

TEXT MINING for PRACTICE

Python을 활용한 비정형 데이터 분석 - WEEK 02
텍스트 데이터 실무 활용사례

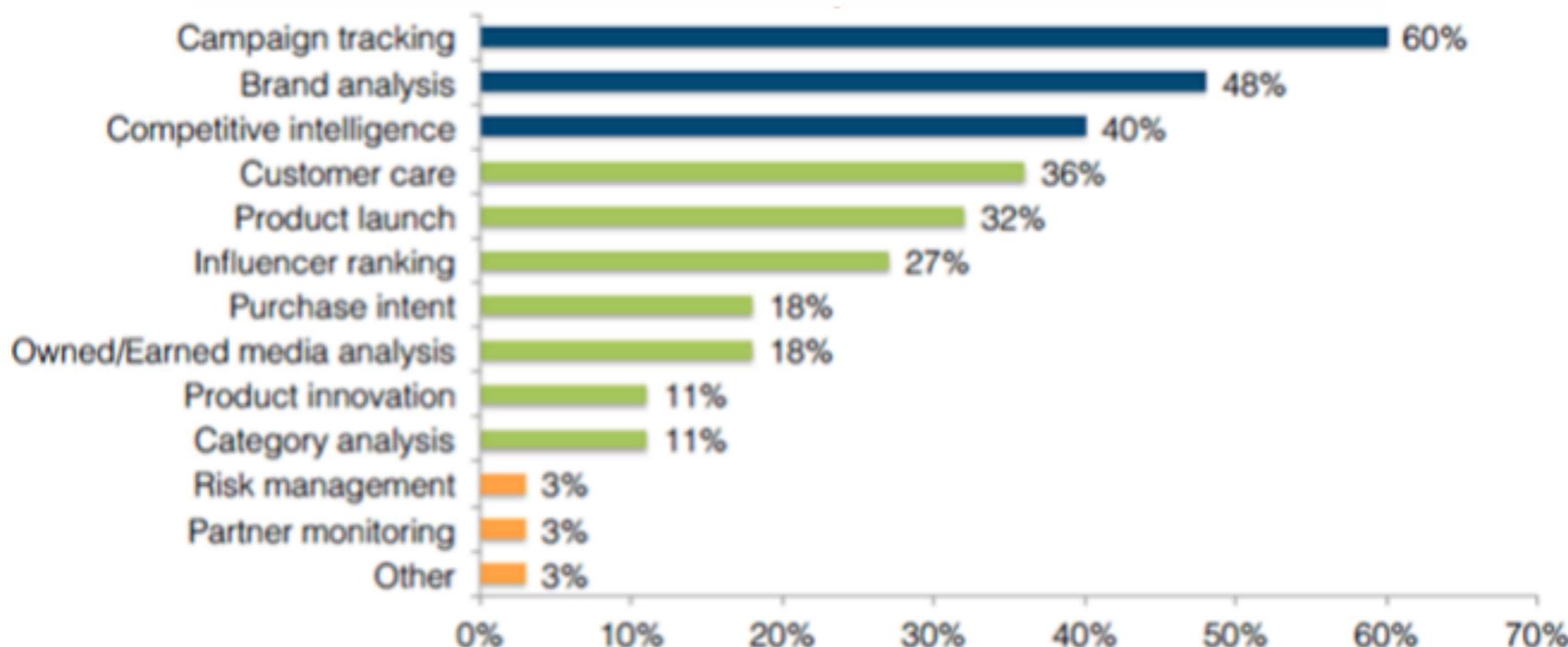
연세대학교 | 서중원

텍스트 마이닝 활용분야

활용분야	내용
경쟁 정보 (Competitive Intelligence)	<ul style="list-style-type: none">경쟁사 정보 수집에 들어가는 작업을 효율화시장 모니터링, 관련 자료수집, 비교 등을 자동화상품에 대한 고객들의 만족도조사를 효율적으로 진행
인적 자원 관리 (Human)	<ul style="list-style-type: none">조직 구성원들의 직무 만족도 등을 텍스트 데이터를 이용해 모니터링이력서 관리 등 대규모 텍스트 데이터를 분석해야 하는 경우에 활용
고객관리 서비스 (Customer Care Service)	<ul style="list-style-type: none">고객 커뮤니케이션에서 발생한 텍스트를 이용하여 고객경험을 향상시키기 위한 자료로 활용신속한 자동 응답제공으로 콜 센터에 대한 의존도 완화 (Predictive)
시장분석 (Market Analysis)	<ul style="list-style-type: none">고객 반응 모니터링 및 잠재고객 탐색 (Predictive)기업의 브랜드 이미지 파악
리스크 관리 (Risk Management)	<ul style="list-style-type: none">문서관리, 취합, 검색 등을 도와주는 텍스트 마이닝 기반의 리스크 관리 소프트웨어 도입으로 리스크 관리 능력 향상
사이버 범죄 예방 (Cybercrime Prevention)	<ul style="list-style-type: none">텍스트 마이닝을 이용한 범죄 예방 애플리케이션 도입 (Predictive)개인 및 기관을 대상으로 한 인터넷 범죄 예방 (Predictive)
의학 (Medical)	<ul style="list-style-type: none">수많은 의학 서적/논문에서 필요한 정보만 추출하거나, 새로운 정보를 찾아냄질병과 유전자 사이의 관계를 네트워크화해서 새로운 관계를 규명
정책수립 (Policy Formulation)	<ul style="list-style-type: none">정부 정책에 대한 대중들의 온라인 의견 분석새로운 정책전략 수립 및 홍보 등에 활용

텍스트 마이닝 활용분야

Text Mining	Language Processing
Spam filtering	Siri (Apple) and Google Now
Document Classification	Language Understanding
Date/time event detection	Spelling Correction
Information Extraction	Statistical Language Modeling
(Web) Search engines	Website translation (Google)
Information Retrieval	Machine Translation
Watson in Jeopardy! (IBM)	"Clippy" Assistant (Microsoft)
Question Answering	Dialog System
Twitter brand monitoring	Finding similar items (Amazon)
Sentiment Analysis (Stat. NLP)	Recommender System



* Source : Florian Leitner, OUTDATED Text Mining 1/5: Introduction, 2014.7.7, <https://www.slideshare.net/asdkfjqlwef/statistical-text-mining-introduction-florian-leitner/>.

