데이터분석

- 감정분석 -

핀인사이트

데이터분석가 김현진

감정분석

감정분석이란?

- 문서, 단락, 문장 내 에서 극성 (예 : 긍정 / 부정 / 중립) 을 감지하는 텍스트 분석 방법
- 문서 내 감정표현을 분석하여 의견, 평가, 태도 등의 특징을 정량화된 자료로 제시함
- 텍스트 간의 비교우위를 밝힘으로써, 상대적 비교를 할 수 있음

이 영화는 너무 재미있다.

POSITIVE

이 영화는 1950년을 배경으로 한다.

NEUTRAL

이 영화는 스토리도 별로고 너무 재미없다

NEGATIVE

감정분석이란?

- 사용자는 자신의 생각과 감정을 블로그나 리뷰 같은 형태로 표현할 수 있고, 이는 비즈니스에서 매우 중요
- 설문 조사, 소셜 미디어 등 고객 피드백을 자동으로 분석하여 고객 의견을 듣고 제품에 반영할 수 있고
- 이를 통해 인사이트를 발견하고 서비스개선과 브랜트 평판 개선에 적용할 수 있음

이 영화는 너무 재미있다.

POSITIVE

이 영화는 1950년을 배경으로 한다.

NEUTRAL

이 영화는 스토리도 별로고 너무 재미없다

NEGATIVE

감정분석의 필요성

대량 데이터 처리 가능
직접 문서를 읽고 주요 감정을 판단할 수 있지만 많은 시간이 소요된다. 감정분석은 효율적인 방식으로 방대한 양의 데이터를 처리하는 데 도움이 된다.

추출의 일관성
텍스트에 감정 태그하는 것은 개인적인 경험, 생각 및 신념에 영향 받는 주관이다. 따라서 텍스트 분석을 직접 수행 할 때 나타나는 불일치를 고려할 필요가 없다. 일관된 기준을 적용하여 정확성을 높이고 더 나은 인사이트를 얻을 수 있다.

실시간 분석 가능
소셜 미디어, 고객 리뷰, 설문 조사 또는 고객 지원에 대한 감정분석을 실시간으로 수행하고 제품에 대한 의견을 얻을 수 있다. 감정분석 모델을 사용하여 문제를 빠르게 식별하고 즉시 조치하여 개선할 수 있다.

감정분석 활용 (1) - 브랜드 모니터링

- 브랜드 혹은 기업에 대한 온라인 반응은 은 브랜드(기업) 가치에 영향을 끼침
- 인터넷 뉴스, 블로그, 포럼 및 기타 텍스트를 분석하여 브랜드의 감정을 분석
- 응답에 적합한 구성원에게 자동 전달
- 브랜드 모니터링에 대해 적절한 조치로 브랜드 이미지 개선

감정분석 활용 (2) - VoC (Voice of Customer)

- 고객 서비스 개선
- 집계된 고객의 리뷰나, 제품사용 후기, 블로그 등을 분석함으로써 설문조사의 새로운 대체재로 활용
- 감정분석을 통해 제품이나 서비스에 대해 '매우 부정적'인 고객을 감지하여 대응하므로써,
- 사전에 파악하지 못했던 실제 고객의 인구통계, 관심사, 페르소나 등 잠재고객 세그먼트를 식별할 수 있음

감정분석 활용 (2) - VoC (Voice of Customer)

✔ 사례 : **Snickers : You're Not You When You're Hungry - Mars global case study(2009)**

사람들이 분노하는 정도가 높아질수록 스니커즈의 가격은 하락한다 / 출처 Ads of the World 공식 유튜브 캡처

- 2006년부터 키켓, 오레오, M&M등 다양한 초콜렛 브랜드가 나오게 되면서 스니커즈의 시장점유율이 계속 하강세에 머물렀다.

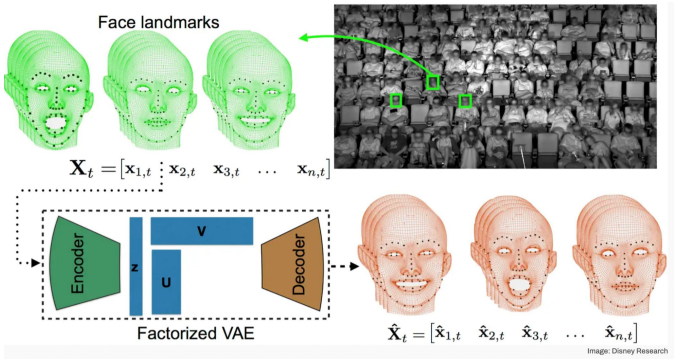
- As-is) 주로 남성들에게 어필하는 Niche 브랜드
- To-be) 전 세계 사람을 대상으로 나라와 지역에 구분없이 Universal Story로 컨셉광고를 전 세계를 대상으로 송출

- **How?**

출처 : https://www.warc.com/Content/41dac56a-1d8a-4dcc-bc65-72b370bc546c
참고 : https://m.blog.naver.com/businessinsight/221218399253

감정분석 활용 (2) - VoC (Voice of Customer)

✔ 사례 : **While you're watching Disney's films at the cinema, Disney can now watch you**



- 관객의 표정 감정분석을 통해 영화의 뒷부분 혹은 관객이 영화를 어떻게 평가할지에 대하여 예측함

- **How?**

출처 : https://emerj.com/ai-sector-overviews/artificial-intelligence-at-disney/

감정분석의 유형

감정 감지 (Emotion detection)

- 기쁨, 슬픔, 행복, 분노 등 감정을 감지
- 감정 사전 기반 분석 혹은 머신러닝/딥러닝 알고리즘을 사용.
- 문맥이해의 어려움으로 사전 기반으로 분석이 잘못 판별할 수 있음.

