



영화 크라우드펀딩 성공과 실패에 텍스트 정보가 미치는 영향

팀 장 : 이 우 창

팀 원 : 남은수, 박종연, 엄선현,
채승완, 한서영, 표규진

목 차

클라우드 펀딩 소개

- 클라우드펀딩의 정의
- 클라우드펀딩의 종류
- 클라우드펀딩의 예시
- 문제 정의

사용 데이터

- 데이터 수집
- 데이터 전처리

데이터 분석

- LDA
- LDA 분석결과
- 인공지능기법을 이용한
클라우드펀딩 성공 예측
- CNN
- RNN
- LSTM
- 정확도 비교

결 론

- LDA 분석을 통한 결론
- 인공지능 기법을 이용
한 펀딩 성공과 실패
예측에 대한 결론

1 크라우드펀딩의 소개

1 크라우드펀딩의 정의

크라우드펀딩이란?

소셜 네트워크 서비스를 이용해 **소규모 후원을 받거나 투자** 등의 목적으로 인터넷과 같은 플랫폼을 통해 다수의 개인들로부터 **자금을 모으는 행위**.



1 크라우드펀딩의 소개

2 크라우드펀딩의 종류

기부형



보상이 없는
순수 기부

보상형



현물 또는 서비스와
같은 보상이 존재

투자형



주식, 채권 등의 증권
보상이 존재

1 크라우드펀딩의 소개

2 크라우드펀딩의 종류

기부형



보상이 없는
순수 기부

보상형



현물 또는 서비스와
같은 보상이 존재

투자형

기부형 또는 보상형 크라우드펀딩은
영화와 음악과 같은 문화예술분야에서
많이 이루어지고 있음

주식, 채권 등의 증권
보상이 존재

1 크라우드펀딩의 소개

3 크라우드펀딩 예시

≡ 프로젝트 둘러보기 프로젝트 올리기

tumbbug



로그인 / 회원가입



단편영화

숨겨야 하는 세상에서, 단편영화 <살인마 숨기기>

정은수



모인금액

1,242,000 원 82%

남은시간

10 일

후원자

38 명

펀딩 진행중

목표 금액인 1,500,000원이 모여야만 결제됩니다.
결제는 2019년 11월 4일에 다함께 진행됩니다.

프로젝트 참여하기



00. 안녕하세요.

저희는 가천대학교에서 단편영화를 만드는 학생들입니다!

심혈을 기울여 준비한 영화 <살인마 숨기기>는 현재 프리프로덕션 마지막 단계로, 메인 프로덕션을 코앞에 두고 있습니다.

그 과정에서 여러분의 도움을 얻고자 이렇게 글을 쓰게 되었습니다.

<살인마 숨기기>는 대한민국 사회 전반에 뿌리내린 '비난'에 대한 의문에서 시작했습니다. 왜 우리는 비난하지 않아도 될 일로 상대방을 비난하고, 비난받아야 하는 걸까요? 그런 풍토를 욕하면서도 왜 우리는 다시 서로를 비난하게 되는 걸까요?

저희는 '보통에 미치지 못하는 것'이 그 이유가 되는 것에 특히 주목했습니다. 신체적, 정신적 장애, 저학력, 심지어는 남들보다 취직을 늦게 했다는 것까지, 보통을 따라가지 못하는 것이 비난의 이유가 되곤 합니다. 하지만 그게 본인의 책임일 순 있어도 남에게 욕먹을 만한 일은 아니잖아요?

주인공 영서는 살인마의 딸입니다. 아빠의 살인이 영서의 죄는 아니지만, 살인마의 딸이라는 낙인이 두려웠던 영서는 아빠의 범죄를 숨기는 데에 집착하게 됩니다. 영서라는 캐릭터는 자

창작자 소개



정은수

창작자는 2017 <피의> 연출부, 2017 <코발트블루> 조연출, 2018 <완벽한 연기> 각본, 감독, 2018 <술오른 밤> 조연출로 활동했습니다. 팀원들 모두 영화에 관심이 많고 저희만의 훌륭한 작품을 만들어보자는 하나의 목표를 가지고 모인 16명의 가천대학교 미디어커뮤니케이션학과 학생들입니다.

