

비정형 데이터분석

유재명

비정형 데이터

- 정형(structured) - 주로 표 형태의 데이터
- 비정형(unstructured) - 자연어, 이미지 등 표 형태가 아닌 데이터

대부분의 데이터는 비정형

왜 비정형 데이터 분석인가?

- 최근 많은 컴퓨터 과학 논문은 정식 출판 전 arXiv에 게재
 - arXiv: 출판 전 논문을 게재하는 사이트
- 주 5일 하루 10편씩 읽으면 월 100편의 논문을 읽을 수 있음.
- arXiv에 월 게재되는 논문은 10,000여편. (100배!)
- 방대한 비정형 데이터를 분석하여 가치있는 정보를 추출할 수 있다면 많은 가치가 있음

텍스트 분석

단어 빈도 분석

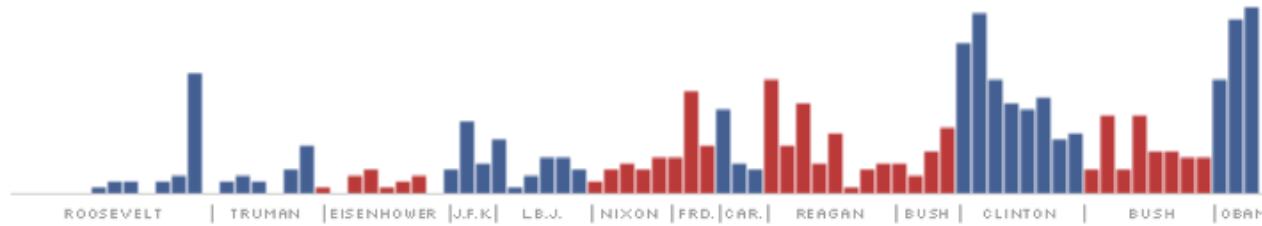
- 단어-문서 행렬을 바탕으로
- 단어 빈도의 총 합계를 구하거나
- 시간에 따른 빈도의 변화를 구하는 것

단어 빈도 분석

'jobs'

With unemployment above 9 percent, jobs were a focus of President Obama's speech. Historically, jobs get mentioned in the speech in rough correlation to the economic cycle, with spikes around 1975, 1981, 1991 and 2002.

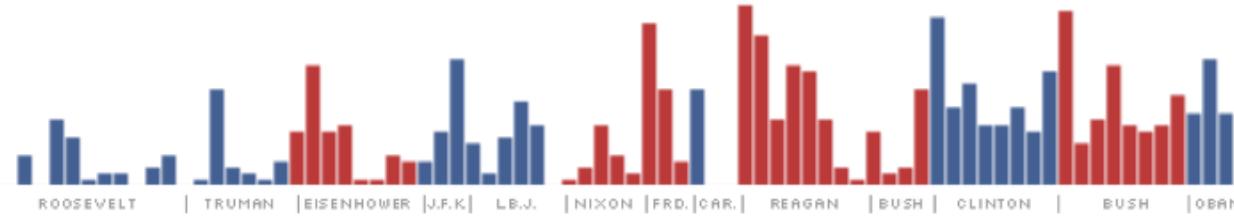
Words included: JOB, JOBS



'tax'

Presidents have used the word every year since 1981, when Mr. Reagan uttered it 30 times, detailing his plan to reduce taxes and government spending.

TAX, TAXED, TAXES, TAXING



'freedom'

President Dwight D. Eisenhower used this popular State of the Union word to describe the values he hoped other countries would adopt. Mr. Reagan spoke about "extending the frontiers of freedom" and Mr. Bush said the "advance of freedom is the great story of our time."

FREEDOM, FREEDOMS



출처: 뉴욕타임즈

의미망 분석

- 두 단어가 함께 나오는 관계를 나타내는 것
- 많이 나오는 단어는 크게, 적게 나오는 단어는 작게 원으로 그림
- 두 단어가 함께 나오는 경향이 있을 수록 굵은 선으로 그림
- 굵은 선으로 연결된 단어일 수록 가깝게 배치