  ex. 미치도록 좋다

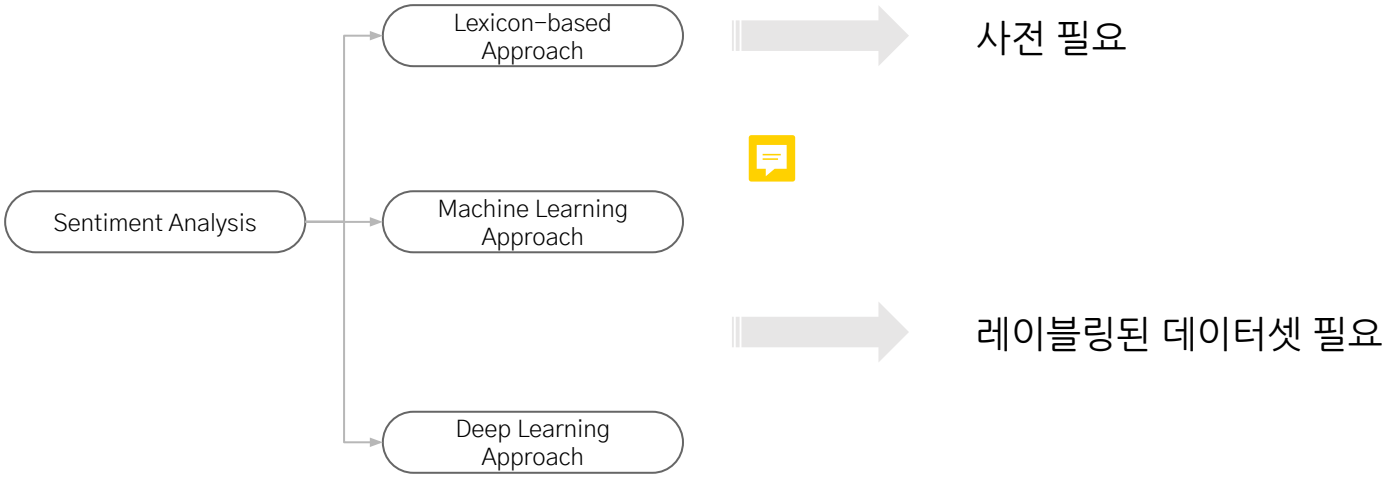
특성 기반 감정분석 (Aspect-based Sentiment Analysis)

- 텍스트 내에서 특성을 기준으로 감정분석
- 예. 카메라의 배터리 수명이 너무 짧습니다. 하지만 화면은 매우 큽니다.

  => 배터리(특성) : 짧다, 화면(특성): 크다

감정분석은 어떻게 할 수 있을까요?

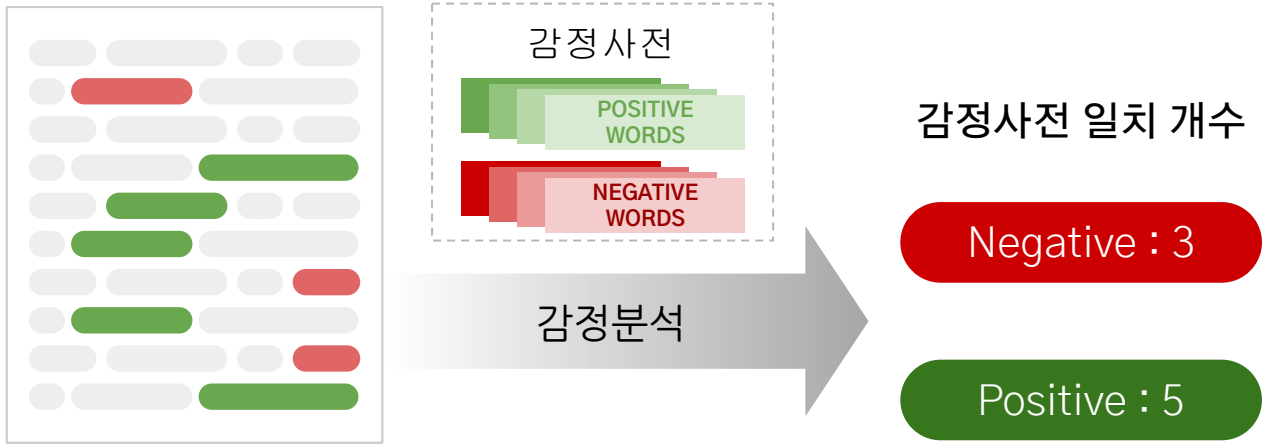
감정분석 방법



Lexicon-based approach

사전기반 감정분석

- 정의된 긍정, 부정 사전을 활용하여 일치 단어 등장 여부를 판단하여 측정하는 방법
- 사전의 질이 분석의 성능을 좌우함

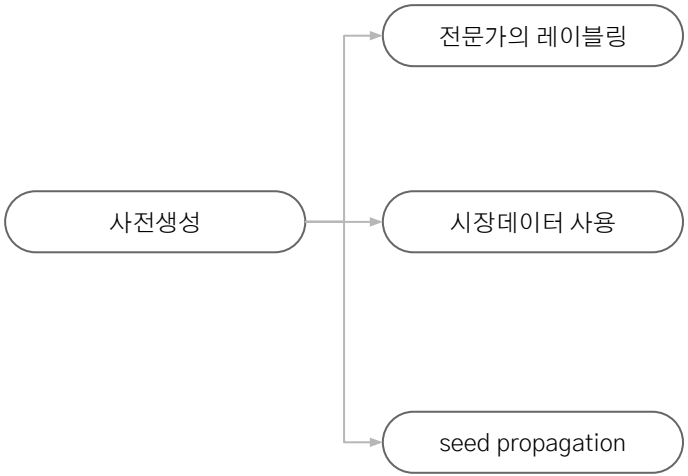


하지만, 사전 기반의 감정분석은 어렵습니다

사전기반 감정분석이 어려운 이유

- **문맥에 따른 감정분석이 어려움**
  - 사전기반 방식은 단순하게 사전에 등록된 단어를 기준으로 극성을 판단하기 때문에 문맥에 따른 감정분석이 어려움
  - 해결안 1 : ngram을 활용하여 문맥을 포함한 사전생성
  - 해결안 2 : Sequence를 처리할 수 있는 딥러닝(RNN, Transformer, BERT 등)
- **범용적 사전적용이 어려움**
  - 도메인에 따라 용어가 다르고, 긍부정 어휘도 다름. => 도메인별 사전 필요
- **한글의 경우 감정 사전이 부족**
  - 한글의 경우 사전 부족으로 사전기반 감정분석이 어려움

사전 생성 방법

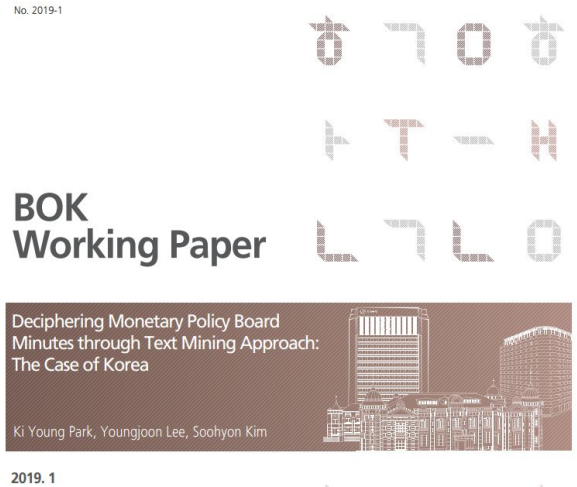


전문가 레이블링

- 생성하고자 하는 사전의 도메인에 전문지식을 가지고 있는 전문가들이 직접 문장을 읽어보며 긍정적인 문장인지, 부정적인 문장인지 직접 레이블을 다는 방법이다.
- 전문가가 직접 레이블을 붙이기 때문에 적은양의 데이터로도 좋을 결과를 얻을 수 있다.