마지막 로그인 한 시간 전

진행한 프로젝트 1 밀어준 프로젝트 0

창작자에게 문의하기

선택할 수 있는 7개의 선물이 있습니다

✓ 17명이 선택

10,000원 +

- 연딩 크레딧 (x 1)
- 비공개 URL (x 1)

예상 전달일 2019년 12월 31일

선물 선택하고 밀어주기

1 크라우드펀딩의 소개

3 크라우드펀딩 예시

≡ 프로젝트 둘러보기 프로젝트 올리기

tumbbug

🔍 로그인 / 회원가입 👤

단편영화

숨겨야 하는 세상에서, 단편영화 <살인마 숨기기>

👤 정은수



모인금액
1,242,000 원 82%

남은시간
10 일

후원자
38 명

펀딩 진행중

목표 금액인 1,500,000원이 모여야만 결제됩니다.
결제는 2019년 11월 4일에 다함께 진행됩니다.

프로젝트 참여하기



크라우드펀딩에서는 목표금액을
100%이상 달성하면, 모금액은
프로젝트진행자에게 조달되지만,
만약 미달되면 후원금은
후원자에게 다시 환급.

1 크라우드펀딩의 소개

3 크라우드펀딩 예시

00. 안녕하세요,

저희는 가천대학교에서 단편영화를 만드는 학생들입니다!

심혈을 기울여 준비한 영화 <살인마 숨기기>는 현재 프리프로덕션 막바지 단계로, 메인 프로모션을 코앞에 두고 있습니다.

그 과정에서 여러분의 도움을 얻고자 이렇게 글을 쓰게 되었습니다.

<살인마 숨기기>는 대한민국 사회 전반에 뿌리내린 '비난'에 대한 의문에서 시작했습니다. 왜 우리는 비난하지 않아도 될 일로 상대방을 비난하고, 비난받아야 하는 걸까요? 그런 풍토를 욕하면서도 왜 우리는 다시 서로를 비난하게 되는 걸까요?

저희는 '보통에 미치지 못하는 것'이 그 이유가 되는 것에 특히 주목했습니다. 신체적, 정신적 장애, 저학력, 심지어는 남들보다 취직을 늦게 했다는 것까지, 보통을 따라가지 못하는 것이 비난의 이유가 되곤 합니다. 하지만 그게 본인의 책임일 순 있어도 남에게 욕먹을 만한 일은 아니잖아요?

주인공 영서는 살인마의 딸입니다. 아빠의 살인이 영서의 죄는 아니지만, 살인마의 딸이라는 낙인이 두려웠던 영서는 아빠의 범죄를 숨기는 데에 집착하게 됩니다. 영서라는 캐릭터는 자

창작자 소개



정은수

창작자는 2017 <피의> 연출부, 2017 <코발트블루> 조연출, 2018 <완벽한 연기> 각본, 감독, 2018 <술오른 밤> 조연출로 활동했습니다. 팀원들 모두 영화에 관심이 많고 저희만의 훌륭한 작품을 만들어보자는 하나의 목표를 가지고 모인 16명의 가천대학교 미디어커뮤니케이션학과 학생들입니다.

마지막 로그인 한 시간 전

진행한 프로젝트 1 | 밀어준 프로젝트 0

✉ 창작자에게 문의하기

선택할 수 있는 7개의 선물이 있습니다

✓ 17명이 선택

10,000원 +

- 엔딩 크레딧 (x1)
- 비공개 URL (x1)

예상 전달일 2019년 12월 31일

선물 선택하고 밀어주기

성공적인 자금조달을 위해서는
프로젝트 창작자가 크라우드펀딩
게시글의 내용을 전략적으로 잘
작성할 필요가 있음.

1 크라우드펀딩의 소개

4 문제 정의

소개글

00. 안녕하세요,

저희는 가천대학교에서 단편영화를 만드는 학생들입니다!

심혈을 기울여 준비한 영화 <살인마 숨기기>는 현재 프리프로덕션 막바지 단계로, 메인 프로모션을 코앞에 두고 있습니다.

그 과정에서 여러분의 도움을 얻고자 이렇게 글을 쓰게 되었습니다.

<살인마 숨기기>는 대한민국 사회 전반에 뿌리내린 '비난'에 대한 의문에서 시작했습니다. 왜 우리는 비난하지 않아도 될 일로 상대방을 비난하고, 비난받아야 하는 걸까요? 그런 풍토를 욕하면서도 왜 우리는 다시 서로를 비난하게 되는 걸까요?

저희는 '보통에 미치지 못하는 것'이 그 이유가 되는 것에 특히 주목했습니다. 신체적, 정신적 장애, 저학력, 심지어는 남들보다 취직을 늦게 했다는 것까지, 보통을 따라가지 못하는 것이 비난의 이유가 되곤 합니다. 하지만 그게 본인의 책임일 순 있어도 남에게 욕먹을 만한 일은 아니잖아요?

