겨울방학 데이터사이언스 특강

빅 데이터 분석 기초

충남대학교 경영학부

이재환

1

BIG DATA

2

빅 데이터

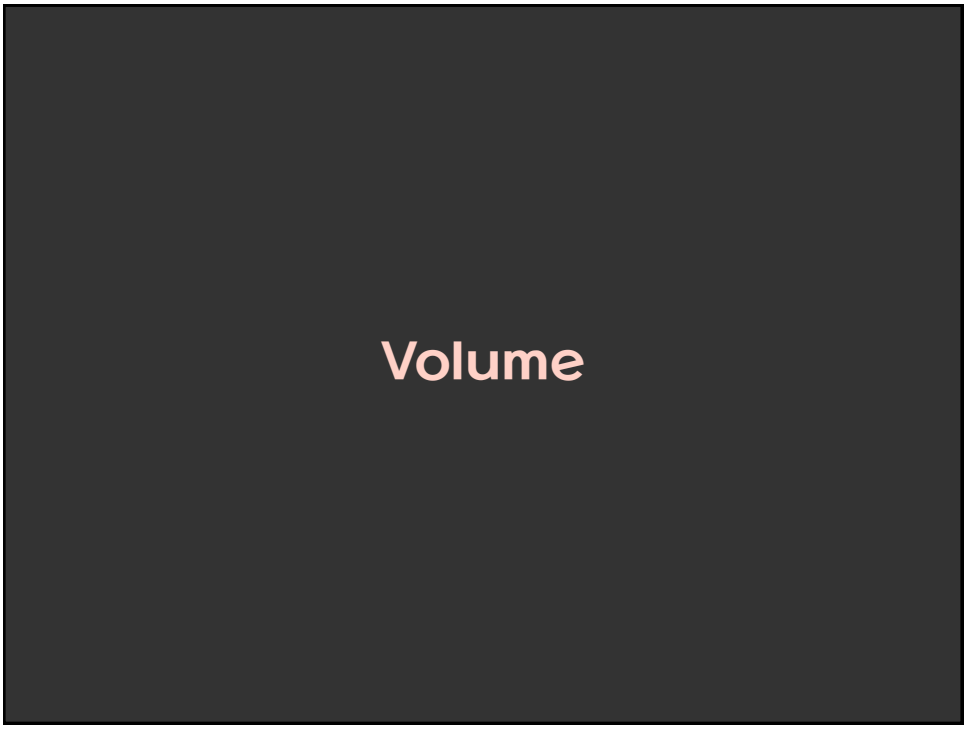
- 기존 데이터베이스 관리 도구로는 수집, 저장, 관리, 분석하기 어려울 정도로 방대하고, 빠르며, 다양한 정형·비정형 데이터
- 핵심 특성
  - Volume(규모): 테라바이트에서 제타바이트 수준의 대용량 데이터
  - Velocity(속도): 실시간 또는 빠른 속도로 생성 및 유통되는 데이터
  - Variety(다양성): 정형(DB), 반정형(XML), 비정형 데이터

3

빅 데이터

- 기존 데이터베이스 관리 도구로는 수집, 저장, 관리, 분석하기 어려울 정도로 방대하고, 빠르며, 다양한 정형·비정형 데이터
- 핵심 특성
  - **Volume**(규모): 테라바이트에서 제타바이트 수준의 대용량 데이터
  - Velocity(속도): 실시간 또는 빠른 속도로 생성 및 유통되는 데이터
  - Variety(다양성): 정형(DB), 반정형(XML), 비정형 데이터

4



5

빅 데이터 – Volume

- Q. 얼마나 많으면 충분히 "빅" 데이터라고 부를 수 있을까?

6

p-Value Problem

Information Systems Research
Articles in Advance, pp. 1–12
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)

http://dx.doi.org/10.1287/isre.2013.0480
© 2013 INFORMS

Too Big to Fail: Large Samples and the *p*-Value Problem

Mingfeng Lin
Eller College of Management, University of Arizona, Tucson, Arizona 85721, mingfeng@eller.arizona.edu
Henry C. Lucas, Jr.
Robert Smith School of Business, University of Maryland, College Park, Maryland 20742, hlucas@rhsmith.umd.edu
Galit Shmueli
Srini Raju Centre for IT & the Networked Economy, Indian School of Business, Hyderabad 500 032, India, galit.shmueli@isb.edu

The Internet has provided IS researchers with the opportunity to conduct studies with extremely large samples, frequently well over 10,000 observations. There are many advantages to large samples, but researchers using statistical inference must be aware of the *p*-value problem associated with them. In very large samples, *p*-values go quickly to zero, and solely relying on *p*-values can lead the researcher to claim support for results of no practical significance. In a survey of large sample IS research, we found that a significant number of papers rely on a low *p*-value and the sign of a regression coefficient alone to support their hypotheses. This research commentary recommends a series of actions the researcher can take to mitigate the *p*-value problem in large samples and illustrates them with an example of over 300,000 camera sales on eBay. We believe that addressing the *p*-value problem will increase the credibility of large sample IS research as well as provide more insights for readers.