시장데이터 사용

- 사전을 생성할때, 사람이 레이블을 붙이지 않고 레이블을 붙이는데 시장 데이터를 사용할 수 있다. 이 방법을 ‘시장 접근법'이라고 하고 ‘market approach’라고 부르기도 한다.
- 사용할 수 있는 시장데이터는 주가, 가격, 지수와 같은 데이터를 시장데이터로 사용할 수 있다.



한국은행 경제 연구원에서 발간하는「BOK 경제연구」2019-1호에 실린 논문

: 텍스트 마이닝을 활용한 금융통화 위원회 의사록 분석

텍스트 마이닝을 활용한 금융통화 위원회 의사록 분석

(Deciphering Monetary Policy Board Minutes through Text Mining Approach : The Case of Korea)

1 중앙은행의 커뮤티케이션은 통화정책의 방향, 경제 상황에 대한 판단등이 포함되어 있어 시장의 기대에 즉각적인 영향을 미칠 수 있다. 또한, 글로벌 금융위기 이후 이에 대한 관심이 고조 되었다.

2 중앙은행의 커뮤니케이션은 절제된 표현이 많아 일반적인 독해만으로는 내제된 정보를 추출하고, 그 영향을 분석하는데 한계가 있다.

텍스트 마이닝을 활용하여

금융통화 위원회 의사록의 어조로 새로운 인덱스를 만들고

이것을 중앙은행의 의도를 파악할 수 있는 도구로 활용할 수 있는지 검증

텍스트 마이닝을 활용한 금융통화 위원회 의사록 분석
(Deciphering Monetary Policy Board Minutes through Text Mining Approach : The Case of Korea)

polarity score를 기반으로 레이블링한 것이

금리상승(Hawkish)이면 “금리상승사전”, 금리하락(Dovish)이면 “금리하락사전” 으로 분류하여

두가지 사전을 만든다.

< 금리상승(Hawkish) 사전 >

1만 2106개

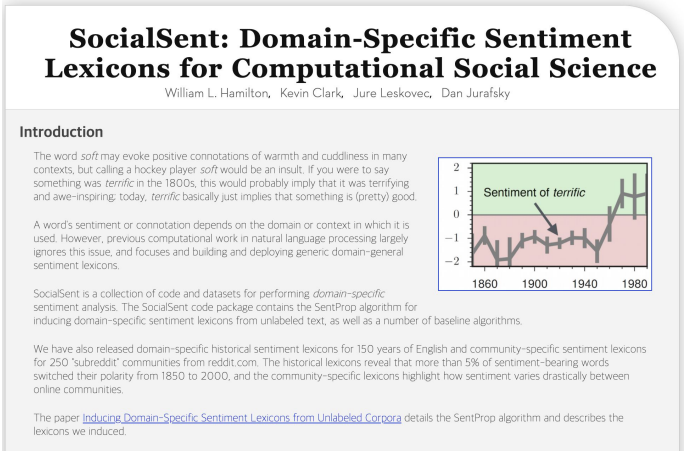
< 금리하락(Dovish) 사전 >

1만 3380개

seed propagation

seed propagation

- 문장에 레이블을 붙이지 않고 사전을 구축할 수 있는 방법
- “seed propagtion”이라는 단어관계 네트워크 그래프를 사용해 단어간의 극성점수를 구하는 것을 통해 사전을 생성한다.



SocialSent: Domain-Specific Sentiment Lexicons for Computational Social Science

William L. Hamilton, Kevin Clark, Jure Leskovec, Dan Jurafsky

Introduction

The word *soft* may evoke positive connotations of warmth and cuddliness in many contexts, but calling a hockey player *soft* would be an insult. If you were to say something was *terrific* in the 1800s, this would probably imply that it was terrifying and awe-inspiring; today, *terrific* basically just implies that something is (pretty) good.

A word's sentiment or connotation depends on the domain or context in which it is used. However, previous computational work in natural language processing largely ignores this issue, and focuses and building and deploying generic domain-general sentiment lexicons.

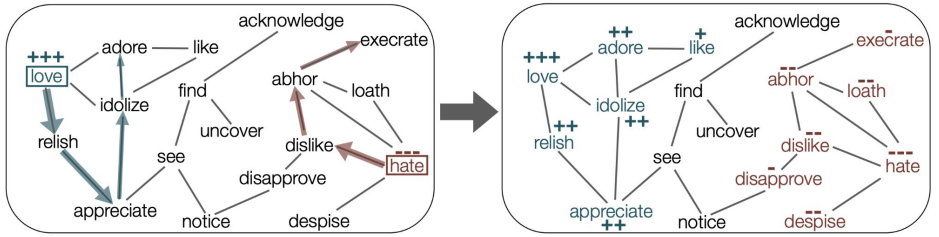
SocialSent is a collection of code and datasets for performing *domain-specific* sentiment analysis. The SocialSent code package contains the SentProp algorithm for inducing domain-specific sentiment lexicons from unlabeled text, as well as a number of baseline algorithms.

We have also released domain-specific historical sentiment lexicons for 150 years of English and community-specific sentiment lexicons for 250 "subreddit" communities from reddit.com. The historical lexicons reveal that more than 5% of sentiment-bearing words switched their polarity from 1850 to 2000, and the community-specific lexicons highlight how sentiment varies drastically between online communities.

The paper Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora details the SentProp algorithm and describes the lexicons we induced.

출처 : https://arxiv.org/pdf/1606.02820.pdf

Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora



a. Run random walks from seed words.

b. Assign polarity scores based on frequency of random walk visits.