** Source : Hope Nguyen, Why Use Social Analytics?, 2014.2.12., <https://www.netbase.com/blog/why-use-social-analytics/>.

텍스트 마이닝 적용사례

1. 고객성향 및 경제상황분석

연합뉴스

벤츠 "E클래스 온라인 키워드는 고소득맞벌이·성공·카푸어"

입력 2019.08.05. 11:56

237



(서울=연합뉴스) 최윤정 기자 = 메르세데스-벤츠 코리아는 고급 세단 모델 E클래스와 관련한 온라인 키워드가 고소득 맞벌이 부부, 인테리어, 성공, 카푸어, 특별한날, 가성비, 브랜드 역사 7가지로 분석됐다고 5일 밝혔다.

벤츠는 10세대 E클래스가 국내 출시 3년 만에 10만대 판매 기록을 세운 데 맞춰서 소셜 빅데이터를 분석했다고 말했다.

벤츠는 다음소프트에 의뢰해서 2016년 1월부터 2019년 6월까지 블로그, 인스타그램 등에서 E클래스 와 관련해 나온 210억건이 넘는 데이터를 분석해 핵심 키워드를 추출했다.



<출처: E-클래스 소셜 빅데이터 분석 리포트, 메르세데스-벤츠 코리아 & 다음 소프트>

주식 커뮤니티 보면 코스피지수도 보인다

조선비즈 | 이종현 기자

입력 2018.12.10 10:19 | 수정 2018.12.10 12:08

주식 커뮤니티 게시글 18만여개 키워드 분석

10월 들어 코스피지수가 급락하자 온라인 주식 커뮤니티에 올라오는 글에서 '수익'이라는 키워드가 빠르게 줄어든 것으로 조사됐다. 대신에 '장외' '비상장' 같은 키워드는 등장하는 빈도 수가 늘었다. 코스피지수가 급락하자 투자자들이 다자이 스이보다 새롭고 트렌디한 차트를 찾느라 고민해기 때문으로 보인다

7월		8월		9월		10월		11월	
주식	1,662	주식	1,711	주식	3,399	주식	4,931	주식	10,298
투자	1,220	투자	1,469	투자	2,392	투자	3,451	투자	3,774
종목	942	수익	1,041	수익	2,008	주식시장	2,527	장외	3,353
주식시장	867	주식시장	954	주가	1,548	시장	2,398	거래	3,074
수익	740	종목	927	종목	1,459	미국	1,898	비상장	2,814
시장	663	시장	865	주식시장	1,360	종목	1,872	시장	2,575
매수	563	상승	669	시장	1,337	수익	1,709	주식시장	2,332
상승	527	주가	602	부동산	1,141	주가	1,667	기업	2,309
기업	483	미국	530	상승	1,083	상승	1,472	종목	2,092
주가	482	부동산	501	정보	1,078	금리	1,464	수익	1,976
미국	424	투자자	494	매매	1,069	기업	1,441	정보	1,757
정보	410	가격	477	미국	873	부동산	1,395	미국	1,688

최근 5개월 동안 주식 커뮤니티에 많이 등장한 키워드 /문데이터 제공

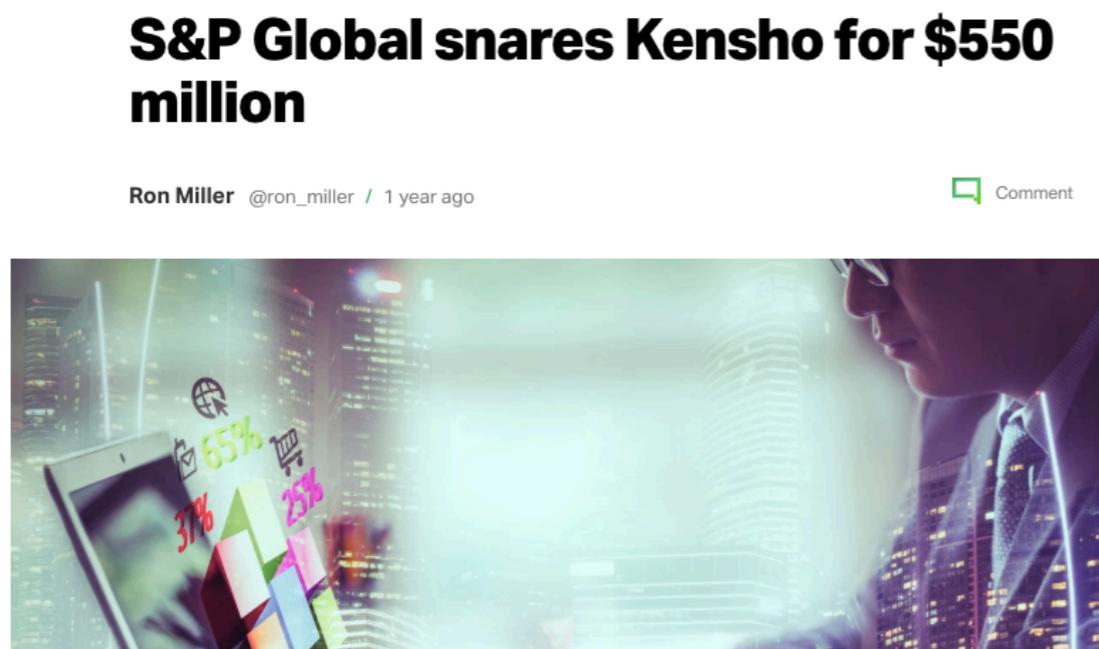
* Source: 최윤정(연합뉴스), 벤츠 "E클래스 온라인 키워드는 고소득맞벌이·성공·카푸어", 2019.08.05, <https://www.yna.co.kr/view/AKR20190805079500003>.

