

자연어 처리 텍스트 마이닝 Overview

자연어처리 텍스트마이닝

자연어 처리란?

자연어 처리 입문

인공지능(AI)의 시작

자연어 처리에 대한 관심은 1950년 앤런 튜링(Alan Turing)이
이른바 튜링 테스트(Turing Test)가 등장한
"Computing Machinery and Intelligence"라는 논문을 발표하면서
본격적으로 시작되었다.

“인간이 컴퓨터와 대화하고 있다는 것을 깨닫지 못하고
인간과 대화를 계속할 수 있다면
컴퓨터는 지능적(Intelligence)인 것으로 간주될 수 있다.”

- 앤런 튜링 -



앨런 매티슨 튜링은 영국의 수학자, 암호학자, 논리학자이자 컴퓨터 과학의 선구적 인물이다. 알고리즘과 계산 개념을 튜링 기계라는 추상 모델을 통해 형식화 함으로써 컴퓨터 과학의 발전에 지대한 공헌을 했다. 튜링 테스트의 고안으로도 유명하다.

자연어 처리(NLP)란?

자연어 처리란?

자연어 처리(自然語處理) 또는 자연 언어 처리(自然言語處理)는 인간의 언어 현상을 컴퓨터와 같은 기계를 이용해서 모사할 수 있도록 연구하고 이를 구현하는 인공지능의 주요 분야 중 하나이다.



WIKIPEDIA
The Free Encyclopedia

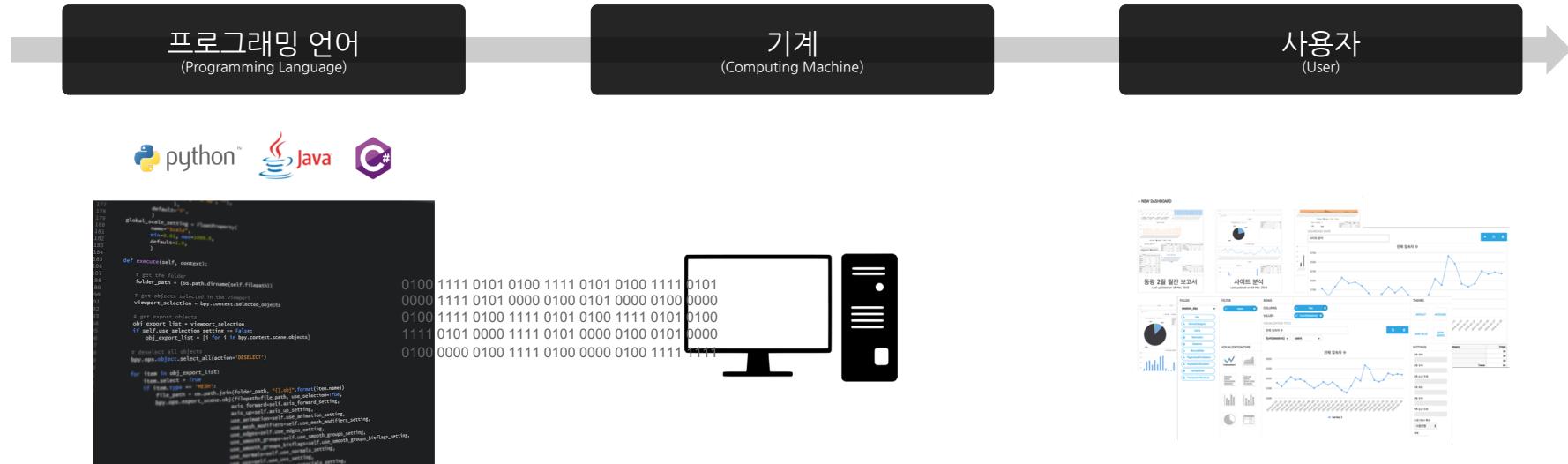
정보 검색, QA 시스템, 문서 자동 분류, 신문기사 클러스터링, 대화형 Agent 등 다양한 응용이 이루어지고 있다.

https://ko.wikipedia.org/wiki/%EC%9E%90%EC%97%B0%EC%96%B4_%EC%B2%98%EB%A6%AC

자연어 처리(NLP)란?

전통적인 프로그래밍 언어

: 기계(혹은 컴퓨터)를 실행하기 위해서 기계가 이해할 수 있는 프로그래밍 언어로 명령을 내리고, 그 결과를 사용자에게 전달



자연어 처리(NLP)란?