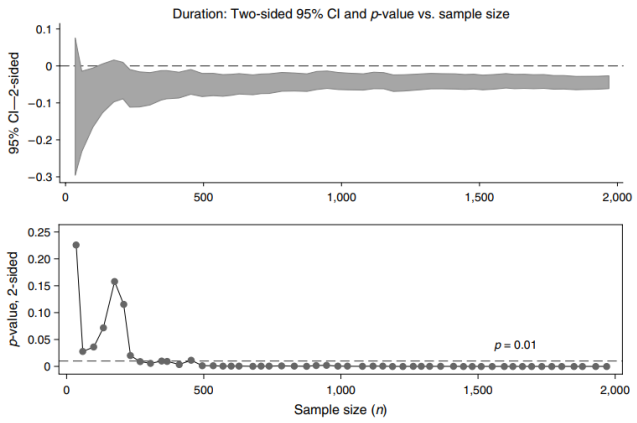
Key words: empirical modeling; practical significance; effect size; *p*-value; statistical significance; inference
History: Alok Gupta, Senior Editor. This paper was received on August 15, 2012, and was with the authors 2 weeks for 1 revision. Published online in *Articles in Advance*.

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information systems research*, *24*(4), 906-917.

7

p-Value Problem

Figure 1 **Two-Sided 95% Confidence Interval (Top) and *p*-Value (Bottom) for Duration vs. Sample Size**



Notes. Zoomed in to $n < 2{,}000$ for illustration. Horizontal dashed line in lower panel: $p = 0.01$.

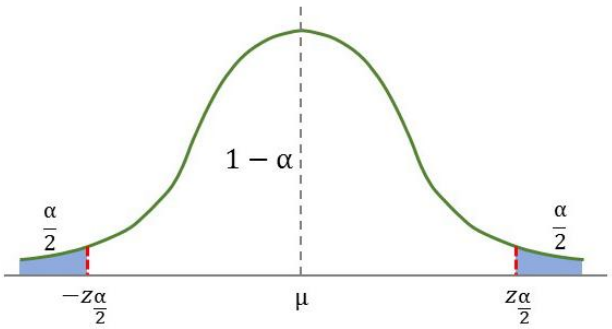
Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information systems research*, *24*(4), 906-917.

8

가설 검정: 1표본 t-검정

- 검정 통계량

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$



9

9

표본과 모집단

- 모집단(Population)
  - 통계전문가가 관심을 가지고 있는 모든 항목들의 그룹(집합)이다.
  - 일반적으로 매우 크며 종종 무한할 수 있다.
    - 예. 예제 12.5에서 플로리다 주의 5백만 명 유권자

- 표본(Sample)
  - 모집단으로부터 추출된 데이터 집합이다.
  - 매우 클 수도 있지만 모집단보다는 작다.
  - 예. 선거일에 출구여론조사가 이루어진 765명의 유권자로 구성된 표본

10

10

기술통계학과 추론통계학

- 기술통계학(Descriptive Statistics)
  - 데이터를 편리하고 정보를 나타내는 방식으로 정리, 요약, 설명하는 방법
  - 데이터의 중심 위치: 평균, 중앙값(median)
  - 데이터의 변동성: 범위, 분산, 표준편차

- 추론통계학(Inferential Statistics)
  - 표본데이터에 기초하여 모집단의 특성에 관한 결론을 얻거나 추론을 하기 위해 사용되는 통계방법론
  - ※. 모집단의 부분집합을 선택하는 표본추출(sampling)은 비용(cost)과 실용성 (practicality) 등의 이유 때문
    - 1억 명의 TV 시청자 인터뷰하기 vs. 1,000명의 시청자 인터뷰 하기
    - 생산된 모든 자동차에 대하여 충돌 실험

11

Variety (1) – Text

12

빅 데이터

- 기존 데이터베이스 관리 도구로는 수집, 저장, 관리, 분석하기 어려울 정도로 방대하고, 빠르며, 다양한 정형·비정형 데이터
- 핵심 특성
  - Volume(규모): 테라바이트에서 제타바이트 수준의 대용량 데이터
  - Velocity(속도): 실시간 또는 빠른 속도로 생성 및 유통되는 데이터
  - **Variety**(다양성): 정형(DB), 반정형(XML), 비정형 데이터

13

빅 데이터 – Variety

- 비정형 데이터
  - 일정한 구조나 형식 없이 저장된 데이터
  - 고정된 형식이나 구조(스키마)가 없는 **텍스트**, 영상, 음성, **이미지**, SNS 게시물 등의 원시 데이터
- 생각해봅시다.
  - 텍스트와 이미지를 어떻게 연산가능한 형태로 표현할 것인가?
  - 숫자로 표현했다면 우리가 사용하던 방법론을 그대로 사용해도 되는가?

14

텍스트 분석의 개념

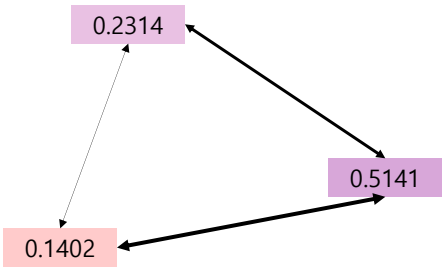
- 대부분의 분석 문제는 다음 두 질문으로 설명 가능
  - 문서를 어떻게 표현할 것인가?
  - 표현된 문서를 어떻게 평가(비교)할 것인가?