주인공 영서는 살인마의 딸입니다. 아빠의 살인이 영서의 죄는 아니지만, 살인마의 딸이라는 낙인이 두려웠던 영서는 아빠의 범죄를 숨기는 데에 집착하게 됩니다. 영서라는 캐릭터는 자

창작자 소개



정은수

창작자는 2017 <피의> 연출부, 2017 <코발트블루> 조연출, 2018 <완벽한 연기> 각본, 감독, 2018 <술오른 밤> 조연출로 활동했습니다. 팀원들 모두 영화에 관심이 많고 저희만의 훌륭한 작품을 만들어보자는 하나의 목표를 가지고 모인 16명의 가천대학교 미디어커뮤니케이션학과 학생들입니다.

마지막 로그인 한 시간 전

진행한 프로젝트 1 | 일어난 프로젝트 0

보상

에게 문의하기

선택할 수 있는 7개의 선물이 있습니다

✓ 17명이 선택

10,000원 +

- 엔딩 크레딧 (x1)
- 비공개 URL (x1)

예상 전달일 2019년 12월 31일

선물 선택하고 밀어주기

소개글, 보상등 크라우드펀딩
게시글의 여러가지 요소들이
펀딩 성공에 영향을 줄 수 있음.

1 크라우드펀딩의 소개

4 문제 정의

소개글

00. 안녕하세요,

저희는 가천대학교에서 단편영화를 만드는 학생들입니다!

보상, SNS 광고, 주기적인 업데이트

등 다양한 수치적 요인들이 펀딩 성공에 영향을 미친다는 것은 기존

에 많은 연구가 있었음.

저희는 '보통에 미치지 못하는 것'이 그 이유가 되는 것에 특히 주목했습니다. 신체적, 정신적 장애, 저학력, 심지어는 남들보다 취직을 늦게 했다는 것까지, 보통을 따라가지 못하는 것이 비난의 이유가 되곤 합니다. 하지만 그게 본인의 책임일 순 있어도 남에게 욕먹을 만한 일은 아니잖아요?

주인공 영서는 살인마의 딸입니다. 아빠의 살인이 영서의 죄는 아니지만, 살인마의 딸이라는 낙인이 두려웠던 영서는 아빠의 범죄를 숨기는 데에 집착하게 됩니다. 영서라는 캐릭터는 자

창작자 소개



정은수

창작자는 2017 <피의> 연출부, 2017 <코발트블루> 조연출, 2018 <완벽한 연기> 각본, 감독, 2018 <술오른 밤> 조연출로 활동했습니다. 팀원들 모두 영화에 관심이 많고 저희만의 훌륭한 작품을 만들어보자는 하나의 목표를 가지고 모인 16명의 가천대학교 미디어커뮤니케이션학과 학생들입니다.

마지막 로그인 한 시간 전

진행한 프로젝트 1 | 잃어준 프로젝트 0

보상

에게 문의하기

선택할 수 있는 7개의 선물이 있습니다

✓ 17명이 선택

10,000원 +

- 엔딩 크레딧 (x1)
- 비공개 URL (x1)

예상 전달일 2019년 12월 31일

선물 선택하고 잃어주기

소개글, 보상등 크라우드펀딩
게시글의 여러가지 요소들이
펀딩 성공에 영향을 줄 수 있음.

1 크라우드펀딩의 소개

4 문제 정의

소개글

00. 안녕하세요,

저희는 가천대학교에서 단편영화를 만드는 학생들입니다!

심혈을 기울여 준비한 영화 <살인마 숨기기>는 현재 프리프로덕션 막바지 단계로, 메인 프로모션을 코앞에 두고 있습니다.

그 과정에서 여러분의 도움을 얻고자 이렇게 글을 쓰게 되었습니다.

<살인마 숨기기>는 대한민국 사회 전반에 뿌리내린 '비난'에 대한 의문에서 시작했습니다. 왜 우리는 비난하지 않아도 될 일로 상대방을 비난하고, 비난받아야 하는 걸까요? 그런 풍토를 욕하면서도 왜 우리는 다시 서로를 비난하게 되는 걸까요?

저희는 '보통에 미치지 못하는 것'이 그 이유가 되는 것에 특히 주목했습니다. 신체적, 정신적 장애, 저학력, 심지어는 남들보다 취직을 늦게 했다는 것까지, 보통을 따라가지 못하는 것이 비난의 이유가 되곤 합니다. 하지만 그게 본인의 책임일 순 있어도 남에게 욕먹을 만한 일은 아니잖아요?