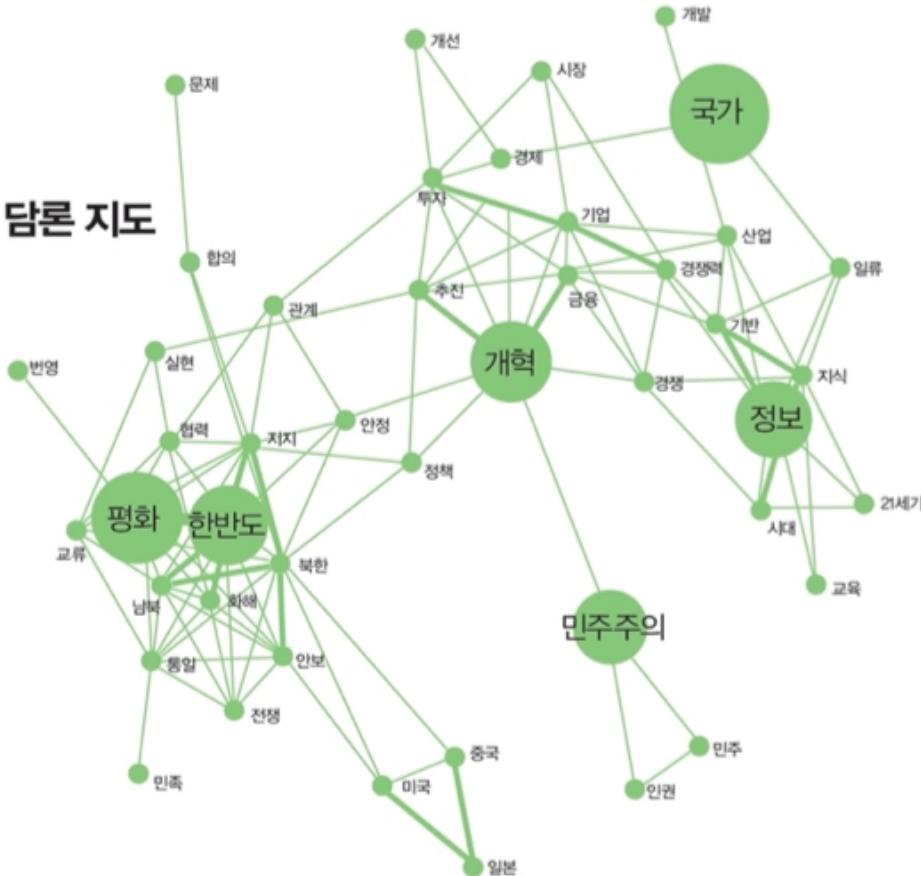
의미망 분석

대통령 연설문 속에 나타난 담론 지도

*원이 클수록 많이 언급했고, 연결선이 굵을수록 단어가 동시에 자주 쓰였다는 뜻이다.



국제통화기금(IMF) 외환위기에 대처하는 열쇳말은 '개혁'. 그는 민주주의와 시장경제라는 두 수레바퀴를 굽혀 '개혁'을 이뤄내고자 했다. 지식·정보사회를 만들어 일류 국가가 되는 게 목표였다. 통일·외교 영역에서는 화해와 협력, 남북 교류를 앞세워 한반도 평화 실현을 꿈꿨다.



의미망 분석



사람과의 대화 “권력을 통째로 내놓을 수도 있다”던 대통령은 언론과도 경쟁하며, 투명하고 공정한 정치와 시장경제를 만들고 싶어 했다. 한반도 평화, 미국과의 관계에서도 ‘원칙’을 내세우다 역풍을 맞기도 했다. 대화와 문제 해결을 좋아한 대통령 답게 담론 지도 속 단어가 가장 다채롭다.