Document 1	Document 2	Document 3
The Moon is a barren, rocky world without air and water. It has dark lava plain on its surface. The Moon is filled wit craters. It has no light of its own. It gets its light from the Sun. The Moo keeps changing its shape as it moves round the Earth. It spins on its axis in 27.3 days stars were named after the Edwin Aldrin were the first ones to set their foot on the Moon on 21 July 1969 They reached the Moon in their space craft named Apollo II.	The sun is a huge ball of gases. It has a diameter of 1,392,000 km. It is so huge that it can hold millions of planets inside it. The Sun is mainly made up of hydrogen and helium gas. The surface of the Sun is known as the photosphere. The photosphere is surrounded by a thin layer of gas known as the chromospheres. Without the Sun, there would be no life on Earth. There would be no plants, no animals and no human beings. As, all the living things on Earth get their energy from the Sun for their survival.	The Solar System consists of the Sun Moon and Planets. It also consists of comets, meteoroids and asteroids. The Sun is the largest member of the Solar System. In order of distance from the Sun, the planets are Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune and Pluto; the dwarf planet. The Sun is at the centre of the Solar System and the planets, asteroids, comets and meteoroids revolve around it.

15

텍스트 분석의 개념

- 대부분의 분석 문제는 다음 두 질문으로 설명 가능
  - 문서를 어떻게 표현할 것인가?
  - 표현된 문서를 어떻게 평가(비교)할 것인가?



16

텍스트를 표현하기

- 문서를 어떻게 연산이 가능한 형태로 표현할 것인가?
  - = 어떻게 임베딩(embedding)할 것인가?

> 충남대학교는 창의·개발·봉사의 교육이념을 기반으로 우수한 졸업생을 배출해온 중부권 최고의 명문 대학입니다. 특히, 충남대학교 경상대학은 글로벌 역량과 윤리의식을 갖춘 전문 인력 양성이라는 미션을 설정하고 국내를 넘어 아시아 상위권 대학으로 도약하기 위해 노력하고 있습니다.

17

텍스트를 표현하기 – BoW

- Bag-of-Words
  - 텍스트를 단어로 쪼갠 후, 단어들의 (순서가 없는) 집합으로 보자!

충남대학교는 창의·개발·봉사의 교육이념을 기반으로 우수한 졸업생을 배출해온 중부권 최고의 명문 대학입니다. 특히, 충남대학교 경상대학은 글로벌 역량과 윤리의식을 갖춘 전문 인력 양성이라는 미션을 설정하고 국내를 넘어 아시아 상위권 대학으로 도약하기 위해 노력하고 있습니다.

→ {충남대학교는, 창의, 개발, 봉사, 교육이념을, 기반으로, 우수한, 졸업생을, 배출해온, 중부권, 최고의, 명문, 대학입니다, 특히, 충남대학교, 경상대학은, 글로벌, 역량과, 윤리의식을, 갖춘, 전문, 인력, 양성이라는, 미션을, 설정하고, 국내를, 넘어, 아시아, 상위권, 대학으로, 도약하기, 위해, 노력하고, 있습니다}

18

형태소 분석의 필요성

19

19

형태소 분석

- 형태소(parts of speech)
  - 뜻을 갖는 최소 언어 단위
  - 단어 종류(word classes) 또는 어휘 범주(lexical categories)로 품사(品詞)라고도 한다.
- 형태소 분석이란?
  - 형태소를 비롯하여 어근, 접두사/접미사 등 문장의 다양한 언어적 속성의 구조를 분석하는 것을 말한다.

20

20

범용 POS 태그 세트

- Universal POS tagset(범용 POS 태그 세트)
  - 여러 나라의 언어에 대해 언어 간 일관성 있는 트리 뱅크(treebank)를 개발하는 프로젝트인 Universal dependencies에서 사용하는 표준 품사(POS) 태그다.
  - https://en.wikipedia.org/wiki/Treebank
  - https://github.com/slavpetrov/universal-pos-tags

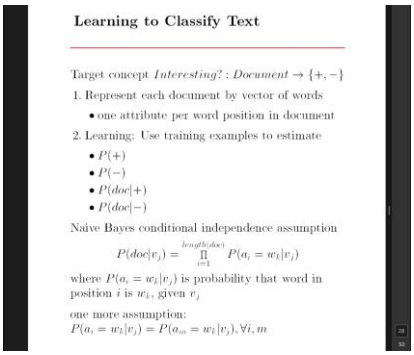
The files in this repository contain mappings from treebank specific tagsets to a set of 12 universal part-of-speech tags. The 12 universal tags are:

VERB - verbs (all tenses and modes)
NOUN - nouns (common and proper)
PRON - pronouns
ADJ - adjectives
ADV - adverbs
ADP - adpositions (prepositions and postpositions)
CONJ - conjunctions
DET - determiners
NUM - cardinal numbers
PRT - particles or other function words
X - other: foreign words, typos, abbreviations
. - punctuation

텍스트 분류 – Naive Bayesian

- Naive Bayesian Classifier
  - 2가지 가정
    - 어떤 단어가 등장할 확률은 다른 단어의 위치와 독립적이다.
    - 어떤 단어가 특정 위치에 등장할 확률은 일정하다.