** Source : 이종현(조선비즈), 주식 커뮤니티 보면 코스피지수도 보인다, 2018.12.10., http://biz.chosun.com/site/data/html_dir/2018/12/10/2018121000888.html/.

텍스트 마이닝 적용사례

2. 로보 어드바이저 (Robo-advisor)

- ▶ 금융 및 법률 분야에는 이미 단계적으로 사용 되는 추세
 - 아마존의 가격이 떨어지면 넷플릭스는 어떻게 되나?
 - 공사판 인부인 아버지가 공사장 사다리차에서 추락하여 사망했을 때 업주에게 어떤 책임을 물을 수 있는지…?



S&P Global snares Kensho for \$550 million

Ron Miller @ron_miller 1 year ago

Comment

JOINS

중앙일보

오피니언

사설칼럼 만평 디지털컬전 e글중심

뉴스검색

[논설위원이 간다]로펌 간 한국 첫 AI 변호사…검사도 놓친 분석 '단 20초'

안혜리 기자

한국의 첫 인공지능(AI) 변호사가 지난 2월 대형 법무법인에 '취직'했다. 변호사만 150여 명인 국내 10위권 로펌인 대륙아주의 AI 변호사 '유렉스' 얘기다. 유렉스는 그동안 담당 변호사와 법률을 비롯한 여러 명이 짧게는 수일에서 길게는 몇달씩 걸려 작업하던 관련 법 조항 검토와 판례 분석 등 사전 리서치 업무를 20~30초만에 해치우는 괴력을 발휘하며 빠르게 업무에 적응하고 있다. 2016년 5월 미국 뉴욕의 100년 전통 로펌인 베이커앤호스텔러가 AI 변호사 로스(ROSS)를 처음 '채용'한 게 전 세계적으로 화제가 됐는데, 그로부터 불과 2년만에 우리에게도 AI 변호사가 현실로 다가온 것이다. 이 추세라면 변호사 상당수가 길거리에 나앉는 게 아니라는 암울한 전망마저 나온다.

추천기사

* Source: Ron Miller (Techcrunch), S&P Global snares Kensho for \$550 million, 2018.03.07, <https://techcrunch.com/2018/03/07/sp-global-snares-kensho-for-550-million/>.

** Source: 안혜리 (중앙일보), 로펌 간 한국 첫 AI 변호사…검사도 놓친 분석 '단 20초', 2018.04.06, <https://news.joins.com/article/22508494>.

텍스트 마이닝 적용사례

3. 챗봇 (Chatbot)

- ▶ 테크나비오는 '16~'21년까지 전 세계 챗봇 시장이 연평균 37% 이상 성장할 것으로 예측하였고, 주요 기대분야로 금융부문(BFSI: Banking, Financial services and Insurance), 정부부문, 리테일 및 이커머스 부문을 선정
- ▶ 또한 페이스북은 메신저 플랫폼 런칭 1년 만에('17년 4월 기준) 10만개의 챗봇을 만들었으며, 그랜드뷰리서치 연구결과에 따르면 2020년 챗봇 시장의 규모는 약 1조 3,600억 원에 달할 것으로 전망

구분	질의 응답형 챗봇	인공지능 상담 챗봇
대화방식	단방향 정보전달 : 사용자 질문에 대한 답변만 가능하며 추가 보완은 어려움	쌍방향 정보 교류 : 대화 에이전트가 주도적 으로 사용자와의 질문-답변 과정을 반복하여 부족한 정보를 보완
정보제공	포괄적 정보제공 : 일반적인 정보만 제공하기 때문에 자신에게 맞는 정보인지 스스로 판단 필요	맞춤형 정보제공 : 대화를 통해 사용자의 상황 정보를 인식하여 사용자에게 적합한 정보 제공
정보획득	사용자 스스로 질문 주도 : 사용자는 자신에게 적합한 정보를 얻기 위해 스스로 질문을 만들고 답변을 찾는 과정을 반복	대화를 통한 전문가의 도움 : 지식이없는사용자도 대화 에이전트의 가이드에 따라 원하는 정보에 빠르고 쉽게 접근

* Source: 한국정보화진흥원, "인공지능 기반 챗봇 서비스의 국내외 동향분석 및 발전 전망", 2018, <https://www.nia.or.kr/common/board/Download.do?bcIdx=20156&cbIdx=37989&fileNo=1>.

* Source: Venturebeat, "Facebook Messenger hits 100,000 bots", 2017.04.18. <https://venturebeat.com/2017/04/18/facebook-messenger-hits-100000-bots/>.