주인공 영서는 살인마의 딸입니다. 아빠의 살인이 영서의 죄는 아니지만, 살인마의 딸이라는 낙인이 두려웠던 영서는 아빠의 범죄를 숨기는 데에 집착하게 됩니다. 영서라는 캐릭터는 자

창작자 소개



하지만 펀딩 소개글에 대한 텍스트 정보가 펀딩 성공에

창작자는 2017 <피의> 연출부, 2017 <코발트블루> 조연출, 2018 <완벽한 연기> 각본, 감독, 2018 <술오른 밤> 조연출로 활동했습니다. 팀원을 모두 고 저희만의 훌륭한 작품을 만들어보자는 취지로 모인 16명의 가천대학교 미디어커뮤니케이션학과 학생들입니다.

마지막 로그인 한 시간 전

진행한 프로젝트 1 | 들어온 프로젝트 0

따라서, 문화예술분야에서 크라우드펀딩 소개글에 대한

선택할 수 있는 7개의 선택이 있습니다

✓ 17명이 선택

10,000원 +

- 연딩 크레딧 (x1)
- 비공개 URL (x1)

예상 전달일 2019년 12월 31일

선물 선택하고 물어주기

미치는 영향에 대한 연구는 미진하였음.

소개글, 보상등 크라우드펀딩

게시글의 여러가지 요소들이

펀딩 성공에 영향을 주 수 있음.

텍스트정보가 펀딩 성공에 미치는 영향에 대해 연구.