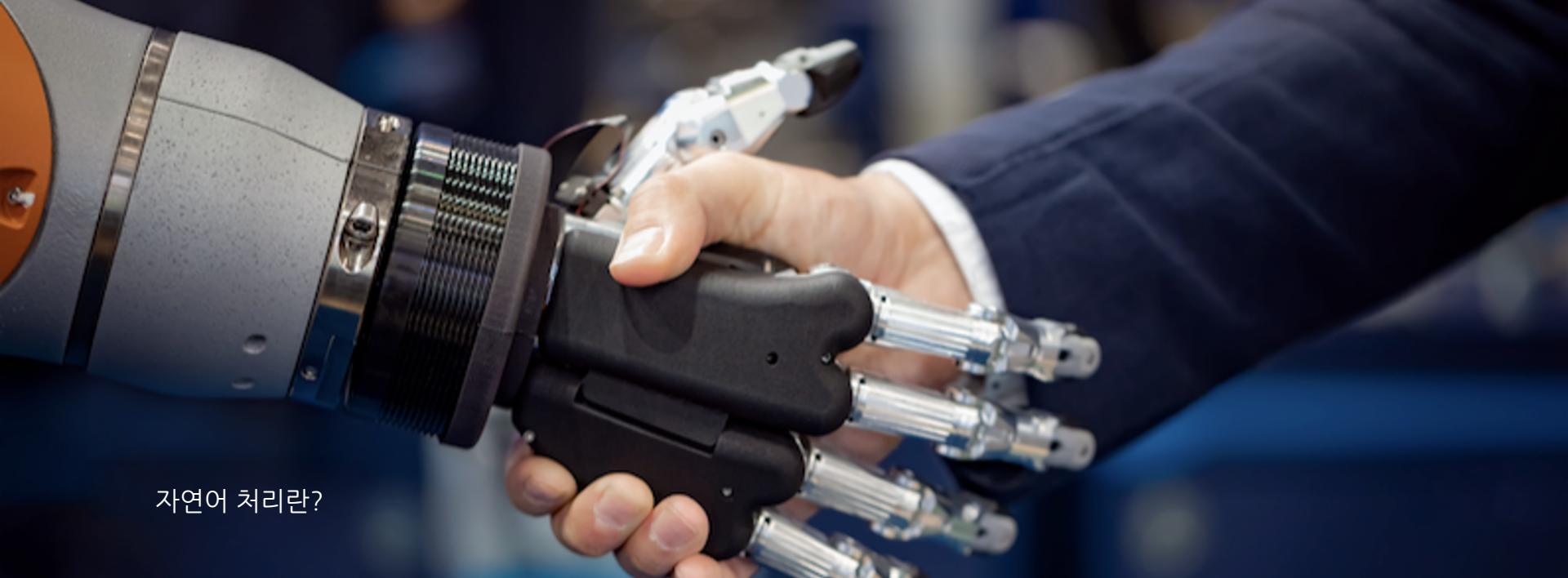
자연어 처리

: 인간의 언어(=자연 언어)로 명령을 내리면 기계가 자연어 처리(NLP)를 통해 이해하여 처리하고, 그 결과를 사용자에게 전달



자연어 처리란?

자연어 처리(NLP)란?



자연어 처리란?

전통적인 프로그래밍 언어가 인간이 기계 언어로 기계(=컴퓨터)를 이해시키는 것이었다면,

자연어 처리는 기계가 인간의 언어(=자연 언어)를 이해하여 소통하는 것을 말한다.

자연어 처리, 왜 관심 가져야 하나

자연어 처리 입문

비정형 데이터의 중요성

- 인터넷과 모바일의 발달로 온라인 매체에 대한 데이터가 급격하게 증가
- 전 세계에서 생성되는 데이터 70~80%가 비정형 데이터(뉴스, SNS, 블로그, 기타 문서 등)
- 의사 결정을 내림에 있어 비정형 데이터 분석은 필수적

정형 데이터 (Structured Data)

사전 정의된 모델을 통해 구조화된 데이터

예시 : 엑셀, RDMS



비정형 데이터 (Unstructured Data)

내부 구조를 갖지만 미리 정의된 데이터 모델을 통해 구조화되지 않음.