Figure 3: Visual summary of the SentProp algorithm.

seed propagation

✔ **seed propagation 방법**

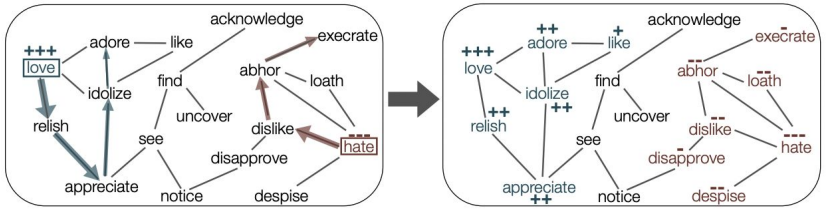
초기 Seed word set 구축 → **단어와 단어의 유사성을 계산하여 어휘그래프 생성** → **단어간의 극성점수 계산을 통한 사전구축**

word embedding

random walk method

Domain	Positive seed words	Negative seed words
Standard English	good, lovely, excellent, fortunate, pleasant, delightful, perfect, loved, love, happy	bad, horrible, poor, unfortunate, unpleasant, disgusting, evil, hated, hate, unhappy
Finance	successful, excellent, profit, beneficial, improving, improved, success, gains, positive	negligent, loss, volatile, wrong, losses, damages, bad, litigation, failure, down, negative
Twitter	love, loved, loves, awesome, nice, amazing, best, fantastic, correct, happy	hate, hated, hates, terrible, nasty, awful, worst, horrible, wrong, sad

Table 1: Seed words. The seed words were manually selected to be context insensitive (without knowledge of the test lexicons).



a. Run random walks from seed words. b. Assign polarity scores based on frequency of random walk visits.

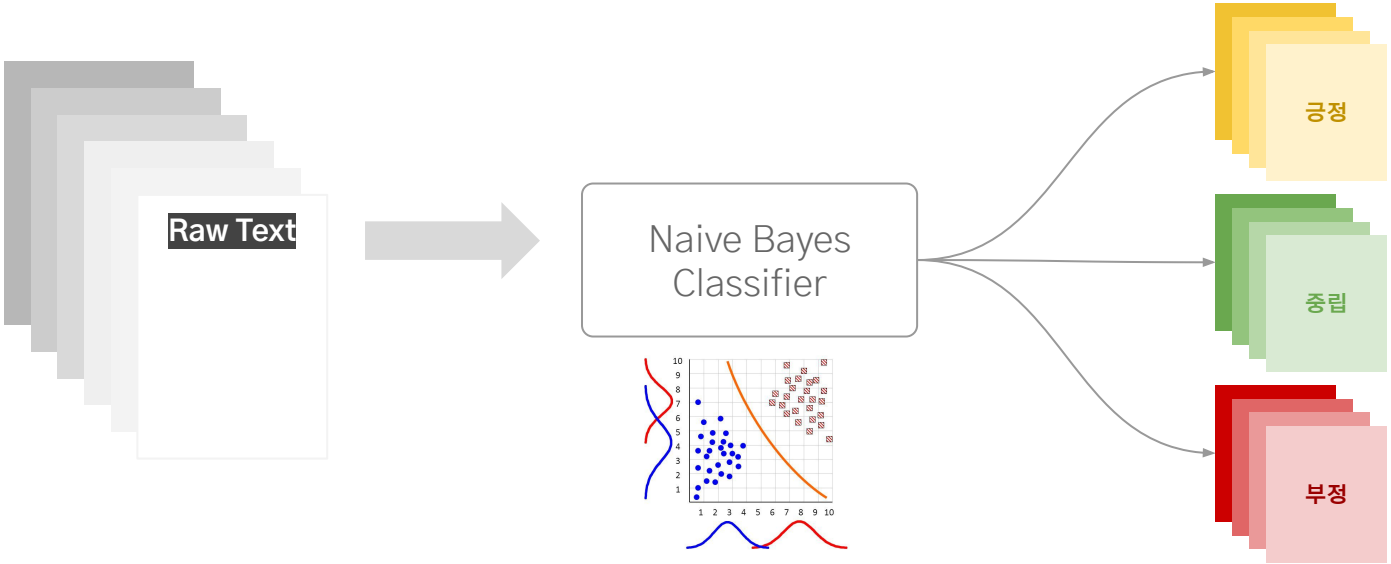
Figure 3: Visual summary of the SENTPROP algorithm.

Machine Learning approach

naive-bayes classification

나이브베이즈 분류기 활용 감정분석

- 감정분석도 분류 문제의 하나로 볼 수 있음
- 따라서 분류모델을 활용하여 감정분석이 가능함. 대신 감정레이블이 부착된 학습용 데이터가 필요



Bayes' theorem

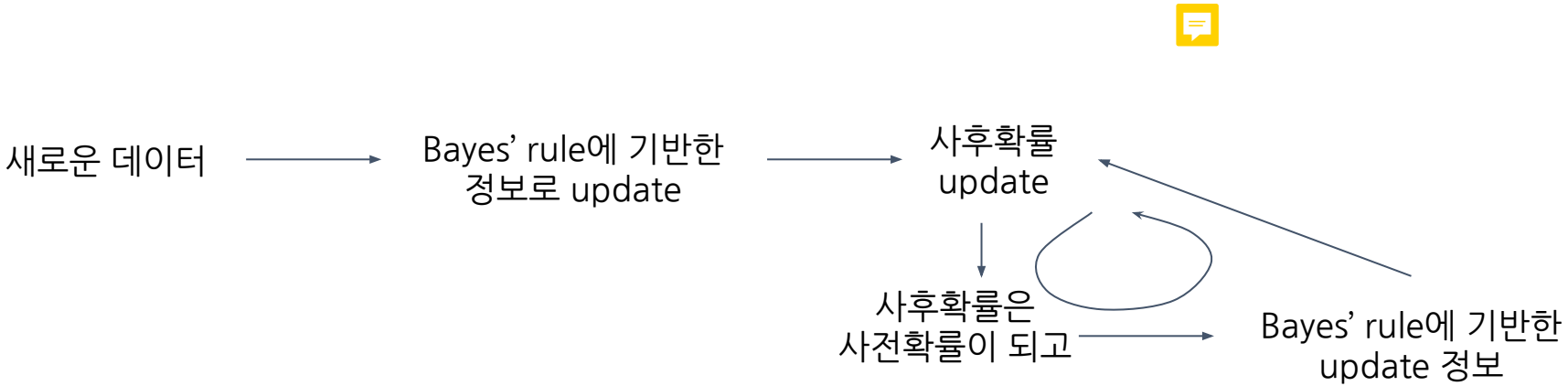
- 조건부 확률에 대한 수학적 정리
- 두 확률변수의 사전확률과 사후확률 사이의 관계를 나타내는 정리

사전확률
(prior)

$$P(X|E) = \frac{P(X,E)}{P(E)} = \frac{P(E|X)P(X)}{P(E)} = \frac{P(E|X)P(X)}{P(X|E)P(E)+P(X|\sim E)P(\sim E)}$$

사후확률
(posterior)

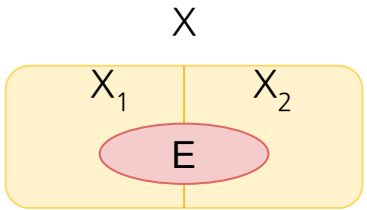
Bayes' theorem



- 어떤 의사결정이나 확률을 구할 때, 계속 발전된 방향으로 업데이트 시켜나가기 때문에 머신러닝 분야나 인공지능분야에서 베이즈 정리를 많이 사용한다.