2 사용 데이터

1 데이터 수집

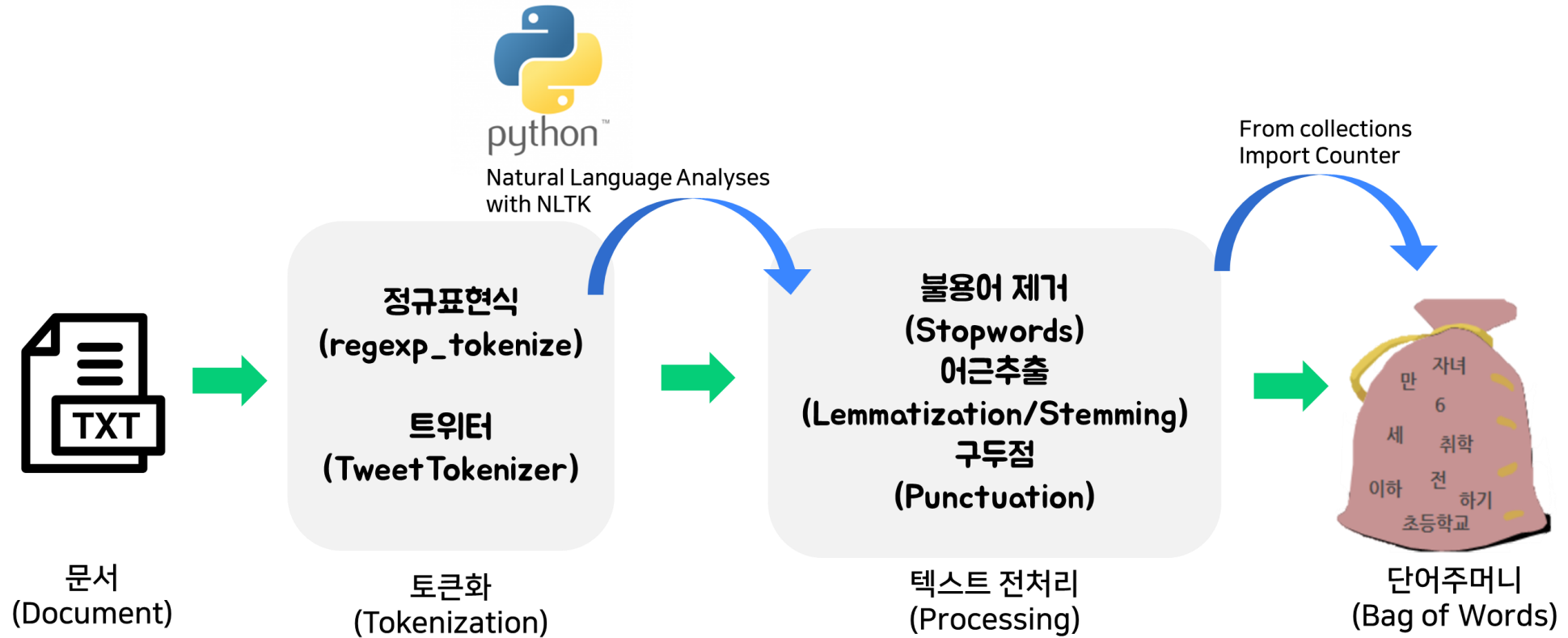


A	B	C	D	E	F	G	H	I	J
label	name	main							
1	AHEAD OF AHEAD OF THE CURVE	With a fist full of credit cards, a lucky run at the horse track, and							
1	THE DISCOWE DID IT! Thank You Discoverers!!!	"A bright, thoughtful indie...it's great to see Griffin							
1	Big Shot	Big Shot is a short film about an 11-year old rapper named Armando, whose viral video							
1	The Death THE STORY	The Death of "Superman Lives": What Happened? is a Feature Length Docu							
1	"Sriracha"	Amazing! We hit \$20,000! Thank you all so much for supporting this film! Today's the la							
1	Married in MARRIED IN SPANDEX	is a love story about Amanda and Rachel, a young lesbian coup							
1	MAKING V	You can be Making Waves too; simply click on the pledge button -->What is MAKING							
1	Automated The Project	Automated Futures is an experimental documentary film about the materi							
1	Heaven A	Thanks for checking out our Kickstarter campaign. If you "like" us, please consider help							
1	231 Mauje 231 MAUJER	is a self-organized, non-institutional, live/work space for filmmakers and a							
1	Morpheus The Morpheus Stabilizer	will be available for pre-order here after the campaign ends. T							
1	Drag Dad	Wow! We have met our goal! Thank you all so very much for making this happen. The							
1	SAVE MY IA Letter from Marianna	This is a true story – you can save my film in 14 days and my c							
1	Interactive Decagon	is an ongoing horror fantasy experience on the web. The film stars supermod							
1	Adagio in	We've had an amazing run here on Kickstarter - thanks to all who've backed and share							

- 문화예술분야에서 가장 규모가 큰 크라우드 펀딩 사이트인 **KickStater**에서 **2010년 1월 ~ 2019년 6월**까지 **“영화”**로 분류된 **크라우드펀딩 게시글 총 5,095개**를 크롤링을 통하여 수집.
- 수집된 항목은 **크라우드펀딩 성공 여부, 크라우드펀딩 소개글로 설정.**

2 사용 데이터

2 데이터 전처리



2 사용 데이터

2 데이터 전처리

1. 정규표현식

1. 정규표현식을 통해 영어 및 숫자를 제외한 문자열 삭제
2. 영어에서는 두 글자 이하 단어는 의미가 없는 경우가 대다수로 삭제

3. 불용어 처리

1. 기본적인 불용어 제거
Ex) a, the 등등
2. 불용어 사전 리스트를 추가하여 제거
Ex) html, https, scott, anna 등등

2. 토큰화

1. TweetTokenizer를 활용한 형태소 추출 및 영어 대소문자 변환
2. I don't like apple
=> I / don't / like / apple

4. 표준화

1. LancasterStemmer 통한 표준화
Ex) running , runs => run
2. 문자열 패딩
=> 행렬 연산을 수행하기 위함

2 사용 데이터

2 데이터 전처리

1. 정규표현식

1. 정규표현식을 통해 영어 및 숫자를 제외한 문자열 삭제
2. 영어에서는 두 글자 이하 단어는 의미가 없는 경우가 대다수로 삭제

⇒ 1-1) 킥스타터의 경우 영어외의 다른 언어로 작성된 내용들이 존재하여 영어와 숫자외의 단어들을 python을 통해 제거하는 작업을 진행하였다

⇒ 2-1) 영어에서 두 글자 이하인 단어는 불용어로 빈도는 높지만 의미가 없는 경우가 많아 분석의 불필요하여 삭제하였다
e.x) to, be, he, if, or, of, no, Mr, in, so 등

2 사용 데이터

2 데이터 전처리

2. 토큰화

1. TweetTokenizer를 활용한 형태소 추출
및 영어 대소문자 변환

2. I don't like apple
=> I / don't / like / apple

⇒ 2-1) He와 he는 같은 단어지만 형태소 추출 후 컴퓨터가 단어를 이해할 수 있도록 one-hot-encoding 또는 정수 인코딩으로 전처리하는 과정에서 두 단어는 다른 단어로 인식되기 때문에 대소문자를 통일하는 작업을 진행하였다.