예시 : 텍스트파일, 전자메일, 소셜미디어, 웹사이트



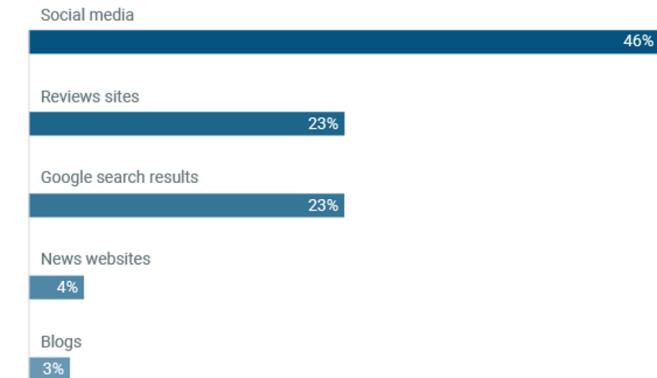
Company Name	Country	KEY FINANCIALS			
		Revenues (billions)	Profits (billions)	Assets (billions)	Market Value As of 3/28/18 (bn)
1. ICBC	China	\$145.300	\$45.700.0	\$4.210.300	\$115.000
2. China Construction Bank	China	\$143.200	\$37.200.0	\$3.831.300	\$85.700
3. JPMorgan Chase	United States	\$118.200	\$26.500.0	\$2.826.300	\$87.700
4. Bank of America	United States	\$125.200	\$35.700.0	\$752.700	\$495.800
5. Agricultural Bank of China	China	\$125.300	\$24.900.0	\$3.436.300	\$194.100
6. Bank of America	United States	\$103.200	\$26.300.0	\$2.526.300	\$715.900
7. Wells Fargo	United States	\$102.100	\$21.700.0	\$1.915.400	\$265.300
8. Apple	United States	\$847.000	\$83.300.0	\$367.300	\$308.800
9. Bank of China	China	\$116.200	\$26.400.0	\$3.204.200	\$158.800
10. Ping An Insurance Group	China	\$143.800	\$15.300.0	\$1.066.400	\$195.400
11. Royal Dutch Shell	Netherlands	\$127.400	\$12.200.0	\$412.700	\$146.300
12. Toyota Motor	Japan	\$285.200	\$32.300.0	\$473.300	\$260.700
13. ExxonMobil	United States	\$291.100	\$25.400.0	\$348.800	\$244.100
14. Samsung Electronics	South Korea	\$224.400	\$41.200.0	\$295.200	\$255.800
15. AT&T	United States	\$118.200	\$30.000.0	\$446.300	\$198.300
16. Volkswagen Group	Germany	\$173.300	\$15.100.0	\$321.400	\$115.400

온라인 데이터의 중요성

- 포브스(Forbes)지에 따르면 “97%의 기업이 온라인 평판 관리(ORM, Online Reputation Management)가 매우 중요하다”
- 온라인 평판은 비정형 데이터(뉴스, SNS, 블로그 등)를 분석하여 평가 가능
- 분석 대상과 관련된 비정형 데이터를 수집하고 자연어 처리를 통해서 문서 내 인사이트 도출 가능

예) 제품에 대한 시장의 반응 (긍정, 부정, 중립)

Top 5 Online Platforms for Monitoring Brand Reputation



None of the above = 1%
Percent of total respondents, N=224 digital marketers
Source: Clutch 2018 Online Reputation Management Survey

