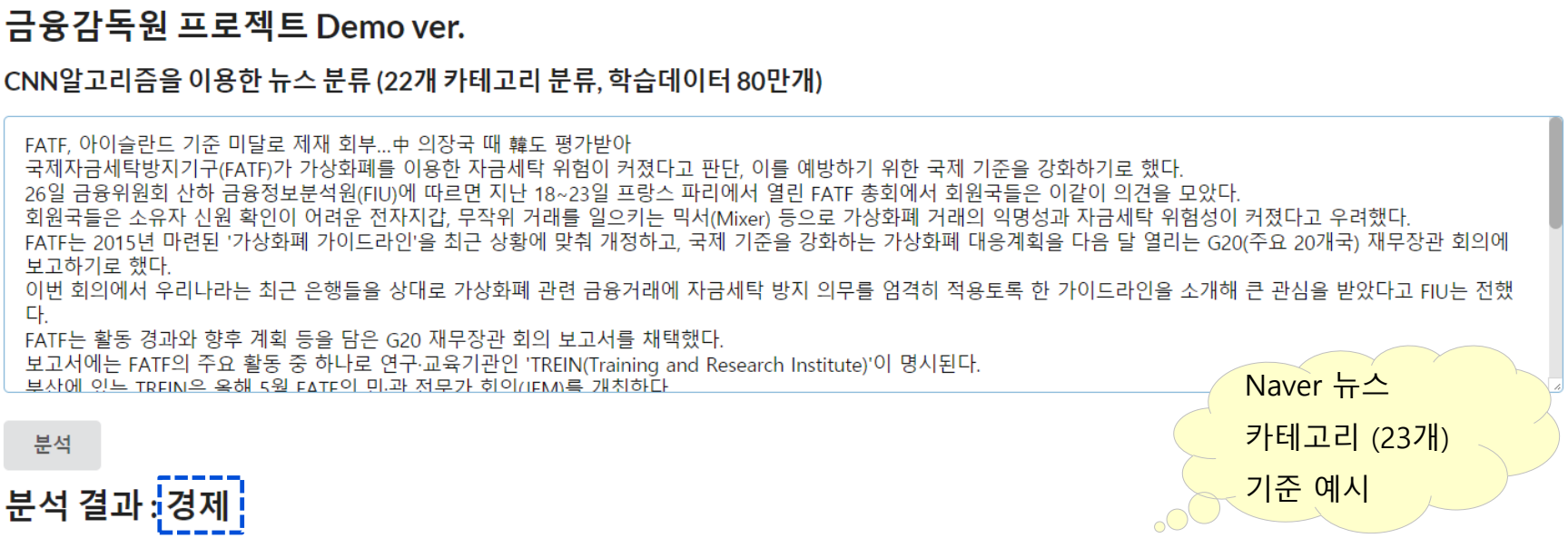
2 문장요약(sentence summarization) 결과



Demo 서버) http://sgcslab.com:3000/

6. Demo(계속)

3 카테고리 분류(category classification) 결과



Demo 서버) http://52.78.25.65:7777/

7. Demo(계속)

4 긍·부정 분석(Sentiment Analysis) 결과

SYSMETIC 긍부정 판별 엔진

극성탐지를 이용한 긍부정 판별 엔진

소규모 자영업자의 몰락이 눈앞에 보이는것 같아 안타깝습니다.좋은 정책이 수립되길 희망 합니다.

분석

긍부정 강도는 -100부터 100까지 나옵니다.