To summarize, the naive Bayes classification v_{NB} is the classification that maximizes the probability of observing the words that were actually found in the document, subject to the usual naive Bayes independence assumption. The independence assumption $P(a_1, \ldots a_{111}|v_j) = \prod_1^{111} P(a_i|v_j)$ states in this setting that the word probabilities for one text position are independent of the words that occur in other positions, given the document classification v_j. Note this assumption is clearly incorrect. For example, the probability of observing the word "learning" in some position may be greater if the preceding word is "machine." Despite the obvious inaccuracy of this independence assumption, we have little choice but to make it—without it, the number of probability terms that must be computed is prohibitive. Fortunately, in practice the naive Bayes learner performs remarkably well in many text classification problems despite the incorrectness of this independence assumption. Domingos and Pazzani (1996) provide an interesting analysis of this fortunate phenomenon.



텍스트를 표현하기 – BoW

• 텍스트를 단어로 쪼갠 후, 단어들의 (순서가 없는) 집합으로 보자!

"충남대학교는 창의·개발·봉사의 교육이념을 기반으로 우수한 졸업생을 배출해온 중부권 최고의 명문 대학입니다. 특히, 충남대학교 경상대학은 글로벌 역량과 윤리의식을 갖춘 전문 인력 양성이라는 미션을 설정하고 국내를 넘어 아시아 상위권 대학으로 도약하기 위해 노력하고 있습니다."

→ {충남대학교, 는, 창의, 개발, 봉사, 교육이념, 을, 기반, 으로, 우수한, 졸업생, 을, 배출해온, 중부권, 최고, 의, 명문, 대학, 입니다, 특히, 충남대학교, 경상대학, 은, 글로벌, 역량, 과, 윤리의식, 을, 갖춘, 전문, 인력, 양성, 이라는, 미션, 을, 설정, 하고, 국내, 를, 넘어, 아시아, 상위권, 대학, 으로, 도약, 하기, 위해, 노력, 하고, 있습니다}

→ {충남대학교:2, 는:1, 창의:1, 개발:1, 봉사:1, 교육이념:1, 을:4, 기반:1, 으로:2, 우수한:1, 졸업생:1, 배출해온:1, 중부권:1, 최고:1, 의:1, 명문:1, 대학:2, 입니다:1, 특히:1, 경상대학:1, 은:1, 글로벌:1, 역량:1, 과:1, 윤리의식:1, 갖춘:1, 전문:1, 인력:1, 양성:1, 이라는:1, 미션:1, 설정:1, 하고:2, 국내:1, 를:1, 넘어:1, 아시아:1, 상위권:1, 도약:1, 하기:1, 위해:1, 노력:1, 있습니다:1}

23

Word Cloud

• (or Tag Cloud)

• Word Cloud is a visual representation of text data which often used to depict keyword metadata on websites, or to visualize free from text

source: https://en.wikipedia.org/wiki/Tag_cloud

24

텍스트를 표현하기 – BoW

- 텍스트를 단어로 쪼갠 후, 단어들의 (순서가 없는) 집합으로 보자!

  → {충남대학교:2, 는:1, 창의:1, 개발:1, 봉사:1, 교육이념:1, 을:4, 기반:1, 으로:2, 우수한:1, 졸업생:1, 배출해온:1, 중부권:1, 최고:1, 의:1, 명문:1, 대학:2, 입니다:1, 특히:1, 경상대학:1, 은:1, 글로벌:1, 역량:1, 과:1, 윤리의식:1, 갖춘:1, 전문:1, 인력:1, 양성:1, 이라는:1, 미션:1, 설정:1, 하고:2, 국내:1, 를:1, 넘어:1, 아시아:1, 상위권:1, 도약:1, 하기:1, 위해:1, 노력:1, 있습니다:1}

- 의문점
  - 단어가 배열된 순서도 중요하지 않나?
  - '으로', '하고'와 같은 조사는 자주 나올 수 밖에 없는데, 자주 나온다고 중요하다고 해도 괜찮은가?

N-gram

- A contiguous sequence of N items from a given sequence of text
- Can be thought as a window placed over a text, so that we only look at N words at a time

- e.g. "I am happy to join with you."
  - Unigram (1-gram)
    - {I, am, happy, to, join, with, you}
  - Bigram (2-gram)
    - {(I am), (am happy), (happy to), (to join), (join with), (with you)}

TF-IDF

- TF-IDF는 여러 개의 문서로 이루어진 문서집단이 있을 때, 특정 단어의 중요도를 계산할 때 사용
- 단어빈도(Term Frequency; TF)
  - 특정 문서에서 특정 단어가 등장한 빈도
- 문서빈도(Document Frequency; DF)
  - 전체 문서 중 특정 단어 t가 등장한 문서의 빈도

$$TF\text{-}IDF_{d,t} = TF_{d,t} \times IDF_t = TF_{d,t} \times \log(\frac{n}{1 + DF(t)})$$

TF-IDF

- 예.
  - 3개의 문서
    - 문서 1: AI is new
    - 문서 2: AI is great
    - 문서 3: Great AI is new

텍스트 분류 – Naive Bayesian

- Naive Bayesian Classifier
  - 2가지 가정
    - 어떤 단어가 등장할 확률은 다른 단어의 위치와 독립적이다.
    - 어떤 단어가 특정 위치에 등장할 확률은 일정하다.