Clutch

소통 패러다임의 변화

- 인터페이스가 점차 인간처럼 자연스러운 방법으로 개선되어 감
- 대화형 인터페이스로 변화
- 예) 인공지능 스피커, 인공지능 챗봇 등

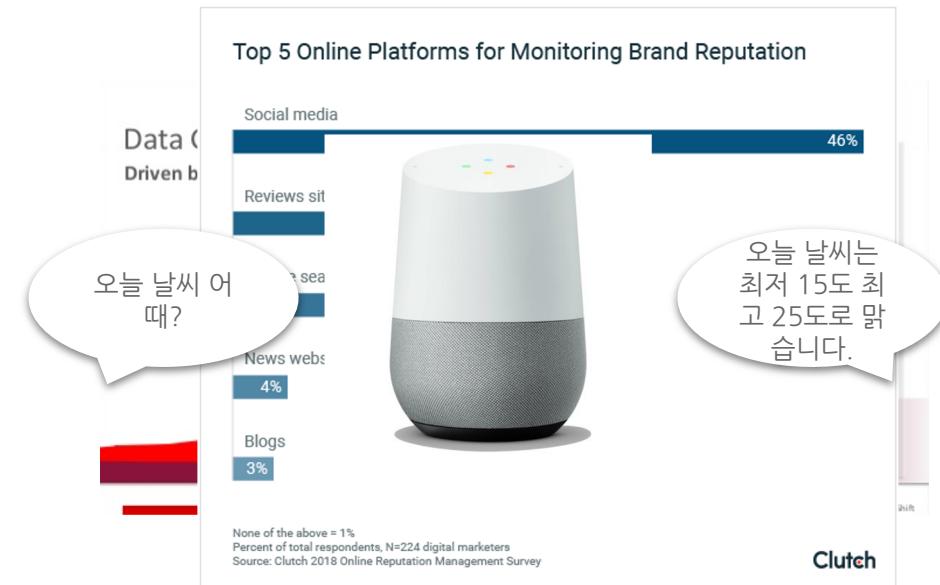


자연어 처리, 왜 관심 가져야 하나

1. 비정형 데이터의 중요성

1. 온라인 데이터의 중요성

1. 소통 패러다임의 변화



자연어 처리가 어려운 이유

자연어 처리 입문

언어의 모호성 - 동음이의어

	동형이의어	동음이형어
의미	철자와 발음이 모두 같은 동음이의어	철자는 다르나 발음이 같은 동음이의어
예시	Turn right (부사, 오른쪽) That's right (형용사, 옳은)	I went to the sea(바다) to see(보다) my friend.
어려움	품사 및 의미파악 어려움	음성인식 어려움

언어의 모호성 - 다의어

하나의 단어가 여러개의 의미를 가질 수 있음

Bolt



Apple



개체명 인식의 어려움

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

1

구어와 문어의 차이

2

띄어쓰기에 어려움

3

청자와 화자의 관계에 따른 높임법

4

동음이의어, 운율적 요소에 따른 의미 변화

5

주어·서술어·목적어 등의 빈번한 생략

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

1

구어와 문어의 차이

문어 : 정돈된 문법을 사용하고 있어 애매모호함이 적음

구어 : 완벽한 문법이나 형식적인 의미에 구애받지 않고 사용

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

2

띄어쓰기에 어려움

아버지 가방에 들어가신다

아버지가 방에 들어가신다

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

3

청자와 화자의 관계에 따른 높임법

김 교수님한테 나 먼저 간다고 문자 보내줘.”

“네 알겠습니다. ‘나 먼저 간다’고 문자를 보냅니다.”

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

4

동음이의어, 운율적 요소에 따른 의미 변화

	의문문	평서문
영어	Did you eat?	I ate
한국어	밥 먹었어?	밥 먹었어

한국어 자연어 처리가 더! 어려운 이유

구글코리아 전산 언어학자 팀에서 발표한 한국어 자연어처리가 힘든 5가지 이유

5

주어·서술어·목적어 등의 빈번한 생략

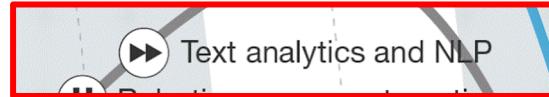
문장의 필수 요소(주어, 서술어, 목적어 등)가 생략되면서 겪는 분석의 어려움

자연어 처리 전망

자연어 처리 입문

자연어 처리 기술 전망

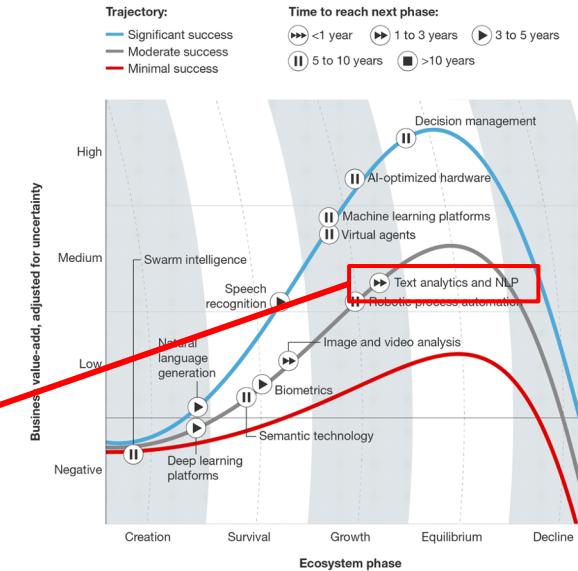
- 인공지능 기술 내에서도 자연어 처리는 빠르게 성장
- 기술전문 매체 테그레이더(TechRadar) 자료를 보면 인공지능 기술 중에서도 가능, 빠르게 성장하는 기술