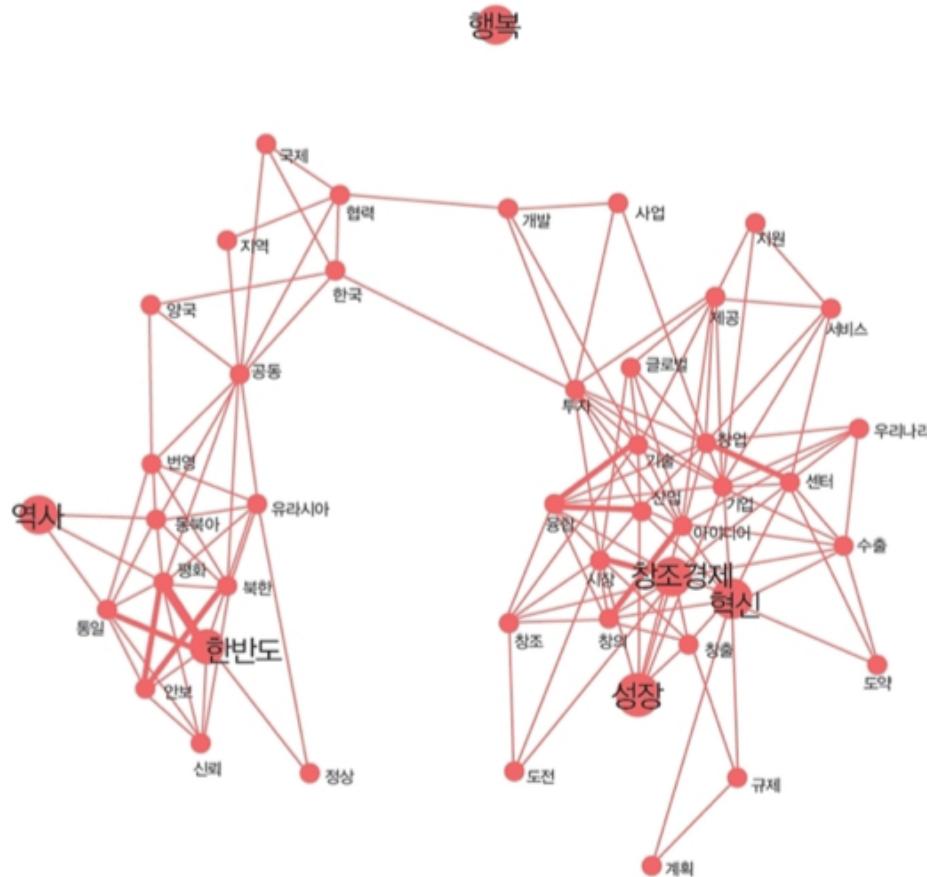


의미망 분석



위기는 나의 힘. 극복은 나의 삶. 샐러리맨에서 대통령으로 '성공 신화'를 쓴 대통령은 선진 일류 국가를 강조하며, 2008년 미국산 쇠고기 반대 촛불집회와 세계 금융위기 사태를 헤쳐나갔다. 친기업 대통령답게 '시장'이라는 건널목을 지나 미국·일본·중국으로, 금융정책과 국제회의로 넘어간다. 녹색과 기술, 투자의 목표는 결국 '성장'이다.

의미망 분석



경제와 안보. 2016년 1월13일 '대국민 담화문'에서 "국가를 지탱하는 두 축이 동시에 위기를 맞았다"고 하더니, 그동안의 연설문에서도 오로지 '두 축만 존재했다. '통일 대박'은 '테러방지법 제정'으로, '창조경제'는 '경제활성화법과 노동개혁법 국회 통과'로 바뀌었을 뿐. 대통령의 소원은 항상 '행복'이다.

감정 분석

- 텍스트에 나타난 감정을 긍정/부정으로 분석하는 것
- 감정 사전 또는 기계학습 사용
- **감정 사전:** 감정을 나타내는 단어 모음
- **기계 학습:** 텍스트의 감정을 사람이 판정 → 컴퓨터가 학습
 - 기계학습으로 감정사전을 만들 수도 있음

부모-자녀 관계 감정 사전

궁금하다 솔직하다
현신적 귀엽다 무관심 심하다
계시다 우리집 말하다 일방적
어렵다 나쁘다 화내다
짜증나다 즐겁다
미다 편하다 쌈우다 소중히
무섭다 안쓰럽다 힘들다 이성적
답답하다 보수적 멋지다 친하다
사랑스럽다 대단하다 잔소리 따뜻하다
슈퍼맨 언제나 아름답다 재미있다
너무나

곡성 감정 사전



곡성 감정 사전

★★★★★ 9 베스트 관람객 나감독.. 숨은 쉬게해줘야지...

seac**** | 2016.05.11 20:38 | 신고

공감 5506 비공감 1309

★★★★★ 10 베스트 관람객 어이없어서 웃기고 그냥 보고난후 뭐지...?이느낌

바보(dpql****) | 2016.05.11 20:12 | 신고

공감 4467 비공감 1385

★★★★★ 1 정신적으로 너무 안좋다

박시영(tldu****) | 2016.05.11 20:18 | 신고

공감 6989 비공감 3317

★★★★★ 1 관람객 ...활말없는 영화임.본사람은 공감할듯..무섭고 잔인하고 징그럽고 소름끼치는 반전
까지 고루 갖췄지만 그 모든걸 넘어서는 쩍찝함.

정연비(baby****) | 2016.05.11 22:05 | 신고

공감 7760 비공감 4451

곡성 감정 사전



긍정	부정
현혹 꿀잼 한국 완전 대박 소름 상영 오랜만	최악 쓰레기 별로 실망 진심 노잼 스트레스 평론가

감정 분석 + 주제 분류: BestBuy

Best Buy TV & Home Theater TVs 4K Ultra HD TVs

Share Print

Samsung - 55" Class (54.6" Diag.) - LED - 2160p - Smart - 4K Ultra HD TV - Black

Model: UN55KU6290FXZA SKU: 5217300 ★★★★★ 4.6 (2,456) 189 Questions, 442 Answers

Add to Cart

PRICE MATCH GUARANTEE See details in checkout Why?

OPEN-BOX Buying Options

FREE DELIVERY on TVs 51" and larger

Geek WE FIX IT OR REPLACE IT Normal Wear & Tear • Power Surges Plan Details

2-Year Standard Geek Squad Protection - \$69.99

Build A Bundle

Save for Later Add to Registry

Screen Size Class: ①

55" 65"

Delivery: Not Available for 96950 Change

Store Pickup: OLYMPIA WA Pick Up Today



감정 분석 + 주제 분류: BestBuy

Overall Customer Rating



4.6

(2459 Reviews)

96% of customers recommend this product.