** Source: 테크엠. "챗봇, 새로운 인터페이스의 부상", 2017.11.13, http://techm.kr/bbs/board.php?bo_table=article&wr_id=4330.

텍스트 마이닝 적용사례



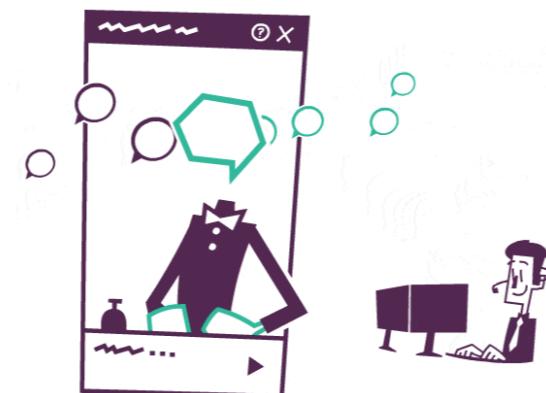
WHY BOOST SOLUTIONS BLOG ABOUT

GET STARTED

Customers can never wait

Boost your customer experience with an artificially intelligent virtual agent that makes everything fast and easy.

GET STARTED



Language ▾ Søk Logg inn

rvice



Recoanized in Gartner

Vi vil gjerne ha deg som kunde

Du velger den løsningen som passer best til ditt behov.

Bli kunde her



Er du medlem i et LO-forbund?

Da får du ekstra gode betingelser på blant annet lån, forsikring og eiendomsmegling.

Sjekk dine medlemsfordeler



Skriv meldinger her



* Source: Boost AI, <https://www.boost.ai/>

** Source: Spare Bank 1 <https://www.sparebank1.no/nb/sr-bank/privat.html>

텍스트 마이닝 적용사례

4. 가짜뉴스 (Fake News)

'진짜 같은 가짜뉴스' 만드는 AI 나왔다

오픈AI, 개발 성공…창립멤버 머스크 "난 관계없다"

이정현 미디어연구소 | 입력: 2019/02/18 13:03 -- 수정: 2019/02/18 13:14 | 인터넷

中 알리바바, AI로 '가짜뉴스' 가려낸다

1초만에 판별…정확도는 81% 수준

유효정 중국 전문기자 | 입력: 2019/04/02 08:19 -- 수정: 2019/04/02 09:59 | 인터넷

Fake News Detection on Social Media: A Data Mining Perspective

Kai Shu[†], Amy Sliva[‡], Suhang Wang[†], Jiliang Tang[§], and Huan Liu[†]

[†]Computer Science & Engineering, Arizona State University, Tempe, AZ, USA

[‡]Charles River Analytics, Cambridge, MA, USA

[§]Computer Science & Engineering, Michigan State University, East Lansing, MI, USA

[†]{kai.shu,suhang.wang,huan.liu}@asu.edu,

[‡]asliva@cra.com, [§] tangjili@msu.edu

* Source: 이정현, '진짜 같은 가짜뉴스' 만드는 AI 나왔다, 2019.02.19, <https://www.zdnet.co.kr/view/?no=20190218113741>.

** Source: 유효정, 中 알리바바, AI로 '가짜뉴스' 가려낸다, 2019.04.02, <http://www.zdnet.co.kr/view/?no=20190402062122>.

*** Source: Shu, Kai, et al. "Fake news detection on social media: A data mining perspective." *ACM SIGKDD Explorations Newsletter* 19.1 (2017): 22-36.

텍스트 마이닝 적용사례

5. 경제 트렌드 예측

한은 '텍스트마이닝 활용해 의사록 분석...기준금리 예측 가능'

전소영 기자 | 승인 2019.01.06 12:00 | 댓글 0

(서울=연합인포맥스) 전소영 기자 = 한국은행 금융통화위원회 의사록을 텍스트마이닝(text mining) 기법으로 활용하면 기준금리의 방향을 예측할 수 있다는 결과가 나왔다.

김수현 한은 경제연구원 국제경제연구실 부연구위원은 6일 '텍스트 마이닝을 활용한 금융통화위원회 의사록 분석' 보고서를 통해 이같이 밝혔다.

텍스트 마이닝은 대규모 텍스트 자료에서 육안으로 읽고 분석하기 힘든 정보를 추출하고 분석하는 기법이다.

중앙은행의 커뮤니케이션은 절제된 표현이 많아 일반적인 독해만으로는 커뮤니케이션에 내재된 정보를 추출하고 그 영향력을 등을 분석하는데 한계가 있다.

논문은 비정형 빅데이터 분석 기법 중 하나인 텍스트 마이닝을 활용해서 금통위 의사록에 담긴 어조를 추출해서 지수로 만들었다. 이를 이용해 기준금리 변동에 대한 설명력과 예측력을 검증했다.

의사록 어조 지수를 만들기 위해 논문은 감성사전을 먼저 구축했다.

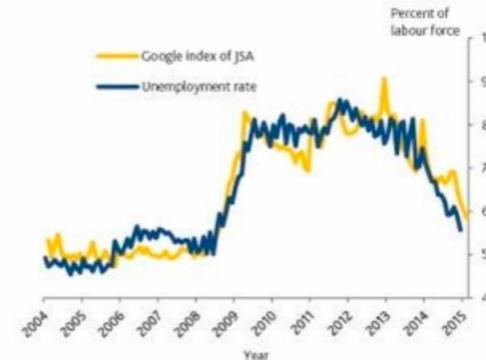
감성사전은 2005년 5월부터 2017년 12월 중 약 23만건의 신문기사와 채권 애널리스트 보고서, 금통위 의사록에서 추출한 형태소 조합을 분석했다.

한은, SNS·댓글·기사 수집해 경제 분석한다

조선비즈 | 조귀동 기자

입력 2017.03.02 08:46

Figure 1 Googling the labour market



Source: McLaren and Sharbhogue (2011).

미국에서 '구직수당(JSA)'나 관련 단어를 검색어로 입력한 빈도와 실제 실업률을 각각 연도별로 표시하면 변화 양상이 비슷하다./영란은행 연구 보고서

한국은행이 페이스북이나 트위터 등 소셜네트워크서비스(SNS), 네이버 등 인터넷 웹사이트에 올라온 댓글, 경제 관련 언론 기사 등을 분석해 경제 상황 해석에 활용한다.