FORRESTER RESEARCH

TechRadar™: Artificial Intelligence Technologies, Q1 '17

TechRadar™: Artificial Intelligence Technologies, Q1 2017

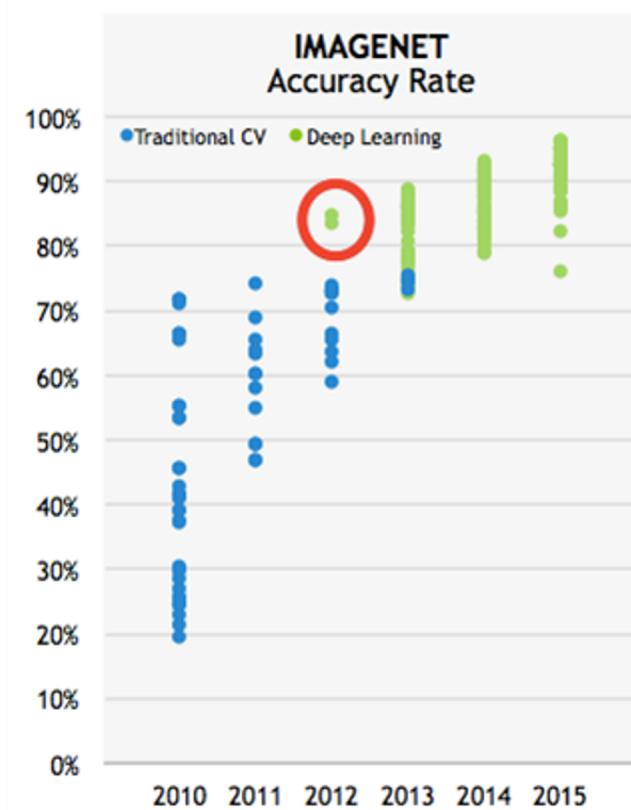
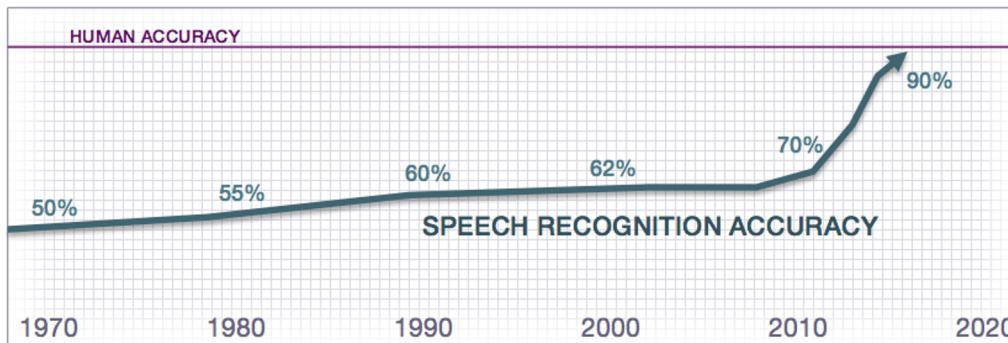