[Write a Review](#)

Product Features Mentioned In Customer Reviews

Pros

[Picture Quality](#)

1326

[Price](#)

804

[Ease Of Use](#)

336

[Sound Quality](#)

328

[Smart Feature](#)

266

Cons

[Stand](#)

14

[Motion Blurring](#)

11

[Bluetooth](#)

7

[Refresh Rate](#)

7

[Compatibility](#)

6

감정 분석 + 주제 분류: BestBuy

rejela



4

Good product for the price

April 18, 2017

Always trust in Samsung products. Only negative is the flimsy *stand* - wish it was more balanced and secure. Picture and sound are great and was easy to connect it to our receiver for a big sound. Satisfied with the purchase.

Atvvstree



5

Great for gaming

July 30, 2016

I have games and watched movies on this tv. It is a fantastic tv. Especially for the price. The **only** thing is the *stand* isn't the best in the world.

Ameya



2

Basic TV

March 4, 2017

Bluetooth is disabled for USA region. *Stand* is not so sturdy. Screws do not fit Properly to hold tv firmly. Otherwise tv is ok. Please cross check tv specifications for USA region, few functionalities are disabled.

VIBE

16 topics
Sorted by date and time

1M

Topic	Impact
React Native	Positive Impact
Coffee Machine	Positive Impact
Firebase Test Lab	Negative Impact
Paired Programming	Negative Impact
API Integrations	Positive Impact
Hardware Server Costs	Positive Impact
Design Sprint	Positive Impact
Expense Report	Negative Impact
AWS EC2 Instance	Negative Impact
Download Office Applications	Negative Impact
Artificial Intelligence	Positive Impact

Scroll for more topics ▾

Positive Impact

Discussed in general

↑ Happiness
Detected Emotion

Irritation Disapproval Disappointment Stress

Team Discussing

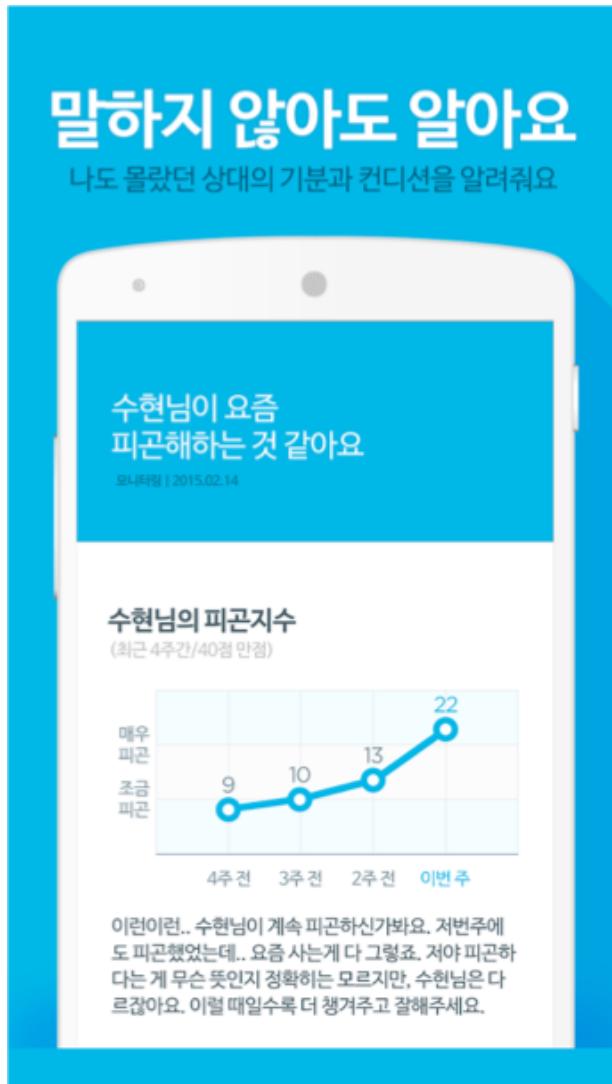
Colin Smithson Simon Ferris James Thompson Sarah Miller +12 others

Read on Slack

Read where this topic was discussed in your Slack chat.

Read on Slack

진저



다면평가 분석

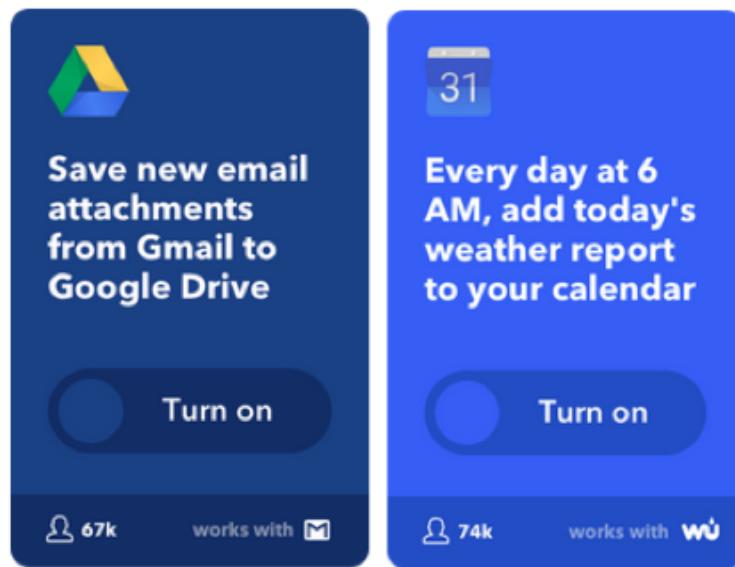
- 다면평가 시 주관식 서술 문항 → 점수 반영 안됨
- 감정 분석을 통해 점수화
- 직급/직무에 맞지 않는 서술 발견