* Source : 전소영(연합인포맥스), 한은 "텍스트마이닝 활용해 의사록 분석...기준금리 예측 가능", 2019.1.6., <http://news.einfomax.co.kr/news/articleView.html?idxno=4010420/>.

** Source : 조귀동(조선비즈), 한은, SNS·댓글·기사 수집해 경제 분석한다, 2017.3.7., http://biz.chosun.com/site/data/html_dir/2017/03/02/2017030200681.html/.

텍스트 마이닝 적용사례

6. 민원 업무 자동화

세계 최초 디지털 콘택트센터 구축한다...KB금융 스타링크 내달 출범

발행일 : 2018.09.17



[AD] 4월 25일 차세대 디스플레이 기술응용 세미나 개최

내달 '스타링크' 본격 상용화...민원처리 업무서 역할 재정립



<KB금융이 아날로그 콜센터를 빅데이터 기반 전문 상담센터로 업그레이드한다. 17일 서울 마포구 KB국민은행 스타링크 직원이 상담내역을 빅데이터화한 정보로 고객과 상담하고 있다. 김동욱기자 gphoto@etnews.com>

은행 창구를 가지 않아도 콘택트센터(일명 콜센터)를 통해 예금, 대출, 자산 관리가 가능해진다. 약 7~8단계에 이르는 자동응답전화(ARS) 기계음을 모두 없애고, 고객이 상담한 음성 내역은 텍스트로 전환돼 빅데이터로 쌓인다. 본부와 은행 창구가 이를 활용, 고객에게 먼저 서비스를 제시하는 '디지털 오퍼' 시대도 열린다. 텔레마케팅 전담 조직도 없어진다.

구분	내용
“ 콜센터 디지털화 ”	△인프라 구축(지능형 전화 연결, 콘텐츠 발송, 新상담관리 시스템 등) △빅데이터 분석 기반 마련(STT/TA 고도화) △채팅 상담 강화 : 다중(1명의 직원이 최대 3명 고객 응대), 멀티(유선+채팅) 상담 운영 △디지털 전문 인력 운영, 디지털화 업무(RPA) 확대
“ 스타링크 고도화 ”	△스타링크 브랜드화 △고객경험관리 프로세스, 비대면 고객관리체계 수립 △스타링크 Plus+ 서비스 시행 : 관리 사각지대 잠재 우수 고객 발굴 및 관리 △상담 직원 캐어 프로그램 구축
“ 비대면 영업 채널화 ”	△비대면 금융 마케팅 허브 시스템 구축 △비대면 대출 운영 집중화 △고객분석(STT/TA, Intelligent Routing 활용) 기반 개인화 마케팅 △비대면 고객 전담 관리 강화
“ 채널 간 연계 강화 ”	△챗봇 운영(4119/대고객) △그룹 계열사 콜센터 간 콘트롤타워 구축 △전문 상담 인력 양성 및 확보

텍스트 마이닝 적용사례

7. 스토리 작성기

StoryAi SEARCH ABOUT LOGIN SIGN UP

This A.I. can create any content. Just start it.

My name is Tommy Duffy. I am an upcoming sophomore at Conestoga High School. School starts in two days and I am sad that summer is coming to an end. G Create

The community has created 5,436 stories...

Machines are getting better at writing. They can finish the...
Machines are getting better at writing. They can finish the sentences we start writing. They ca...
By Anonymous and John Tales

I'm not a cat person, but I can see the point of this
Hey.. you are good for nothing...!?!? screamed by wife with the best possible decibel a human ear can...
10 By Anonymous and John Tales

Eden Hazard is the best football player in the world right now
Eden Hazard has created 70 goal moments in this BCL season. In our opinion he is the best football playe...
2 By Anonymous and John Tales

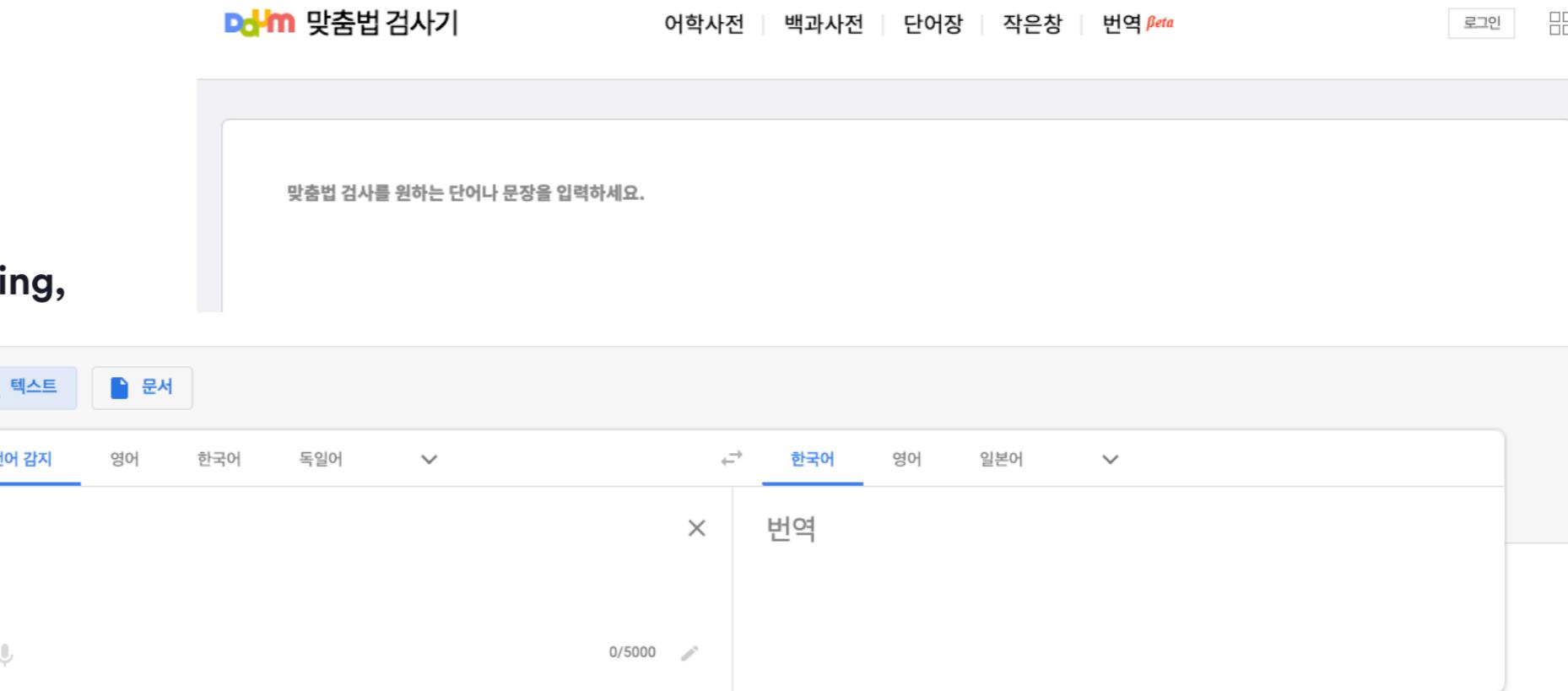
I believe we can do this
I want to meet her as soon as possible. I hope she also wants to.
...