Bayes' theorem 확장

사건 X_1, 사건 X_2가 서로 배타적(교집합 없음)이고 완전(X_1,X_2=X)한 경우



전체 확률 법칙에 의해 P(E)를 X_1과 X_2로 표현할 수 있다.

$$P(E) = P(E,X) = P(E, X_1) + P(E, X_2)$$
$$= P(E|X_1)P(X_1) + P(E|X_2)P(X_2)$$

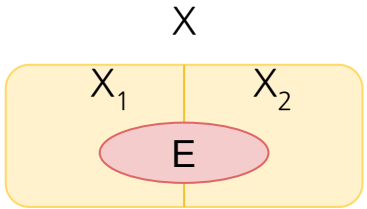
조건부 확률로 각각의 확률을 구하면?

$$P(X_1|E) = \ ?$$

$$P(X_2|E) = \ ?$$

Bayes' theorem 확장

사건 X_1, 사건 X_2가 서로 배타적(교집합 없음)이고 완전(X_1,X_2=X)한 경우



전체 확률 법칙에 의해 P(E)를 X_1과 X_2로 표현할 수 있다.

$$P(E) = P(E,X) = P(E, X_1) + P(E, X_2)$$
$$= P(E|X_1)P(X_1) + P(E|X_2)P(X_2)$$

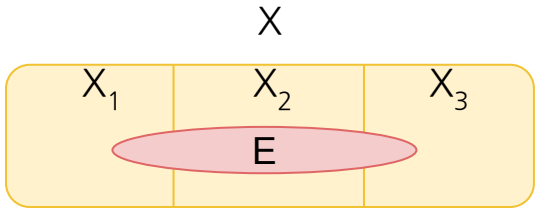
조건부 확률로 각각의 확률을 구하면?

$$P(X_1|E) = \frac{P(X_1,E)}{P(E)} = \frac{P(E|X_1)P(X_1)}{P(E)} = \frac{P(E|X_1)P(X_1)}{P(E|X_1)P(X_1)+P(E|X_2)P(X_2)}$$

$$P(X_2|E) = \frac{P(X_2,E)}{P(E)} = \frac{P(E|X_2)P(X_2)}{P(E)} = \frac{P(E|X_2)P(X_2)}{P(E|X_1)P(X_1)+P(E|X_2)P(X_2)}$$

Bayes' theorem 확장

사건 X_1, 사건 X_2 사건 X_3가 서로 배타적(교집합 없음)이고 완전(X_1,X_2,X_3=X)한 경우



전체 확률 법칙에 의해 P(E)를 X_1 , X_2 ,X_3으로 표현할 수 있다.

$$P(E) = P(E,X) = P(E, X_1) + P(E, X_2) + P(E, X_3)$$
$$= P(E|X_1)P(X_1) + P(E|X_2)P(X_2) + P(E|X_3)P(X_3)$$

조건부 확률로 각각의 확률을 구하면?

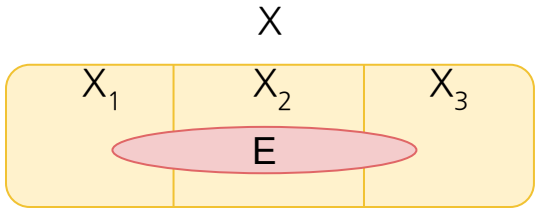
$$P(X_1|E) = \ ?$$

$$P(X_2|E) = \ ?$$

$$P(X_3|E) = \ ?$$

Bayes' theorem 확장

사건 X_1, 사건 X_2 사건 X_3가 서로 배타적(교집합 없음)이고 완전(X_1,X_2,X_3=X)한 경우



전체 확률 법칙에 의해 P(E)를 X_1 , X_2 ,X_3으로 표현할 수 있다.

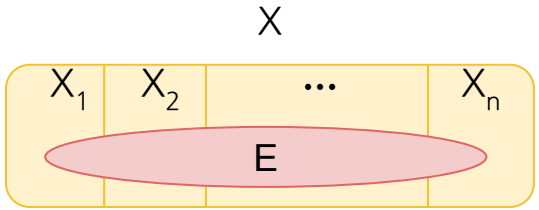
$$P(E) = P(E,X) = P(E, X_1) + P(E, X_2) + P(E, X_3)$$
$$= P(E|X_1)P(X_1) + P(E|X_2)P(X_2) + P(E|X_3)P(X_3)$$

조건부 확률로 각각의 확률을 구하면?

$$P(X_1|E) = \frac{P(X_1,E)}{P(E)} = \frac{P(E|X_1)P(X_1)}{P(E)} = \frac{P(E|X_1)P(X_1)}{P(E|X_1)P(X_1)+P(E|X_2)P(X_2)+P(E|X_3)P(X_3)}$$
$$P(X_2|E) = \frac{P(X_2,E)}{P(E)} = \frac{P(E|X_2)P(X_2)}{P(E)} = \frac{P(E|X_2)P(X_2)}{P(E|X_1)P(X_1)+P(E|X_2)P(X_2)+P(E|X_3)P(X_3)}$$
$$P(X_3|E) = \frac{P(X_3,E)}{P(E)} = \frac{P(E|X_3)P(X_3)}{P(E)} = \frac{P(E|X_3)P(X_3)}{P(E|X_1)P(X_1)+P(E|X_2)P(X_2)+P(E|X_3)P(X_3)}$$

Bayes' theorem 확장

사건 X_1, 사건 X_2, … , 사건 X_n이 서로 배타적(교집합 없음)이고 완전(X_1,X_2, … ,X_n=X)한 경우



전체 확률 법칙에 의해 P(E)를 X_1 , X_2 , ⋯ , X_n으로 표현할 수 있다.

$$P(E) = P(E,X) = P(E, X_1) + P(E, X_2) + \cdots + P(E, X_n)$$

$$= P(E|X_1)P(X_1) + P(E|X_2)P(X_2) + \cdots + P(E|X_n)P(X_n)$$

조건부 확률로 각각의 확률을 구하면?