의미 해석

- Semantic Parsing
- 자연어 문장을 논리식, 프로그램 코드 등으로 변환



- <http://ifttt.com>



의미 해석: ifttt

- 마이크로소프트의 연구

	INPUT	Park in garage when snow tomorrow	
(a)	IFTTT	Weather : Tomorrow's_forecast_calls_for	⇒ SMS : Send_me_an_SMS
	OUTPUT	Weather : Tomorrow's_forecast_calls_for	⇒ SMS : Send_me_an_SMS
	INPUT	Suas fotos do instagr.am salvas no dropbox	
(b)	IFTTT	Instagram : Any_new_photo_by_you	⇒ Dropbox : Add_file_from_URL
	OUTPUT	Instagram : Any_new_photo_by_you	⇒ Dropbox : Add_file_from_URL
	INPUT	Foursquare check-in archive	
(c)	IFTTT	Foursquare : Any_new_check-in	⇒ Evernote : Create_a_note
	OUTPUT	Foursquare : Any_new_check-in	⇒ Google_Drive : Add_row_to_spreadsheet
	INPUT	if i post something on blogger it will post it to wordpress	
(d)	IFTTT	Blogger : Any_new_post	⇒ WordPress : Create_a_post
	OUTPUT	Feed : New_feed_item	⇒ Blogger : Create_a_post
	INPUT	Endless loop!	
(e)	IFTTT	Gmail : New_email_in_inbox_from	⇒ Gmail : Send_an_email
	OUTPUT	SMS : Send_IFTTT_any_SMS	⇒ Philips_hue : Turn_on_color_loop

Table 4: Example output from the posclass system. For each input instance, we show the original query, the recipe originally authored through IFTTT, and our system output. Instance (a) demonstrates a case where the correct program is produced even though the input is rather tricky. Even the Portuguese query of (b) is correctly predicted, though keywords help here. In instance (c), the query is underspecified, and the system predicts that archiving should be done in Google Drive rather than evernote. Instance (d) shows how we sometimes confuse the trigger and action. Certain queries, such as (e), would require very deep inference: the IFTTT recipe sets up an endless email loop, where our system assembles a strange interpretation based on keyword match.

Channel	+Func
<i>(a) All: 4,294 recipes</i>	
retrieval	28.2
phrasal	17.3
sync	16.2
classifier	46.3
posclass	47.4
mturk	33.4
oracleturk	48.8
	37.8

빅데이터 인문학



- 빅데이터 인문학: 진격의 서막
- 에레즈 에이든, 장바티스트 미셸 공저
- 사계절
- 2015년
- 구글 n그램을 이용한 연구 사례
- 교양서로 읽어볼만