To summarize, the naive Bayes classification v_{NB} is the classification that maximizes the probability of observing the words that were actually found in the document, subject to the usual naive Bayes independence assumption. The independence assumption $P(a_1, \ldots a_{111}|v_j) = \prod_1^{111} P(a_i|v_j)$ states in this setting that the word probabilities for one text position are independent of the words that occur in other positions, given the document classification v_j. Note this assumption is clearly incorrect. For example, the probability of observing the word "learning" in some position may be greater if the preceding word is "machine." Despite the obvious inaccuracy of this independence assumption, we have little choice but to make it—without it, the number of probability terms that must be computed is prohibitive. Fortunately, in practice the naive Bayes learner performs remarkably well in many text classification problems despite the incorrectness of this independence assumption. Domingos and Pazzani (1996) provide an interesting analysis of this fortunate phenomenon.

Learning to Classify Text

Target concept *Interesting?* : *Document* $\rightarrow \{+, -\}$

1. Represent each document by vector of words
   - one attribute per word position in document
2. Learning: Use training examples to estimate
   - $P(+)$
   - $P(-)$
   - $P(doc|+)$
   - $P(doc|-)$

Naive Bayes conditional independence assumption

$$P(doc|v_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k|v_j)$$

where $P(a_i = w_k|v_j)$ is probability that word in position i is w_k, given v_j

one more assumption:
$P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$

29

29

텍스트 표현 – Word2Vec

- Word2Vec
  - https://word2vec.kr/search/

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

30

30

텍스트 표현 – ChatGPT

- OpenAI API – Embeddings
  - https://platform.openai.com/docs/guides/embeddings?lang=python
  - small 모델의 경우 1,536차원의 벡터 생성
  - large 모델의 경우 3,072차원의 벡터 생성

Embedding models

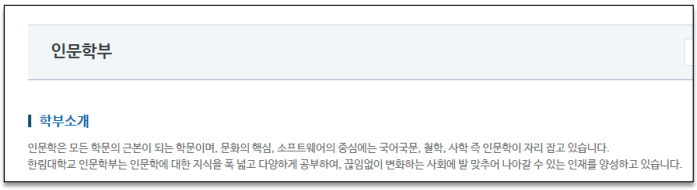
OpenAI offers two powerful third-generation embedding model (denoted by `-3` in the model ID). You can read the embedding v3 announcement blog post for more details.

Usage is priced per input token, below is an example of pricing pages of text per US dollar (assuming ~800 tokens per page):

MODEL	~ PAGES PER DOLLAR	PERFORMANCE ON MTEB EVAL	MAX INPUT
text-embedding-3-small	62,500	62.3%	8191
text-embedding-3-large	9,615	64.6%	8191
text-embedding-ada-002	12,500	61.0%	8191

텍스트 표현 – ChatGPT

- 인문학부



- 임베딩 결과 (길이 = 1,536)

```
[-0.022846616804599762, 0.014188679866492748, 0.00049104721983894 71, 0.020899659022688866, 0.07801619917154312, -0.039214856922626495,
0.002145100850611925, 0.02968681789934635, -0.030565553564965725, 0.03413208946585655, -0.04531418159604072 6, 0.022777698934078217, -0.015127699822187424,
0.004184669815003872, 0.032098978757858276, 0.007817988283932209, -0.04572769254446603, 0.040558777749538 42, 0.005733191501349211, 0.018625333905220032,
0.050276342779397964, -0.009519424289464 95, -0.015377530828118324, 0.005044002551585436, -0.010131078772246838, -0.015368916094303131, -
0.011207937262952328, -0.013809625990688 8, -0.025603374466330001, 0.050896611064672 47, 0.031702697277069 09, 0.012353713624179363, -0.019228374585 5093,
0.013826855458319187, -0.023122293874621 39, 0.042591884732246 4, 0.024552360177040 1, 0.014989862218499184, 0.003094889689236 8793, -0.005453208461403847,
0.004391426686197519, 0.001744509791024 0293, 0.014412666670978 07, -0.029772967100143433, -0.021743914112448692, 0.03232296555623436, .022054050117731094,
0.014576348476111889, 0.049208097159862 52, 0.066024310886859 9, … -0.024138847365975 38, 0.032805398106575 01, -0.043418910354375 84, -0.037181749939918 52,
0.017014354467391968, -0.018211821095155 31, 0.018694253638386726, -0.019952023401856422, -0.023329049348831177, -0.014533273875713348]
```

Variety (2) – ANN

33

빅 데이터 분석 방법

- 데이터를 우리가 분석할 수 있는 형태로 표현한 후에, 어떤 분석 방법을 사용할 것인가
- 예. gpa.csv

	A	B	C	D
1	평균공부시간(1주일)	평균수면시간(1일)	학점(4.5만점)	
2	21	7	3.8	
3	5	5	2.75	
4	5	8	2.39	
5	5	6	2.8	
6	24	7	4	
7	3	9	2	
8	7	7	3	
9	3	7.5	2.7	
10	26	6	3.35	
11	6	7	2.8	
12	12	6	3.5	
13	2.5	5	2.75	
14	3.5	6	2.78	

34

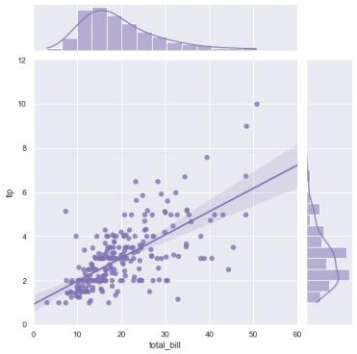
34

선형 회귀 분석

• 선형 회귀는 주어진 데이터 집합에 대해, 종속 변수와 독립 변수 사이의 선형 관계를 모델링한다. 모델은 다음과 같은 형태를 갖는다.