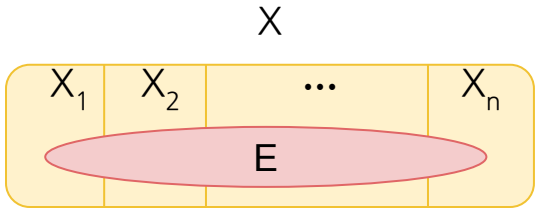
$P(X_1|E) =$?

일반화 >?

$P(X_i|E) =$?

Bayes' theorem 확장

사건 X_1, 사건 X_2, … , 사건 X_n이 서로 배타적(교집합 없음)이고 완전(X_1,X_2, … ,X_n=X)한 경우



전체 확률 법칙에 의해 P(E)를 X_1 , X_2 , ⋯ , X_n으로 표현할 수 있다.

$$P(E) = P(E,X) = P(E, X_1) + P(E, X_2) + \cdots + P(E, X_n)$$
$$= P(E|X_1)P(X_1) + P(E|X_2)P(X_2) + \cdots + P(E|X_n)P(X_n)$$

조건부 확률로 각각의 확률을 구하면?

$$P(X_1|E) = \frac{P(X_1,E)}{P(E)} = \frac{P(X_1|E)P(X_1)}{P(E)} = \frac{P(X_1|E)P(X_1)}{P(X_1|E)P(X_1)+P(X_2|E)P(X_2)+\ldots+P(X_n|E)P(X_n)}$$

일반화 >?

$$P(X_i|E) = \frac{P(X_i,E)}{P(E)} = \frac{P(X_i|E)P(X_i)}{P(E)} = \frac{P(X_i|E)P(X_i)}{P(X_1|E)P(X_1)+P(X_2|E)P(X_2)+\ldots+P(X_n|E)P(X_n)}$$

$$= \frac{P(X_i|E)P(X_i)}{\Sigma_i^n P(E|X_i)P(X_i)}$$

Bayes’ theorem 예제

총 50건의 메일이 왔습니다.

- 전체 메일 중 스팸메일 15건
- 전체 메일 중 “로또"라는 단어가 등장한 건수는 30건
- 스팸메일에 “로또"라는 단어가 등장한 건수는 10건

새로운 메일에 “로또"라는 단어가 포함되어 있을 경우, 이 메일이 스팸메일일 확률은?

그럼, “경품"이라는 단어조건이 추가되는 경우는 어떨까요?

❝ 새로운 메일이 왔습니다. “로또" 와 “경품"을 포함하고 있는경우 이 메일이 스팸일 확률은 얼마일까요? ❞

‘로또'라는 단어가 등장할 확률과 ‘경품'이라는 단어가 등장할 확률을 서로 관련이 없다.
이러한 경우를 생각할때,
‘로또'라는 단어가 등장할 확률과 ‘경품'이라는 단어가 등장할 확률이 서로 독립이다
라는 가정이 필요합니다

이러한 문제를 해결할 수 있는 모델이
나이브베이지안 분류기입니다

Naive Bayes Model

- Bayes' theorem에 특징(feature)들이 서로 확률적으로 독립(independent)이라는 가정을 추가

<독립가정>
사건 A와 사건 B가 서로 독립(independent)인 경우, 아래 식이 성립한다.

$$P(A \cap B) = P(A, B) = P(A)P(B)$$

A와 B가 서로 독립인 경우, B사건이 발생하는 것이 A사건에 전혀 영향을 주지 않기 때문에, 조건부 확률과 원래 확률이 같아지는 것을 확인 할 수 있다.

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

일반적인 독립가정은 앞 장과 같지만,
나이브베이지안은 베이지안에서 파생된 조건부확률 모형이죠?

그러기 때문에 '조건부독립'도 알아야합니다

Naive Bayes Model

< 조건부독립>

세 사건 X_1, X_2, E에 대해 다음 조건을 만족할때, X_1, X_2가 E에 대해

조건부독립(conditional independent)라고 한다.

$$P(X_1, X_2|E) = P(X_1|E)P(X_2|E)$$

일반화하면,

$$P(X_1, X_2, \ldots, X_{n-1}, X_n|E) = P(X_1|E)P(X_2|E)\ldots P(X_{n-1})P(X_n)$$

Naive Bayes Model 예제

총 50건의 메일이 왔습니다.

- 전체 메일 중 스팸메일15건
- 전체 메일 중 “로또"라는 단어가 등장한 건수는 30건
- 전체 메일 중 “경품"이라는 단어가 등장한 건수는 10건
- 스팸메일에 “로또"라는 단어가 등장한 건수는 10건
- 스팸메일에 “경품"이라는 단어가 등장한 건수는 7건 이라고 주어질 때,

새로운 메일에 “로또"와”경품"이라는 단어가 포함되어 있을 경우, 이 메일이 스팸메일일 확률은?

Naive Bayes Model 예제 풀이

총 50건의 메일이 왔습니다.

- 전체 메일 중 스팸메일15건 → P(스팸) = 15/50
- 전체 메일 중 “로또"라는 단어가 등장한 건수는 30건 → P(로또) = 30/50
- 전체 메일 중 “경품"이라는 단어가 등장한 건수는 10건 → P(경품) = 10/50
- 스팸메일에 “로또"라는 단어가 등장한 건수는 10건 → P(로또|스팸) = 10/15
  → P(로또|스팸아님) = 20/35

- 스팸메일에 “경품"이라는 단어가 등장한 건수는 7건 이라고 주어질 때, → P(경품|스팸) = 7/15
  → P(경품|스팸아님) = 3/35

새로운 메일에 “로또"와”경품"이라는 단어가 포함되어 있을 경우, 이 메일이 스팸메일일 확률은?

→ P(스팸|로또, 경품) = ?

만약, 기존에 존재하지 않은 새로운 단어가 등장하면 어떨까요?

Laplace Smoothing

Laplace Smoothing이란?

- 기존에 등장하지 않았던 새로운 단어가 입력되는경우 새로운 입력 단어의 확률은 0이다.
  그러므로 다른 단어들과 조건부확률을 구하게 되면 모든 확률값이 0이 나오게 된다.
  이러한 경우를 방지하기 위해 사용하는 방법이 “Laplace Smoothing”이다.
- Laplace Smoothing은 모든 단어가 1번이상 등장한다고 가정한다. 그러므로 계산시 분자에 ‘1’을 더해주고, 분모에는 ‘모든 unique 한 데이터의 수'를 더해준다.

만약, 메일에 등장하는 무수히 많은 단어중에서
단어 등장 빈도수가 극히 적은 경우는 어떨까요?

Underflow

Underflow 란?