빅데이터 인문학



- 음식의 언어
- 댄 주라프스키
- 어크로스
- 2015년
- 음식과 관련된 텍스트 분석 사례
- 교양서로 읽어볼만

이미지 분석

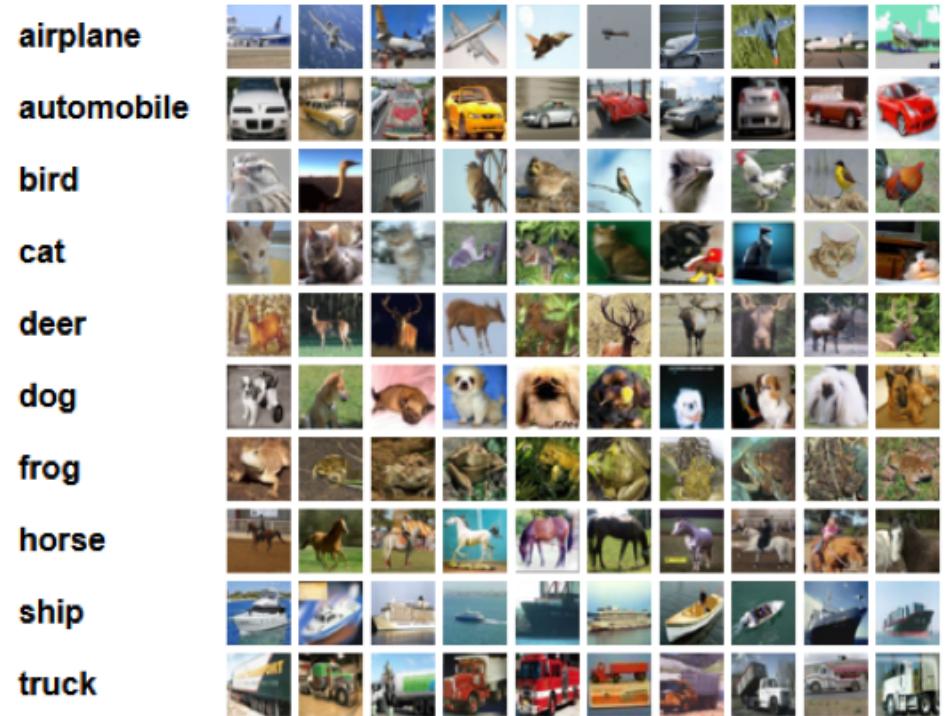
문자 인식

- 손글씨 인식(손글씨 이미지 → 글자)
- MNIST 숫자 손글씨 데이터
- 딥러닝 오류율: 0.21% (인간 수준)