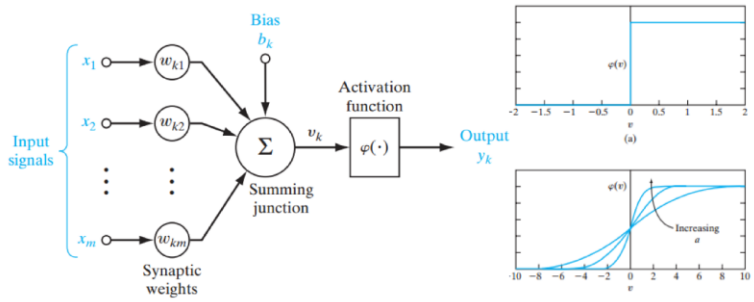
$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n$$

- 회귀계수의 계산은 잔차의 제곱합이 최소가 되는 최소제곱법을 많이 사용
- 예.
  - 종속 변수: 웨이터에게 지불한 팁
  - 독립 변수: 전체 음식 가격



인공신경망

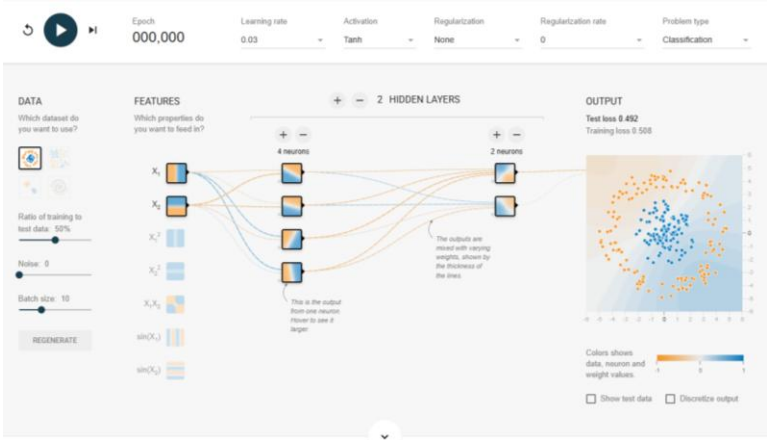
• 신경세포(neuron)을 모방한 구조에서 시작



$$y_k = \varphi\left(b_k + \sum_m x_k w_k\right)$$

인공신경망 체험

- A Neural Network Playground
  - https://playground.tensorflow.org/



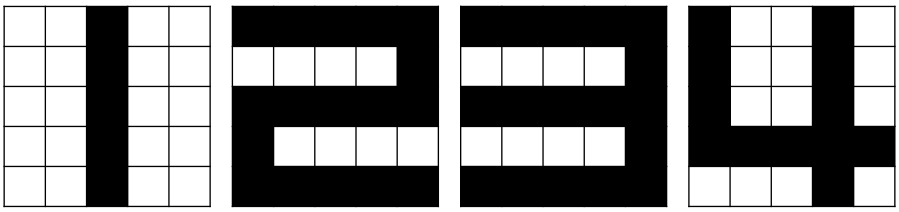
37

37

Variety (3) - Image

38

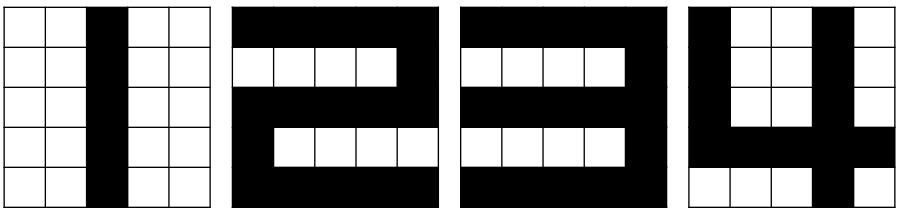
이미지 데이터



39

39

이미지 데이터



0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0

1	1	1	1	1
0	0	0	0	1
1	1	1	1	1
1	0	0	0	0
1	1	1	1	1

1	1	1	1	1
0	0	0	0	1
1	1	1	1	1
0	0	0	0	1
1	1	1	1	1

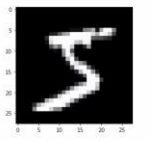
1	0	0	1	0
1	0	0	1	0
1	0	0	1	0
1	1	1	1	1
0	0	0	1	0

40

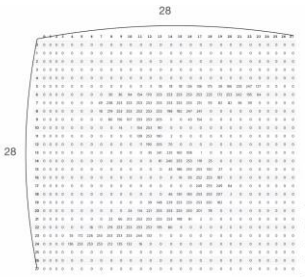
40

이미지 데이터

흑백



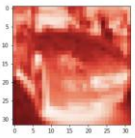
(28, 28)

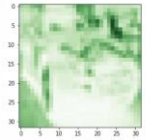


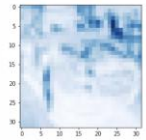
784개의 숫자

컬러

(32, 32, 3)