부정

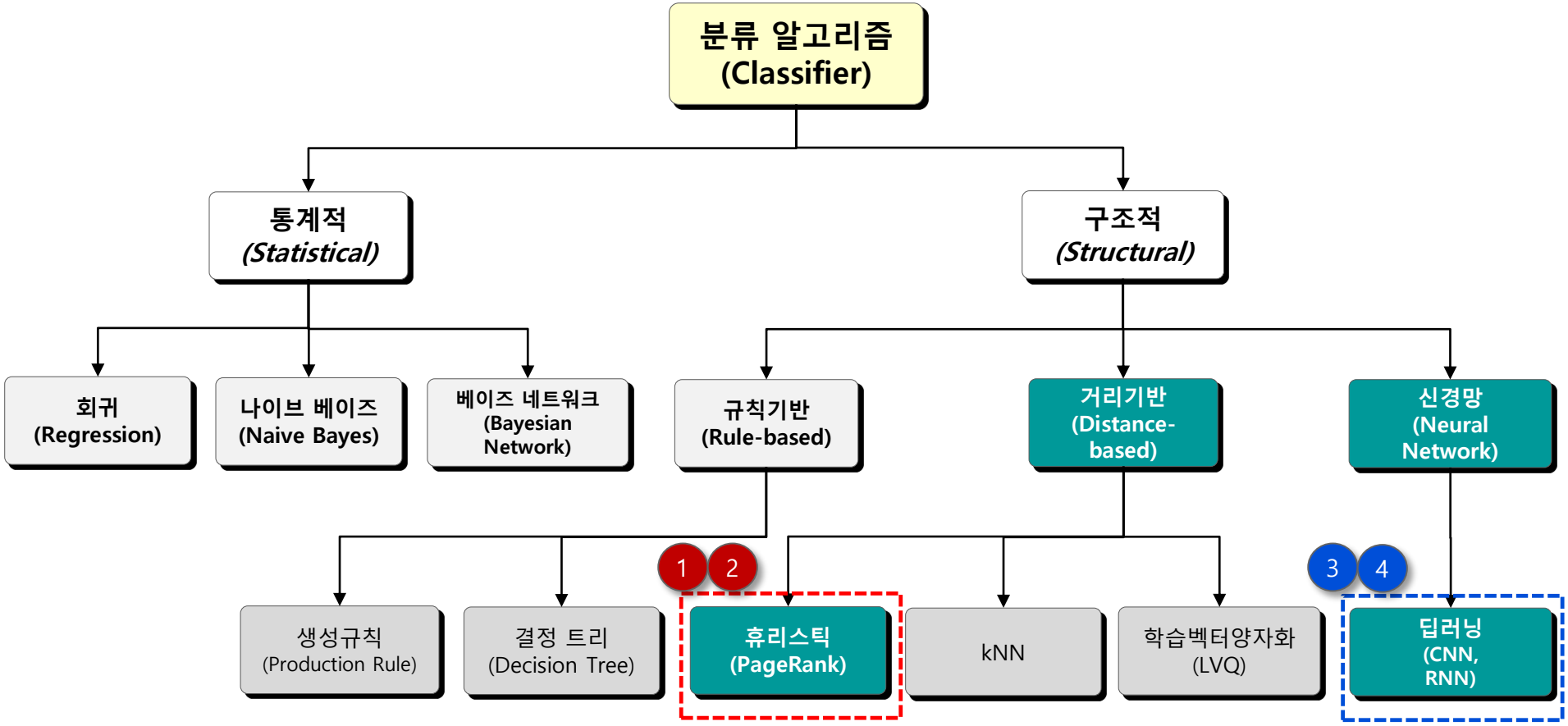
News Polarity : -1.67

Demo 서버) http://moya.ai/sentiment

7. 적용 알고리즘

1 키워드추출(keyword extraction), 2 문장요약(sentence summarization), 3 카테고리 분석(category classification), 4 감성분석 알고리즘이 개발·적용