- 수학적 계산으로 취급할 수 있는 범위보다 작아지는 경우를 Underflow라고한다.
- 즉, 0은 아니지만 0에 매우가까운수로 무한히 작아지는 경우

- Underflow는 로그를 취함으로써 방지 할 수 있다.

$$log(P(X_i|E)) = \frac{log(P(E|X_i)P(X_i))}{log(P(E))} = \frac{log(P(E|X_i)+log(P(X_i)))}{log(P(E))}$$

그럼 나이브베이지안 모델을 감성분석에 적용해 볼까요?

Naive Bayes 감정 분류

	토큰화 및 정제 된 단어들	분류
1	I love you	긍정
2	love happy weekend	긍정
3	bore work job	부정
4	I hate you	부정
5	bore weekend	부정
6	happy together	긍정

위와 같이 데이터가 주어진 경우,
“happy weekend”라는 단어가 포함된 문장이 긍정일 확률과 부정일 확률은 각각 얼마일까요?

→ P(positive|happy,weekend) = ?

→ P(negative|happy,weekend) = ?

Naive Bayes 감정 분류

1단계. 각 단어들이 긍정, 부정 분류에 등장한 빈도수를 구한다.

	토큰화 및 정제 된 단어들	분류
1	I love you	긍정
2	love happy weekend	긍정
3	bore work job	부정
4	I hate you	부정
5	bore weekend	부정
6	happy together	긍정

tokens | 분류	positive	negative
I	1	1
love	2	0
you	1	1
happy	2	0
weekend	1	1
bore	0	2
work	0	1
job	0	1
hate	0	1
together	1	0
합계	8	8

Naive Bayes 감정 분류

2단계. 각 단어의 조건부확률 값을 구한다.

tokens | 분류	positive	negative	P(w|positive)	P(w|negative)
you	1	1		
happy	2	0		
weekend	1	1		

$$P(happy|positive) = \frac{P(happy,positive)+1}{P(positive)+\text{전체}unique\text{한토큰수}} = \frac{2+1}{8+10} = 0.1667$$

$$P(happy|negative) = \frac{P(happy,negative)+1}{P(negative)+\text{전체}unique\text{한토큰수}} = \frac{0+1}{8+10} = 0.0556$$

Laplace smoothing 적용

Laplace Smoothing

Laplace Smoothing이란?

- 기존에 등장하지 않았던 새로운 단어가 입력되는경우 새로운 입력 단어의 확률은 0이다.

  그러므로 다른 단어들과 조건부확률을 구하게 되면 모든 확률값이 0이 나오게 된다.

  이러한 경우를 방지하기 위해 사용하는 방법이 “Laplace Smoothing”이다.

- Laplace Smoothing은 모든 단어가 1번이상 등장한다고 가정한다. 그러므로 계산시 분자에 ‘1’을 더해주고, 분모에는 ‘모든 unique 한 데이터의 수'를 더해준다.

- ‘긍정과 부정', ‘스팸과 스팸아님'과 같이 이진 분류의 경우 다음과 같은 일반화된 식을 사용하기도 한다.

$$P(X_i|E) = \frac{P(X_i,E)+k}{P(E)+2k}$$

Naive Bayes 감정 분류

2단계. 각 단어의 조건부확률 값을 구한다.

	토큰화 및 정제 된 단어들	분류
1	I love you	긍정
2	love happy weekend	긍정
3	bore work job	부정
4	I hate you	부정
5	bore weekend	부정
6	happy together	긍정

tokens | 분류	positive	negative	P(w|positive)	P(w|negative)
I	1	1	0.1111	0.1111
love	2	0	0.1667	0.0556
you	1	1	0.1111	0.1111
happy	2	0	0.1667	0.0556
weekend	1	1	0.1111	0.1111
bore	0	2	0.0556	0.1667
work	0	1	0.0556	0.1111
job	0	1	0.0556	0.1111
hate	0	1	0.0556	0.1111
together	1	0	0.1111	0.0556
합계	8	8		

Naive Bayes 감정 분류

3단계. 각 단어의 조건부 확률값을 사용하면 P(positive|happy,weekend)와 P(negative|happy,weekend)를 구할수 있다

tokens | 분류	positive	negative	P(w|positive)	P(w|negative)
you	1	1	0.1111	0.1111
happy	2	0	0.1667	0.0556
weekend	1	1	0.1111	0.1111
but	1	1	0.0556	0.1667

$$P(positive|happy, weekend) = \frac{P(happy,weekend|positive)P(positive)}{P(happy,weekend)} = \frac{P(happy|positive)P(weekend|positive)P(positive)}{P(happy,weekend|positive)+P(happy,weekend|negative)}$$

$$= \frac{exp(log(P(happy|positive)P(weekend|positive)P(positive)))}{exp(log(P(happy|positive)P(weekend|positive)P(positive)))+exp(log(P(happy|negative)P(weekend|negative)P(negative)))}$$

$$= \frac{exp(log(0.1667)+log(0.1111)+log(0.5))}{exp(log(0.1667)+log(0.1111)+log(0.5))+exp(log(0.0556)+log(0.1111)+log(0.5))} = \frac{exp(-1.792-2.197-0.693)}{exp(-1.792-2.197-0.693)+exp(-2.890-2.197-0.693)}$$

$$= \frac{0.00926}{0.00926+0.00309} = 0.75$$

Naive Bayes 감정 분류

3단계. 각 단어의 조건부 확률값을 사용하면 P(positive|happy,weekend)와 P(negative|happy,weekend)를 구할수 있다

tokens | 분류	positive	negative	P(w|positive)	P(w|negative)
you	1	1	0.1111	0.1111
happy	2	0	0.1667	0.0556
weekend	1	1	0.1111	0.1111
but	1	1	0.0556	0.1667

$$P(negative|happy, weekend) = \frac{P(happy,weekend|negative)P(negative)}{P(happy,weekend)} = \frac{P(happy|negative)P(weekend|negative)P(negative)}{P(happy,weekend|negative)+P(happy,weekend|negative)}$$

$$= \frac{exp(log(P(happy|negative)P(weekend|negative)P(negative)))}{exp(log(P(happy|positive)P(weekend|positive)+P(happy|negative)P(weekend|negative)))}$$

$$= \frac{exp(log(0.0556)+log(0.1111)+log(0.5))}{exp(log(0.1667)+log(0.1111)+log(0.5))+exp(log(0.0556)+log(0.1111)+log(0.5))} = \frac{exp(-2.890-2.197-0.693)}{exp(-1.792-2.197-0.693)+exp(-2.890-2.197-0.693)}$$

$$= \frac{0.00309}{0.00926+0.00309} = 0.25$$

실습 1 - Naive Bayes 감정분석

	토큰화 및 정제 된 단어들	분류
1	I love you	긍정
2	love happy weekend	긍정
3	bore work job	부정
4	I hate you	부정
5	bore weekend	부정
6	happy together	긍정

위와 같이 데이터가 주어진 경우,
“happy weekend”라는 단어가 포함된 문장이 긍정일 확률과 부정일 확률은 각각 얼마일까요?