⇒ 2-2) 자연어처리 패키지인 NLTK에서 여러 가지 토큰화 라이브러리를 제공해준다. 그중에서도 TweetTokenizer를 활용한 이유는 어퍼스트로피 때문이다. TweetTokenizer의 경우 I don't like apple을 I / don't / like / apple로 형태소를 추출한다. word_tokenize의 경우 I / do / n't / like / apple로 추출하여 단어 본 뜻이 왜곡될 수 있으므로 본 연구에서는 TweetTokenizer를 사용하였다.

2 사용 데이터

2 데이터 전처리

3. 불용어 처리

1. 기본적인 불용어 제거

Ex) a, the 등등

2. 불용어 사전 리스트를 추가하여 제거

Ex) html, https, scott, anna 등등

⇒ 3-1) 단어 빈도는 많이 등장하지만 의미가 없는 단어를 불용어라고 한다. Text 분석시 불용어를 제거해주지 않으면 단어 가중치를 구하는 과정에서 모수가 증가하여 왜곡되기 때문에 NLTK에서 제공하는 불용어 사전을 이용하여 불용어를 제거해주었다.

⇒ 3-2) 또한 NLTK에서 제공하는 불용어 사전에만 의존하지 않고 scott, anna 등의 대명사 및 불용어라고 판단되는 단어들을 불용어 사전에 추가하여 제거해주었다.

2 사용 데이터

2 데이터 전처리

4. 표준화

1. LancasterStemmer 통한 표준화
Ex) running , runs => run
2. 문자열 패딩
=> 행렬 연산을 수행하기 위함

⇒ 4-1) running, runs의 기본형은 run이다. 하지만 컴퓨터가 단어를 인식하는 방법은 one-hot-encoding 또는 정수인코딩으로 runs와 run을 다른 단어로 인식하기 때문에 단어의 뿌리가 같은 단어들을 표준화 시켜주었다.

⇒ 4-2) 본 연구에서는 정수인코딩을 통해 text데이터를 수치 데이터로 변환하였다. 정수인코딩의 경우 각 단어들을 고유 인덱스 번호를 부여한다. 하지만 사례별로 단어들의 길이가 달라 행렬연산을 수행할 수 없다. 그래서 단어가 가장 긴 문장을 기준으로 나머지 문장들을 0으로 처리하는 과정을 거친다.

3 데이터 분석

1 LDA

LDA : 문서에서 단어들 간의 관련성에 따라 문서 내의 **잠재된 주제**를 찾아내는 토픽 모델링 기법

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

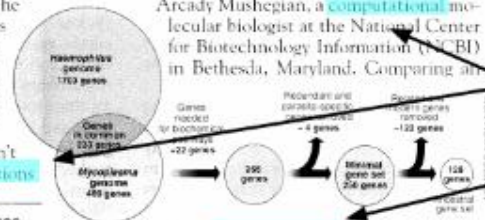
Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