1 1 5 4 3
7 5 3 5 3
5 5 9 0 6
3 5 2 0 0

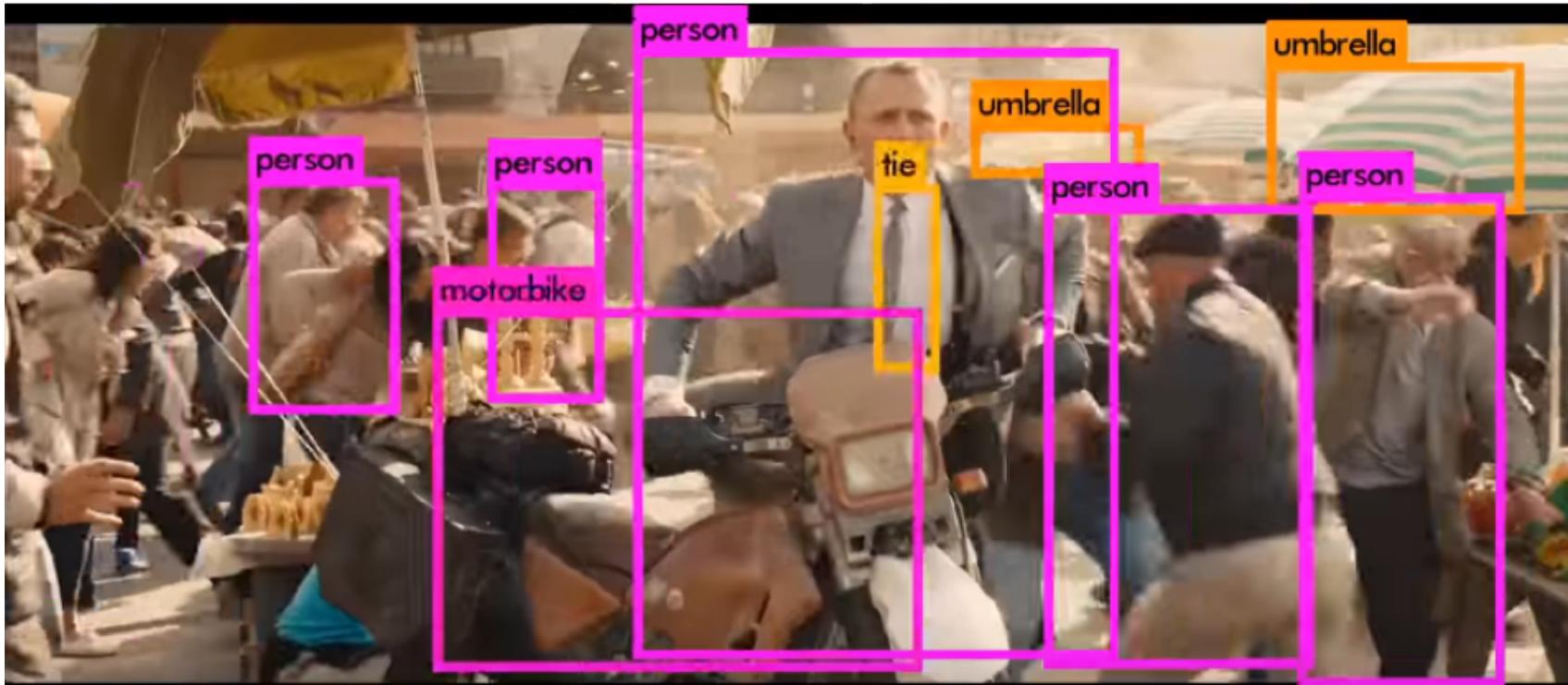
물체 인식

- 물체 인식(물체 이미지 → 물체 종류)
- CIFAR10 물체 이미지 데이터
- 딥러닝 오류율: 3.47%



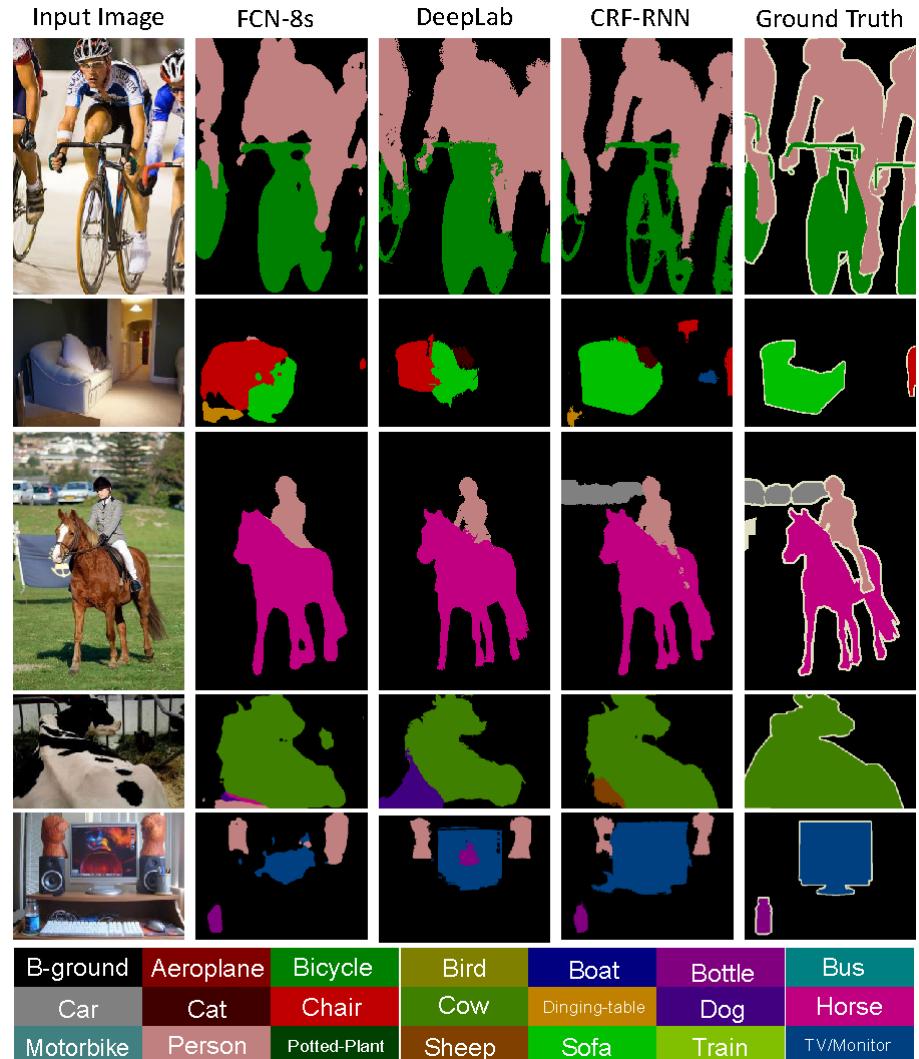
물체 탐지

- <https://www.youtube.com/watch?v=VOC3huqHrss>



시맨틱 세그멘테이션

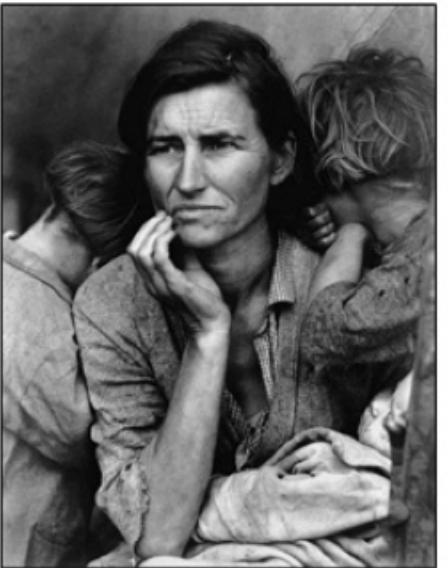
- semantic segmantation



Colorization



Colorization



SNS

- 소셜 미디어에 올라온 사진에서 브랜드 인식



@emilymily
Las Vegas, NV

Today 1:32 pm
3.2k followers 290 following 159 posts

BRANDS
50% Adidas 100%

CONTEXT