P(positive|happy,weekend) = 75%

P(negative|happy,weekend) = 25%

실습2 - Naive Bayes 감정분석

- 사전기반 감정분석
- 네이버 영화리뷰데이터
  - 나이브베이지안

```
$ head ratings_train.txt
id	document	label
9976970 아 더빙.. 진짜 짜증나네요 목소리	0
3819312 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나	1
10265843	너무재밋었다그래서보는것을추천한다	0
9045019 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정	0
6483659 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다	1
5403919 막 걸음마 뗀 3세부터 초등학교 1학년생인 8살용영화.ㅋㅋㅋ...별반개도 아까움.	0
7797314 원작의 긴장감을 제대로 살려내지못했다.	0
9443947 별 반개도 아깝다 욕나온다 이응경 길용우 연기생활이몇년인지..정말 발로해도 그것보단 낫겟다 납치.감금만반복반복..이드라마는 가족도없다 연
7156791 액션이 없는데도 재미 있는 몇안되는 영화 1
```

링크 : https://github.com/e9t/nsmc

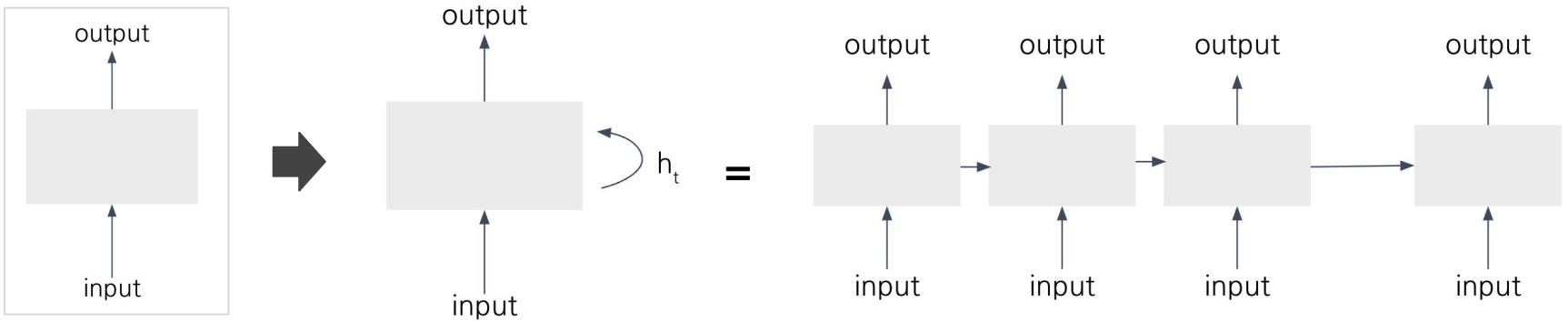
Deep Learning approach

RNN

RNN

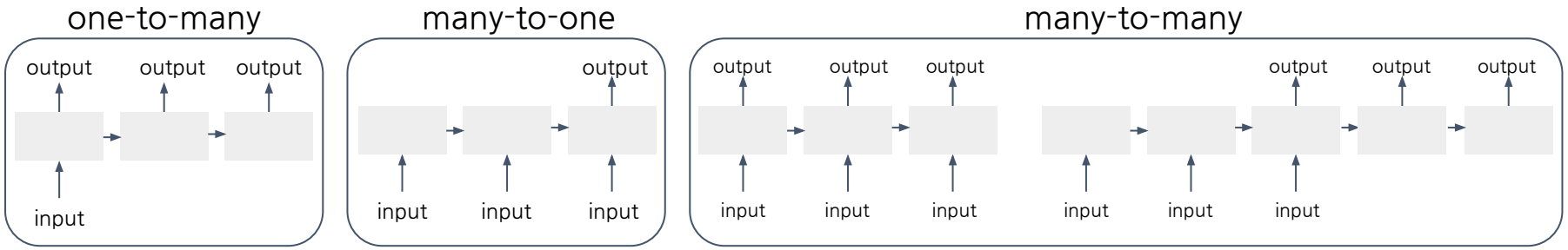
- RNN(Recurrent Neural Network)은 sequence 데이터나 Time-series를 처리하기 위해 만들어진 모델로 데이터 처리에 시간의 흐름을 반영할 수 있다.

  - 정보지속성



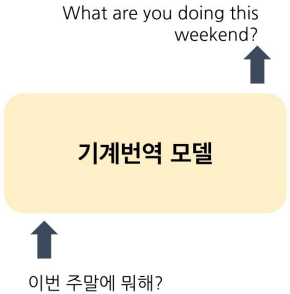
RNN의 종류

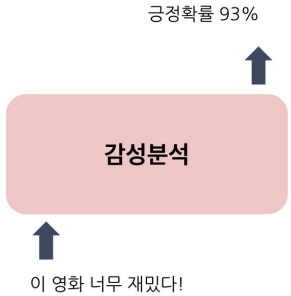
- RNN 응용모형



이미지 캡셔닝

[Google: Autonomous Image Captioning, '14. 11]





RNN의 한계

- 장기의존성 문제

  예측하려는 문장의 길이가 길어지는 경우 히든 스테이트에 충분한 정보가 담기지 못한다.

- Gradient Vanishing & Gradient Exploding

  RNN 학습시 발생하는 문제

  학습시 가중치에 따라 Gradient가 사라지거나 폭발하는 문제가 생긴다.

실습3 - RNN 감정분석

- 네이버 영화리뷰데이터
  - 딥러닝 (LSTM)

```
$ head ratings_train.txt
id      document        label
9976970 아 더빙.. 진짜 짜증나네요 목소리        0
3819312 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나        1
10265843        너무재밋었다그래서보는것을추천한다       0
9045019 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정        0
6483659 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다  1
5403919 막 걸음마 뗀 3세부터 초등학교 1학년생인 8살용영화.ㅋㅋㅋ...별반개도 아까움.      0
7797314 원작의 긴장감을 제대로 살려내지못했다.  0
9443947 별 반개도 아깝다 욕나온다 이응경 길용우 연기생활이몇년인지..정말 발로해도 그것보단 낫겠다 납치.감금만반복반복..이드라마는 가족도없다 연
7156791 액션이 없는데도 재미 있는 몇안되는 영화 1
```

링크 : https://github.com/e9t/nsmc