129161

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

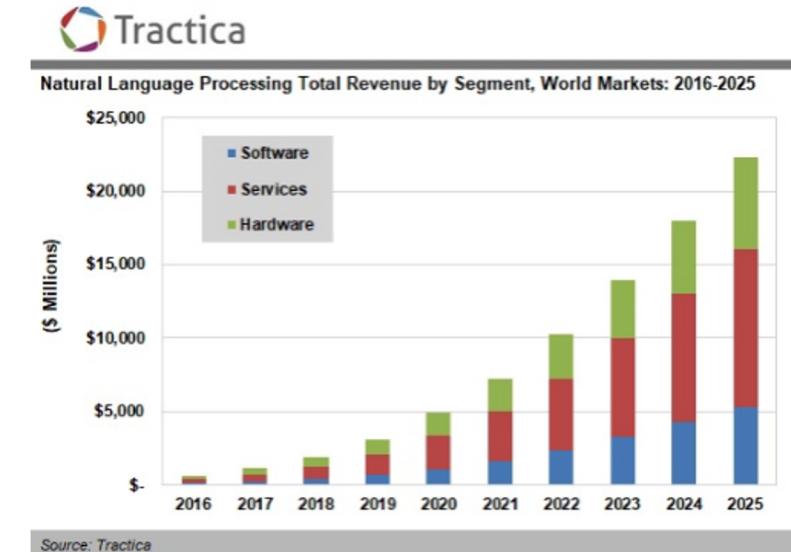
자연어 처리 기술 전망

- 이미지 프로세싱 - 이미지 넷 (2012)
- Speech Recognition - AM → LSTM
- NLP - Seq2Seq



자연어 처리 시장 전망

- 2016년 \$500M(한화 5,600억원)에서
- 2025년 \$22.3B(한화 24.9조)로 증가 (10년 내 44.6배 성장)
- 자연어 처리 시장 성장 동력은 “수요 증가”
 - 인공지능 스피커와 같은 스마트 장치 사용 증가
 - 웹 및 클라우드 기반 비즈니스 응용프로그램 증가
 - 비정형 데이터(Unstructured data)로부터 인사이트를 도출 Needs 증가



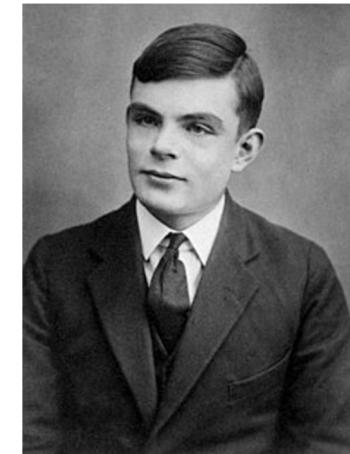
자연어 처리 현재 한계점

첫째, 도메인(산업)에 독립적인 범용 자연어 처리 솔루션이 없음

둘째, 자연어 처리 교육이 얼마나 오래 걸릴지, 결과가 얼마나 정확하며, 비즈니스 이점을 제공하기 위해 얼마나 정확해야 하는지를 예측하고 평가하기가 어려움

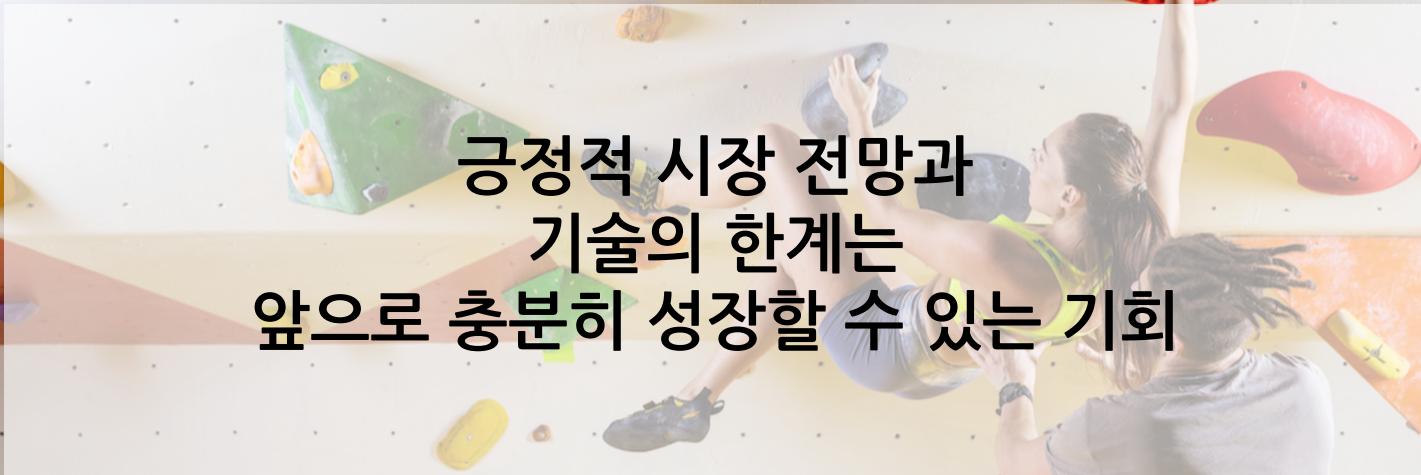
“인간이 컴퓨터와 대화하고 있다는 것을 깨닫지 못하고
인간과 대화를 계속할 수 있다면
컴퓨터는 지능적(Intelligence)인 것으로 간주될 수 있습니다.”