그림1) 디자인 기준 분류 알고리즘 체계

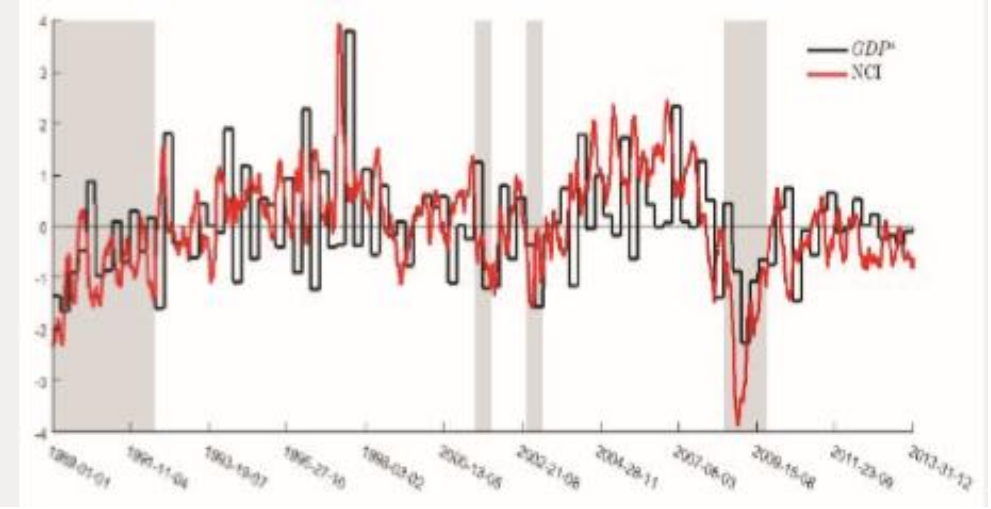


Contents

1. 경기 지수 Case

뉴스기사 긍·부정 분석 ☞ 경기동행지수

NCI와 GDP 비교



자료) Thorsrud (2016a)
주 : 1) 회색 음영은 경기침체기

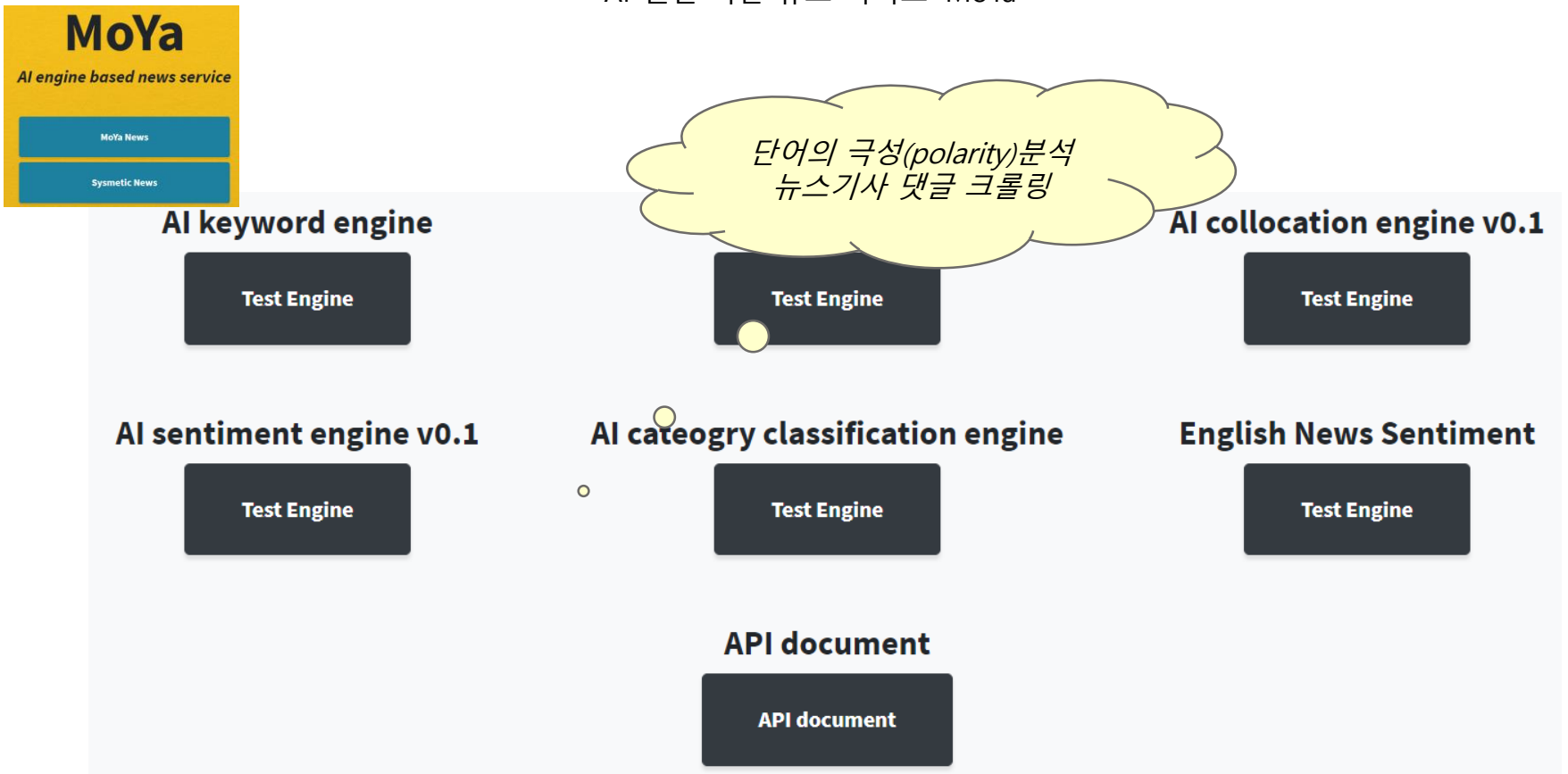
뉴스기사 경기동행지수

- 신문기사로부터 추출한 텍스트 정보도 경기순환지수 및 분기 GDP 성장률을 추정 Case
- Thorsrud는 노르웨이의 주요 일간 경제신문인 Dagen Naringsliv(DN)의 뉴스기사를 Latent Dirichlet Allocation(LDA) 모형을 이용하여 기사에 나타나는 단어들을 바탕으로 신문기사를 여러 주제(topic)들로 분해하고 각 주제가 기사에 언급된 빈도수를 계산
- 하버드 IV-4 심리사전에 정의된 긍정/부정 단어목록 을 이용하여 기사에서 각 주제에 대해 사용된 단어들을 긍정단어와 부정단어로 분류하고 그 차이를 계산
- 이러한 과정을 통해 신문기사의 텍스트 자료는 최종적으로 어떤 주제에 대한 기사가 많았는지, 각 주제에 대한 기사의 전반적인 분위기는 어떠했는지가 반영 된 시계열 자료로 변환
- Thorsrud (2016a)는 시계열 자료로 표현된 텍스트 정보를 동적인자모형(Dynamic Factor Model, DFM)에 적용하여 일별 경기동행지수(Newsy Coincident Index of Business Cycles, NCI)를 추정

2. AI 뉴스 서비스 산학기업

AI엔진기반 뉴스 서비스 벤처기업 Moya
- 뉴스매체 크롤러, 형태소분석기(Mecab), 문맥엔진, 카테고리, 긍·부정 딥러닝 엔진 等

AI 엔진 기반 뉴스 서비스 'MoYa'



사이트) moya.ai

감사합니다

서강대학교
SOGANG UNIVERSITY

황 문기

산학협력교수, 혁신과 경쟁 연구센터
책임교수, 서강 머신러닝 Lab

서울시 마포구 백범로 35(신수동), 남덕우경제관 803호 우) 04107
Mobile : 010-7685-3889 E-mail : mkhwang@sogang.ac.kr
moonkihw@naver.com