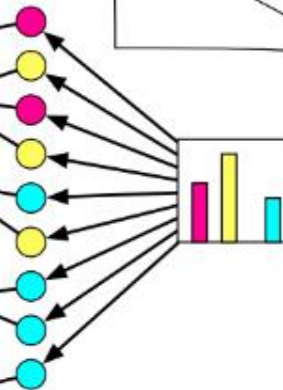


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 271 • 24 MAY 1996

Topic proportions & assignments



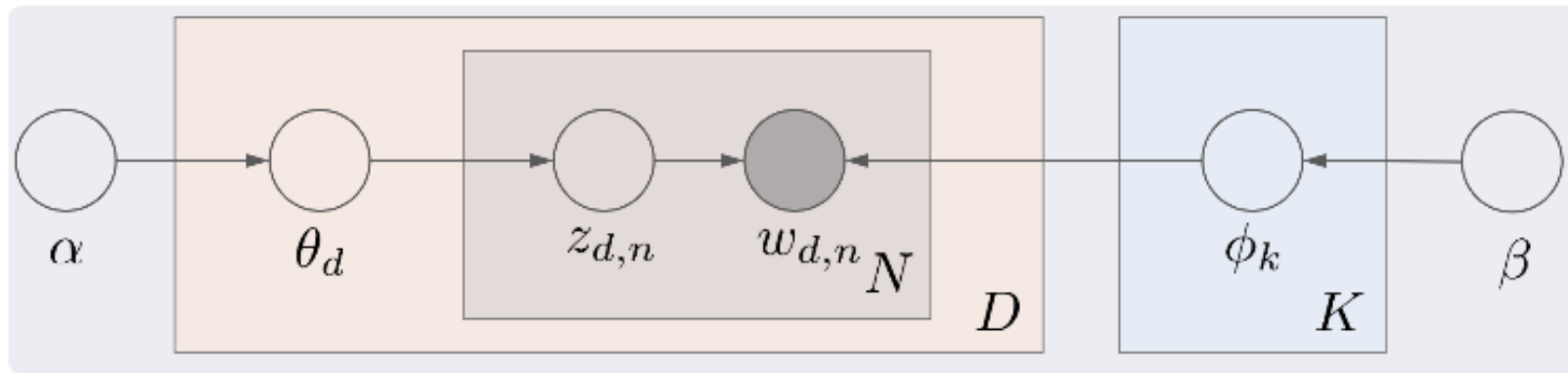
3 데이터 분석

1 LDA

LDA : 토픽의 단어 분포와 문서의 토픽 분포의 결합으로 문서 내 단어들이 생성.

LDA에서는 문서 내 단어를 가지고 토픽의 단어분포, 문서의 토픽분포를 추정.

LDA에서의 확률 과정과 결합 확률을 그림과 수식으로 나타내면 다음과 같음.



$$p(\phi_{1:k}, \phi_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\phi_i | \beta) \prod_{d=1}^D p(\theta_d | \alpha) \left\{ \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \phi_{1:k}, z_{d,n}) \right\}$$

3 데이터 분석

1 LDA

$$p(\phi_{1:k}, \phi_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\phi_i | \beta) \prod_{d=1}^D p(\theta_d | \alpha) \left\{ \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \phi_{1:k}, z_{d,n}) \right\}$$

여기서, α 는 문서의 토픽 분포 생성을 위한 디리클레 분포의 파라미터이며 β 는 토픽 단어 분포 생성을 위한 디리클레 분포의 파라미터이다. θ_d 는 d 번째 문서가 가진 토픽 비중을 나타내는 벡터이며, 전체 토픽 개수인 K 만큼의 길이를 가진다. ϕ_k 는 k 번째 토픽에 해당하는 벡터이며 말뭉치 전체의 단어 개수만큼의 길이를 가진다. $z_{d,n}$ 은 d 번째 문서의 n 번째 단어가 어떤 토픽에 해당하는지 할당해주는 역할을 한다. $w_{d,n}$ 은 d 번째 문서에 나타난 n 번째 단어를 의미한다.

위의 수식의 경우, 말뭉치로부터 관찰 가능한 $w_{d,n}$ 을 제외한 모든 변수가 미지수가 된다. LDA는 $p(z, \phi, \theta | w)$ 를 최대로 만드는 z, ϕ, θ 를 찾게 된다. $p(w)$ 는 z, ϕ, θ 의 모든 경우의 수를 고려한 각 단어 w 의 등장 확률을 말한다.

3 데이터 분석

1 LDA

실제 d 번째 문서 i 번째 단어의 토픽 $z_{d,i}$ 가 j 번째에 할당될 확률은 다음과 같다.

$$p(z_{d,i} = j | z_{-i}, w) = \frac{n_{d,k} + \alpha_j}{\sum_{i=1}^K (n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^V (v_{k,j} + \beta_j)}$$

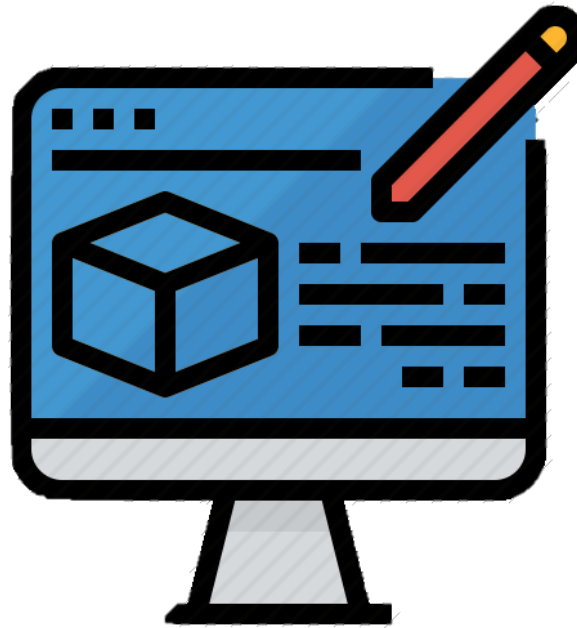
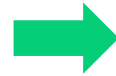
여기서 $n_{d,k}$ 는 k 번째 토픽에 할당된 d 번째 문서의 단어 빈도이며, $v_{k,w_{d,n}}$ 는 전체 말뭉치에서 k 번째 토픽에 할당된 단어 $w_{d,n}$ 의 빈도를 나타낸다. 여기서 $w_{d,n}$ 은 d 번째 문서에 나타난 n 번째 단어를 의미한다. K 는 사용자가 지정하는 토픽의 수이며 V 는 말뭉치에 등장하는 전체 단어 수이다.