The hamster ran through the forest
The hamster ran through the forest. He was looking for berries to eat. All

The Navy was on time and on budget for the X-23 project
All the preparation for the upgrade to the space port were on target. Delays

텍스트 마이닝 적용사례

8. 문법/맞춤법 검사기



The screenshot shows the Grammarly web interface. At the top, there is a navigation bar with the Grammarly logo, the text "Ddm 맞춤법 검사기", and links for "어학사전", "백과사전", "단어장", "작은창", and "번역 beta". On the far right are "로그인" and a user icon. Below the navigation is a large input field with the placeholder "맞춤법 검사를 원하는 단어나 문장을 입력하세요.". To the left of the input field, there is a sidebar with the text "Great Writing, Simplified" and "Compose better with Grammarly's writing assistant." It includes a "Compose" button, a "Compose with Grammarly" button, a "Text" tab (selected), a "Document" tab, a "Grammarly Assistant" section with a 5-star rating and "20 million people", and a "Add to Grammarly" button. The main input area has language selection dropdowns for "언어 감지" (selected), "영어", "한국어", "독일어", and "한국어" (selected for translation). A "번역" (Translation) button is also present. At the bottom right of the input area is a "Feedback" link. Below the input field are three circular icons with Korean labels: "기록" (Record), "저장됨" (Saved), and "커뮤니티" (Community).

<https://www.grammarly.com>

https://alldic.daum.net/grammar_checker.do

<https://translate.google.co.kr>

실무적용 사례:

뜻밖의 텍스트 마이닝 - 네오플



뜻밖의 텍스트마이닝 유저동향분석에서 인게임데이터까지

네오플 데이터분석팀

김대영 emotionalcode@gmail.com

NEXON COMPANY

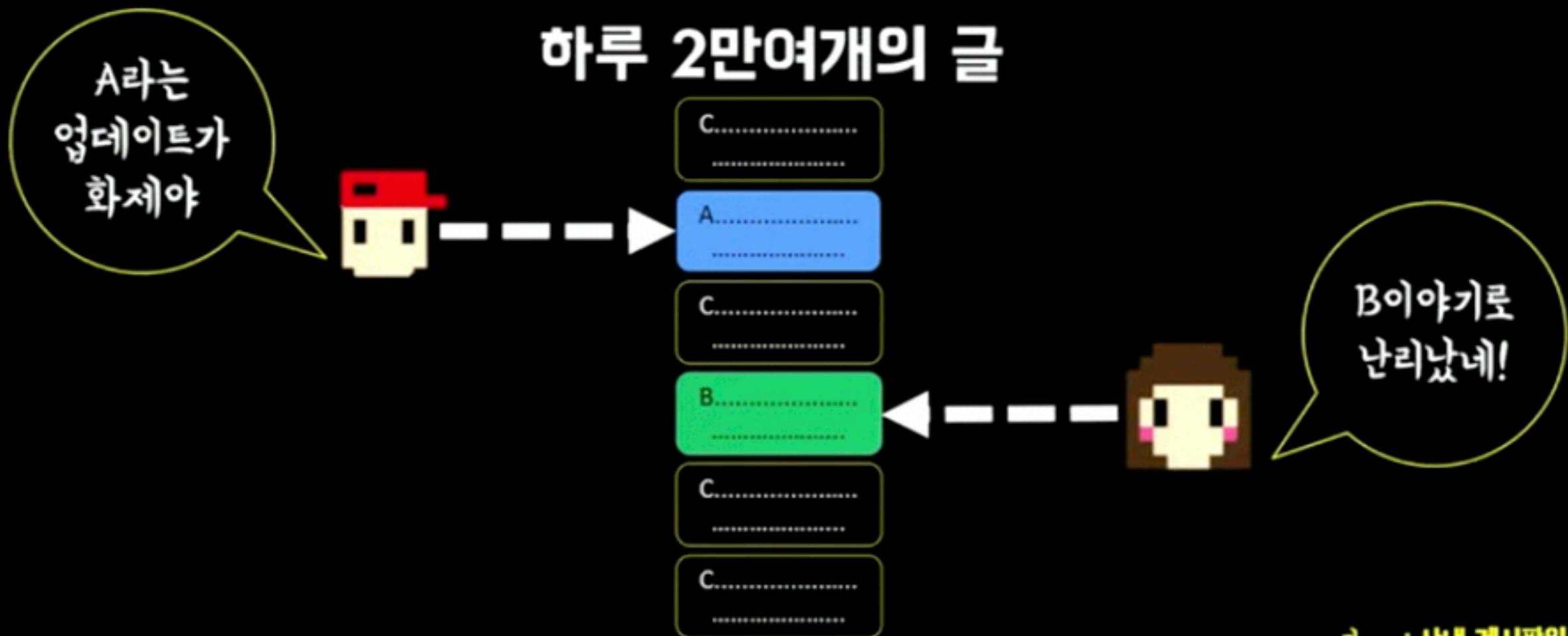


발표자 소개

**통계학, 언어학, 수학 전문가가 아닙니다.
자연언어처리 분야의 개발을 해 본 경험이 없습니다.
머신러닝을 다뤄본 적이 없습니다.**

NEXON COMPANY

유저동향파악을 보다 객관적으로



NDC NEXON DEVELOPERS CONFERENCE

- 기 : 사내 게시판의 한 아이디어
- 승 : 텍스트데이터의 시각화
- 전 : VOC시스템 개발
- 걸 : 인게임데이터 마이닝

텍스트레이터의 시각화

NLP HanNanum

(<http://semanticweb.kaist.ac.kr/home/index.php/HanNanum>)

워드카운팅



Wordcloud Tagxedo

(<http://www.tagcedo.com/>)

- 기 : 사내 게시판의 한 아이디어
- 승 : 택스트데이터의 시각화
- 전 : VOC시스템 개발
- 개 : 인게임데이터 마이닝

NEXON
DEVELOPERS
CONFERENCE

텍스트데이터의 시각화

**NLP
HanNanum**

(<http://semanticweb.kaist.ac.kr/home/index.php/HanNanum>)

워드카운팅

**Wordcloud
Tagxedo**

(<http://www.tagxedo.com/>)