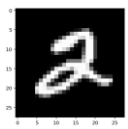




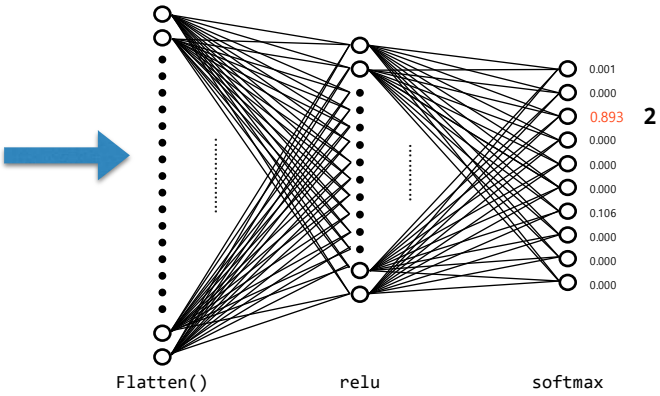
3,072개의 숫자

41

이미지 데이터 in 인공신경망



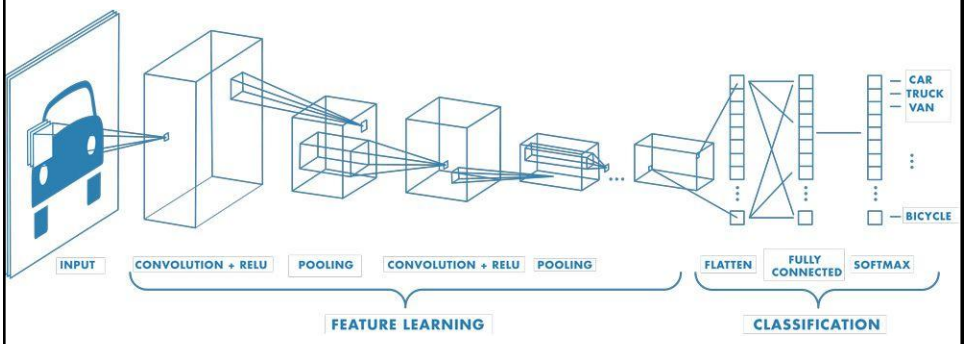
(28, 28)



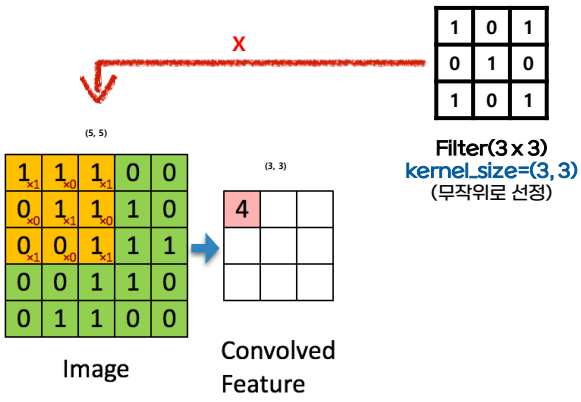
42

합성곱 신경망

• Convolutional Neural Network

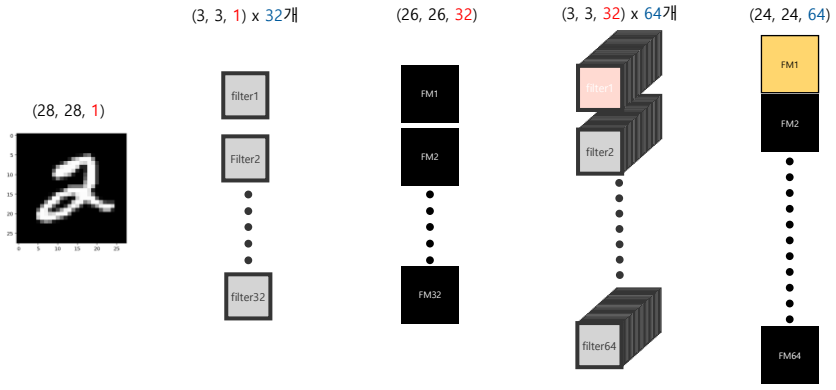


합성곱 연산 처리 절차



[출처 : http://deeplearning.stanford.edu/wiki/index.php/Feature_extraction_using_convolution]

필터



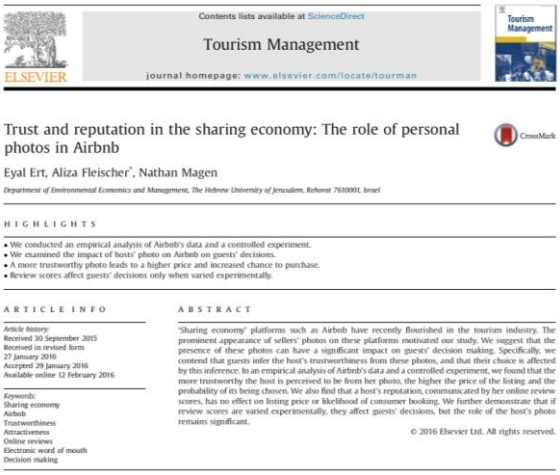
45

합성곱 신경망 체험

- Teachable Machine
  - https://teachablemachine.withgoogle.com/

46

연구 사례



Contents lists available at ScienceDirect

Tourism Management

journal homepage: www.elsevier.com/locate/tourman

Trust and reputation in the sharing economy: The role of personal photos in Airbnb

Eyal Ert, Aliza Fleischer*, Nathan Magen

Department of Environmental Economics and Management, The Hebrew University of Jerusalem, Rehovot 7610001, Israel

HIGHLIGHTS

- We conducted an empirical analysis of Airbnb's data and a controlled experiment.
- We examined the impact of hosts' photo on Airbnb on guests' decisions.
- A more trustworthy photo leads to a higher price and increased chance to purchase.
- Review scores affect guests' decisions only when varied experimentally.