23% T-shirt 100% Shoes 100%
Sport 57%

HUMANS
99% Female 99% Brunette 87%

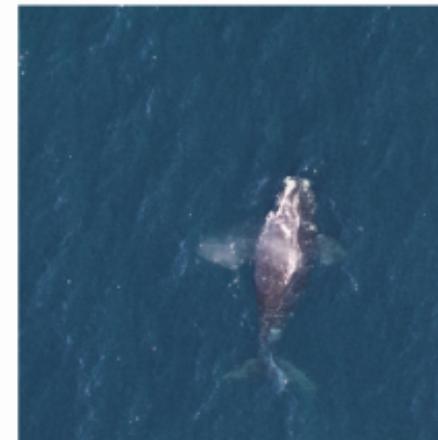
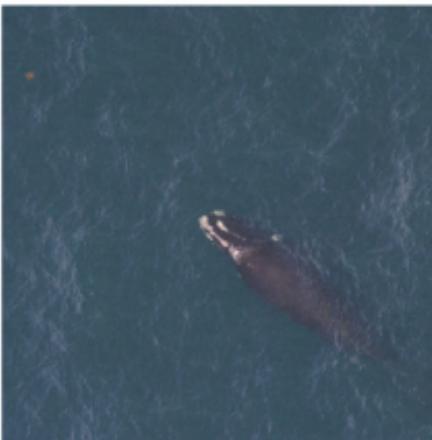
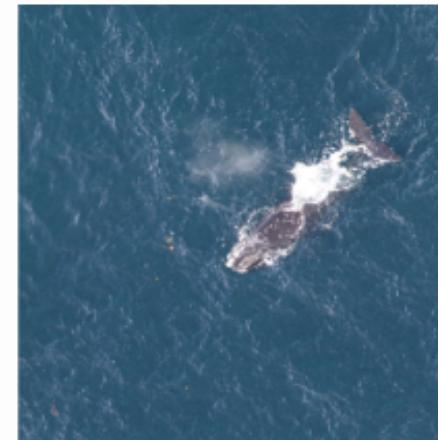
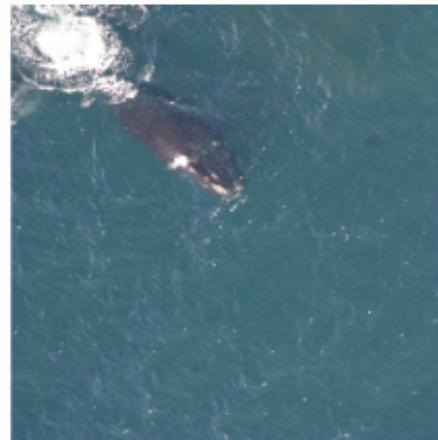
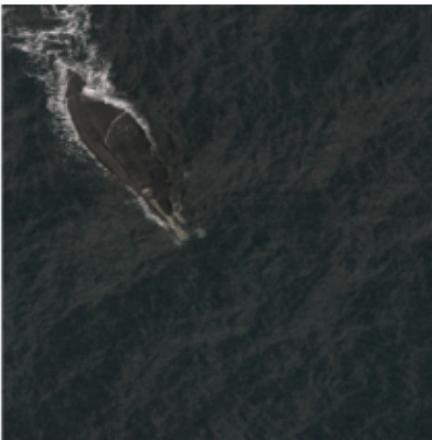
USER TAGS
#girl #adidas #lovesport #selfie #workout #iphone
#runtheworld @fitarea

위성 사진 segmentation

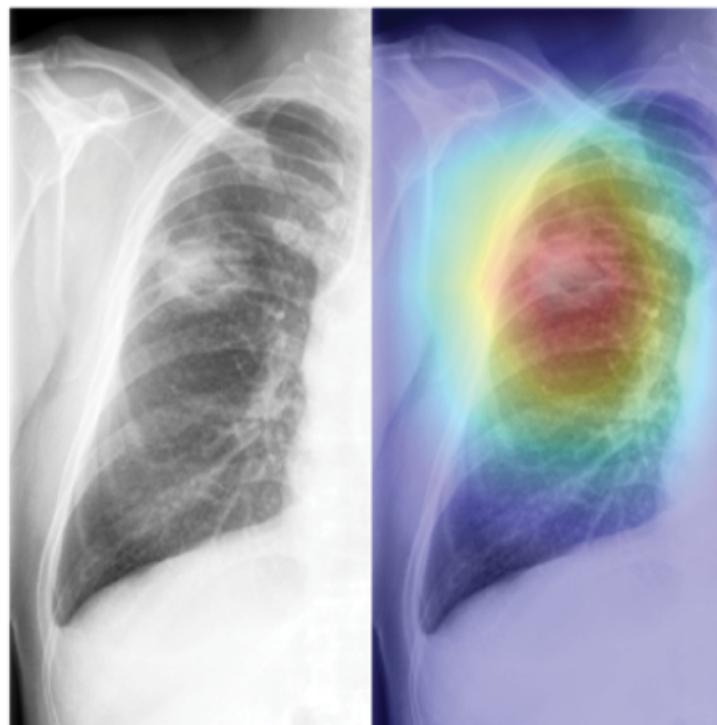
- 위성 사진의 물체를 구별



멸종 위기종 탐지



헬스케어



ORIGINAL CR

DIB APPLIED

패션

옷을 사진으로 찍으면 쇼핑몰 구매 링크를 제시