- 앤런 튜링 -



자연어 처리 전망

긍정적 시장 전망과
기술의 한계는
앞으로 충분히 성장할 수 있는 기회

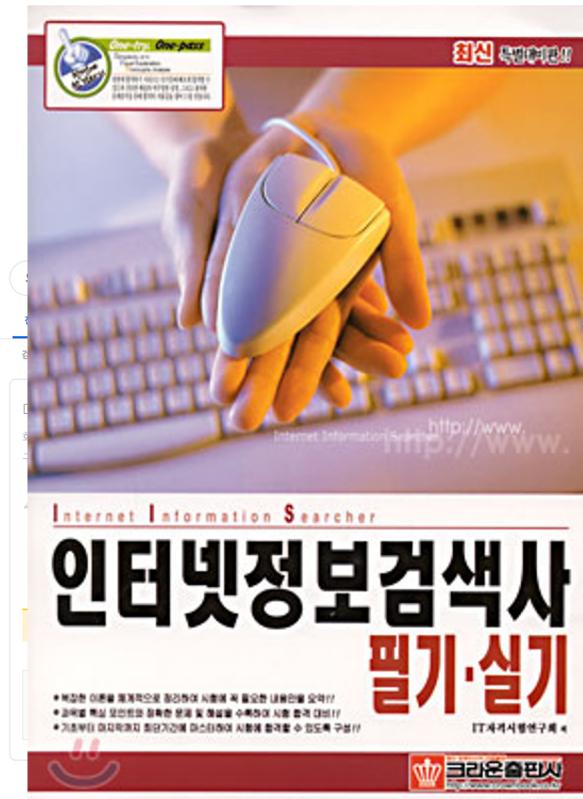
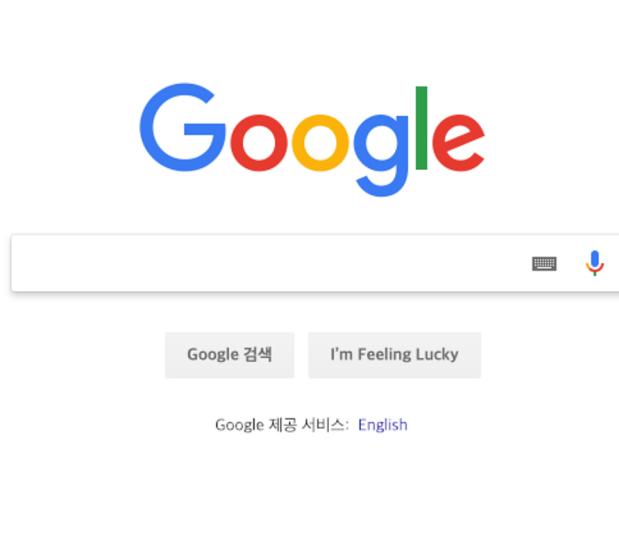


일상 속 자연어 처리

자연어 처리 시작하기

검색 엔진

- 과거 검색 엔진은 연산자(and, or 등)를 통한 검색이 가능
- 최근 검색 엔진은 검색창에 자연어 질의를 입력하면 적합한 답변을 제공

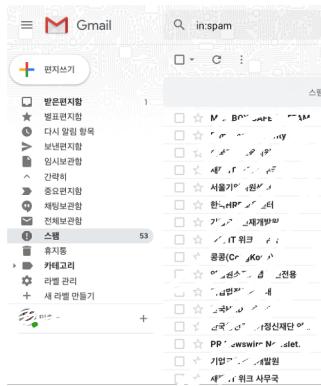


스팸 메일 분류

- 온라인 메일서비스로 메일이 보내지면 그 메일이 스팸이거나 아닐까?
- 설정을 하지 않으면?

나는 스팸설정을 한
적이 없는데..

어떻게 분류 했지?