ARTICLE INFO

Article history:
Received 30 September 2015
Received in revised form
27 January 2016
Accepted 29 January 2016
Available online 12 February 2016

Keywords:
Sharing economy
Airbnb
Trustworthiness
Attractiveness
Online reviews
Electronic word of mouth
Decision making

ABSTRACT

'Sharing economy' platforms such as Airbnb have recently flourished in the tourism industry. The prominent appearance of sellers' photos on these platforms motivated our study. We suggest that the presence of these photos can have a significant impact on guests' decision making. Specifically, we contend that guests infer the host's trustworthiness from these photos, and that their choice is affected by this inference. In an empirical analysis of Airbnb's data and a controlled experiment, we found that the more trustworthy the host is perceived to be from her photo, the higher the price of the listing and the probability of its being chosen. We also find that a host's reputation, communicated by her online review scores, has no effect on listing price or likelihood of consumer booking. We further demonstrate that if review scores are varied experimentally, they affect guests' decisions, but the role of the host's photo remains significant.

© 2016 Elsevier Ltd. All rights reserved.

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism management*, *55*, 62-73.

47

47

연구 사례

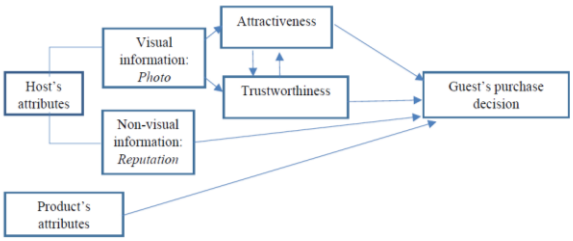


Fig. 1. Basic conceptual framework.

"In order to evaluate the perceived attractiveness of each apartment based on its main photo, we hired 260 workers (i.e., participants) on Amazon Mechanical Turk (Mturk) from the United States and Canada."

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism management*, *55*, 62-73.

48

48

연구 사례

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A computational framework for understanding antecedents of guests' perceived trust towards hosts on Airbnb

Le Zhang[a], Qiang Yan[a,*], Leihan Zhang[b]

[a] School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, PR China
[b] Institute of Computer Science and Technology, Peking University, Beijing 100871, PR China

ARTICLE INFO

Keywords:
Airbnb
Perceived trust
Antecedents
Sentiment
Self-description

ABSTRACT

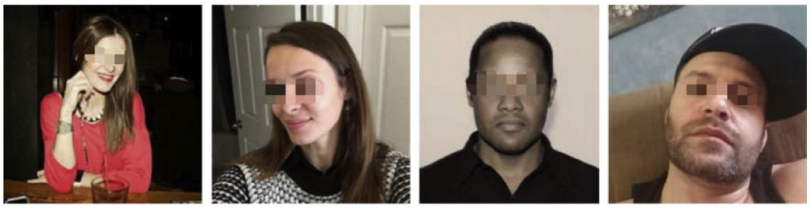
Several studies have researched the antecedents influencing the perceived trust of guests towards hosts on Airbnb typically relying on survey data. However, the contribution of these antecedents to trust building in a practical context remains unclear. To fill this gap, we focused on the antecedents within the manageable information about hosts and proposed a computational framework for understanding the antecedents influencing perceived trust. Specifically, perceived trust was proxied by the growth rate of bookings and the validity of the proxy method was proved through comparing with human labeled data. From the snapshot information about hosts, the antecedents were quantified through text mining and face recognition methods. The least square regression was applied to analyze and compare the influence of these antecedents. We found that the contribution of reputation is not less than the summation of all the other antecedents. Additionally, in terms of self-descriptions, it is worthwhile to pay more attention to interactions and services. Expressing positive sentiment in either self-descriptions or profile photo is also helpful. The response behavior pattern and the number of verifications also matter. At last, several effective trust prediction models were built by using deep neural network and the ensemble method. The findings shed light on the working of the antecedents in trust formation and can provide instructions for the transaction partners, designers and managers of online services in the sharing economy.

Zhang, L., Yan, Q., & Zhang, L. (2018). A computational framework for understanding antecedents of guests' perceived trust towards hosts on Airbnb. Decision Support Systems, 115, 105-116.

49

49

연구 사례



A B C D

Fig. 2. The profile photos of four hosts.

Table 2
Facial expressions detected by Face++.

Photo	Sad	Neutral	Disgust	Angry	Surprise	Fear	Happy	Key emotion	Emotion polarity
A	0.08	0.00	0.17	0.01	0.00	0.01	99.73	Happy	1
B	0.07	0.61	0.73	0.07	0.17	0.06	98.30	Happy	1
C	0.00	99.84	0.00	0.16	0.00	0.00	0.00	Neutral	0
D	90.34	9.21	0.02	0.01	0.01	0.22	0.20	Sad	−1

Zhang, L., Yan, Q., & Zhang, L. (2018). A computational framework for understanding antecedents of guests' perceived trust towards hosts on Airbnb. Decision Support Systems, 115, 105-116.

50

50

감사합니다.

51