오늘자
데이터로도
만들어주세요

지난주말 이후
현재까지
데이터로
만들어주세요

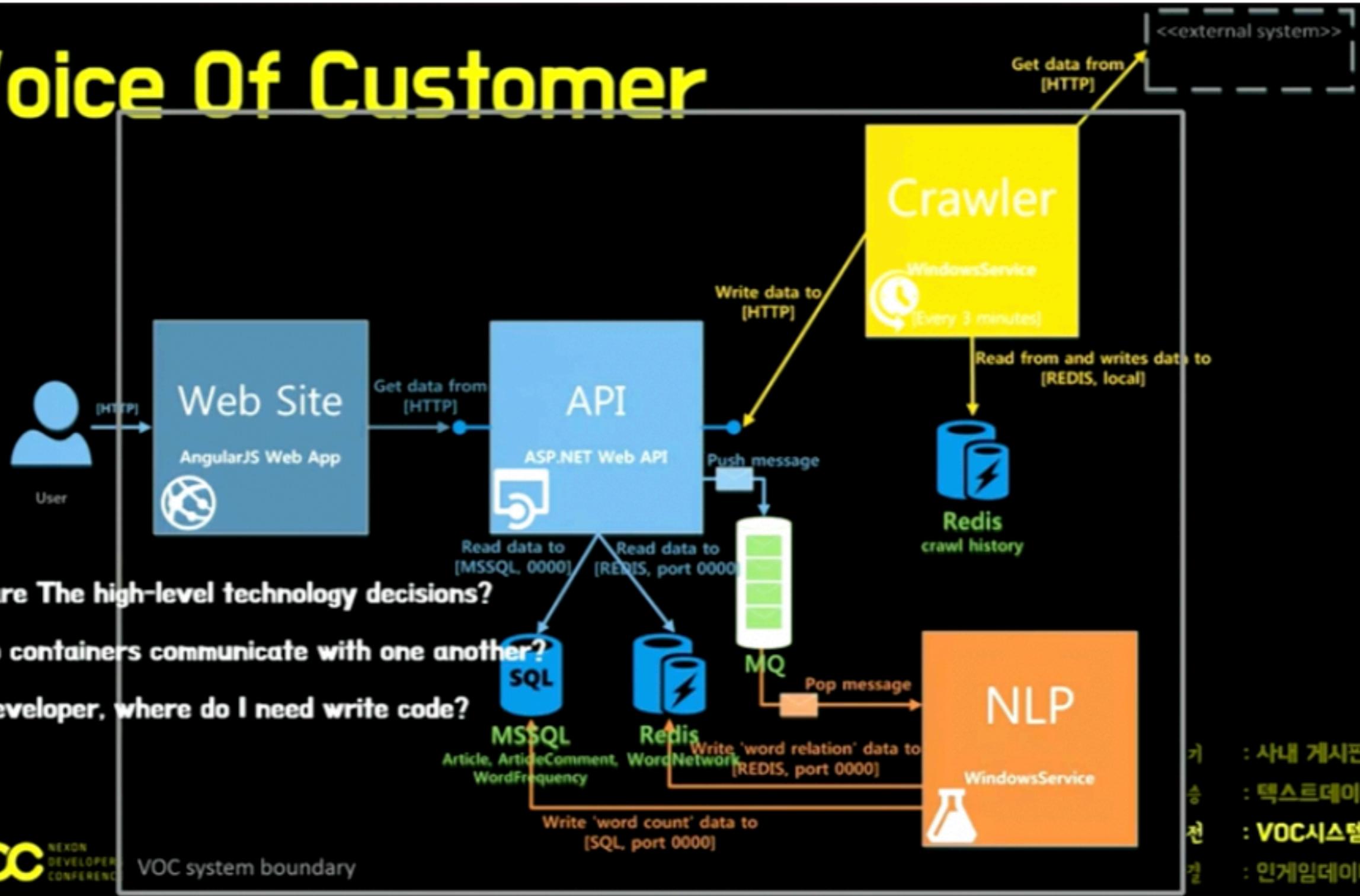
이번 업데이트
이후 데이터로
만들어주세요



이미지출처 - <http://blog.naver.com/jisun22ing/220506931005>

기 : 사내 게시판의 만 아이디
승 : 텍스트데이터의 시각화
전 : VOC시스템 개발
결 : 인게임데이터 마이닝

Voice Of Customer



2017-03-19 (2)

온라인 사이트

Word Filter

제3장 해석한 풀타를 저장

전체선택

전체선택해제

frequency.

- 루크 (822)
 - 탄핵 (739)
 - 어록 (604)
 - 일 (597)
 - 만화 (577)
 - 달 (542)
 - 레이드 (535)
 - 무기 (513)
 - 덤 (485)
 - 안전 (394)
 - 가족 (373)

액 - 2017-03-10 (금)

▣ **탄핵 인용으로 인한 굴드 시세 변동??**

학습시간 2017-03-10 (월) 오후 11:36:00

오늘 **날짜** 일정으로

한국 시장에 대한 관심을 이끌 수 있는 경쟁력을 갖춘다.

萬人以上之勞軍及慰問團等項，均係各該處之公事。

● 이날을 기다렸네요

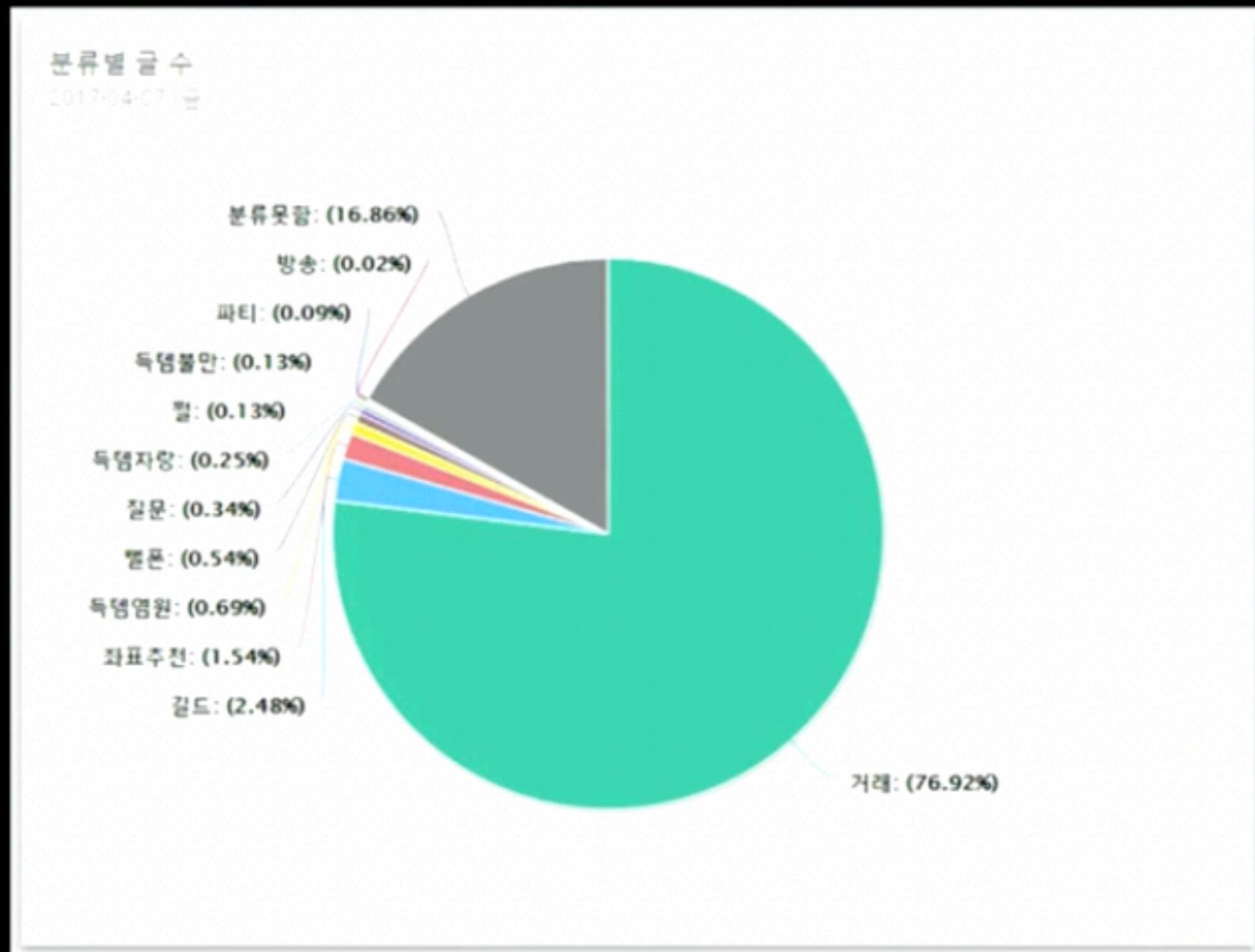
한국시간 2017-03-10 (월) 오후 10:33:07



148-150



하트비트 메가폰 - 텍스트분류 결과



E.O.D