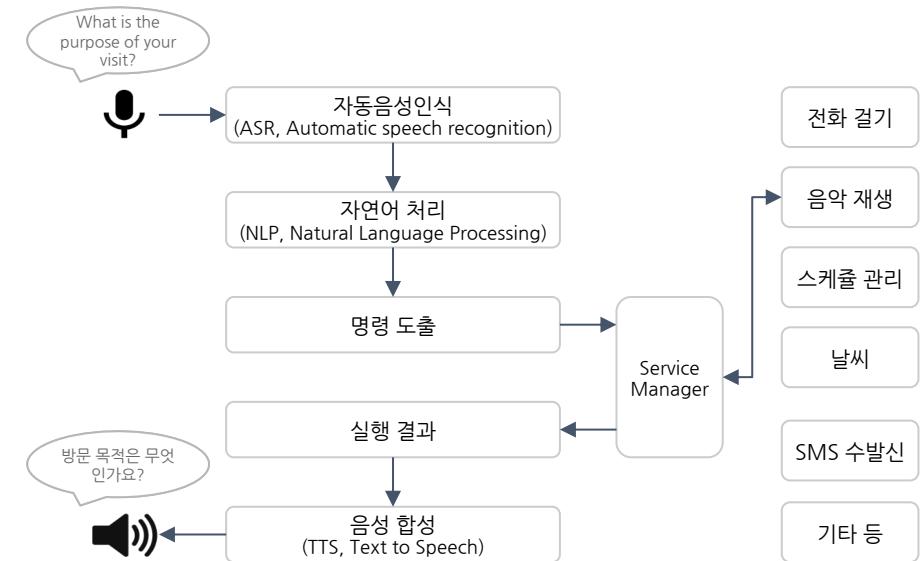


인공지능 비서

- 시리(Siri), 알렉사(Alexa), 구글 어시스턴스, 빅스비 등 음성기반의 인공지능 비서
- 음성으로 요청을 하면 문자로 변환하여 자연어 처리 엔진이 질의를 이해하여 처리하고 답변

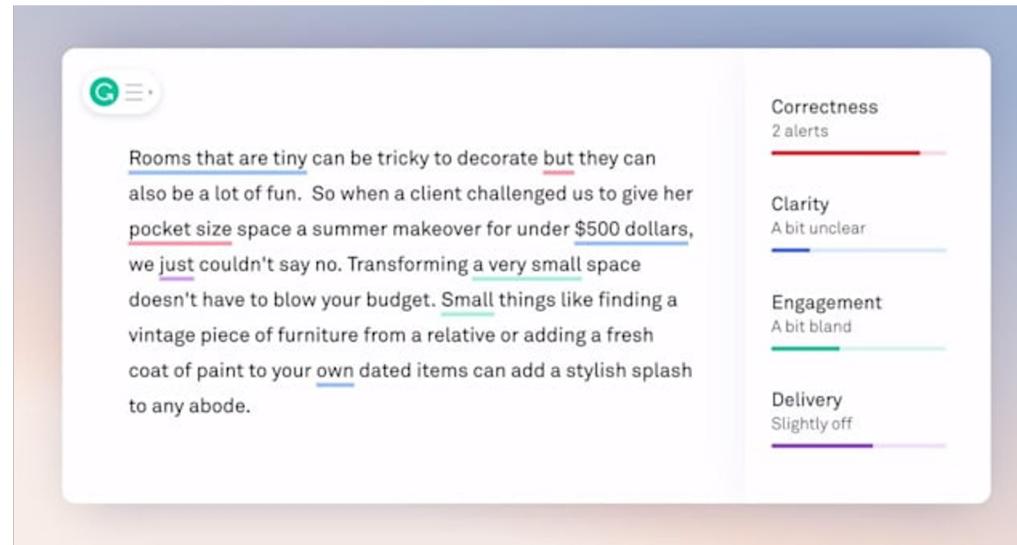
INPUT

Raymond Wong
HEY GOOGLE
Google Pixel Buds 2 review:
real-time translation is the killer app
They're very small, very comfy, and sound very good. But you should only get them if you love the Google Assistant and want to marry it.



문법 검사기

- 철자 검사
- 문법 검사기
- 단어 추천
- 뉴앙스 파악



SNS 내 인지도 분석

- 온라인 미디어(뉴스, 블로그, SNS, 리뷰 등) 데이터를 수집하여 베즈량 및 감성분석
- 분석하고자 하는 대상의 시장 반응(긍정, 부정, 중립) 여부를 판단하여 전략수립