3 데이터 분석

1 LDA



고객이 상품을 구매



상품 후기를 작성



Topic

구두, 뽕족함, 예쁨,
높은 굽, 편함

가방, 무거움,
가성비, 깔끔함

옷자락, 섬세함,
치수, 후드티, 멋짐

LDA 분석

3 데이터 분석

1 LDA

크라우드펀딩의 **성공사례**에 나타난 펀딩 주제

주제(토픽)	단어
애니메이션의 장면연출	animation, build, motion, artwork, trailer
정치, 심리 영화	politic, justice, mental, psychology, interview
어려움을 극복하는 영화	perspective, achieve, struggle, child, senior
대중음악과 관련된 영화	band, rock, song, club, guitar
영화캐스팅 관련 펀딩	cast, theater, degree, graduated, academy
청소년 문화를 다룬 영화	youth, communicate, culture, research, climate
DVD 제작	dvd, exclusive, awesome, signed, episode

크라우드펀딩의 **실패사례**에 나타난 펀딩 주제

주제(토픽)	단어
학생이 제작하는 영화	art, editor, director, student, college
인종차별, 인종역사를 다룬 영화	black, women, men, culture, history
어린이, 학생들의 건강교육과 관련된 영화	education, student, health, children, public
공포 영화	action, horror, zombie, heart, dead
가족 영화	kid, mother, father, fun, song
해학적인 애니메이션의 녹음	animation, voice, comic, clown, record
보상 관련 내용이 주로 명시된 펀딩	receive, pledge, reward, credit, dvd

3 데이터 분석

2 LDA 분석 결과 - 성공사례와 실패사례



애니메이션



장면연출



펀딩 성공



해학적인 녹음



펀딩 실패

3 데이터 분석

2 LDA 분석 결과 - 성공사례와 실패사례



학생, 청소년



문화



펀딩 성공



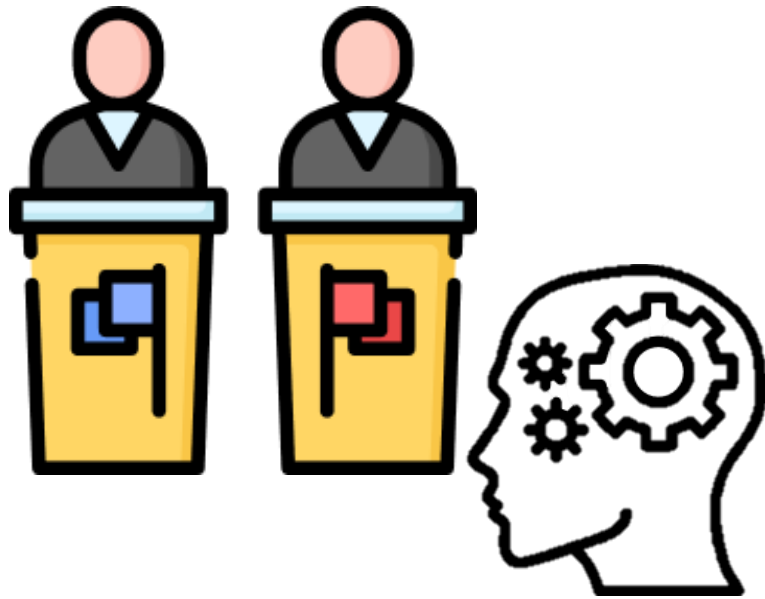
건강 교육



펀딩 실패

3 데이터 분석

3 LDA 분석 결과 - 성공사례



정치, 심리 영화를 다룬 편딩이나 등장인물들이 어려움을 극복하는 영화와 같이
스토리에 긴장감을 주는 영화 편딩은 성공하는 것으로 나타남.

3 데이터 분석

3 LDA 분석 결과 - 성공사례



영화 캐스팅과 같이 실력있는 배우를 캐스팅하거나 대중음악을 주제로 영화 제작을 위한 펀딩은 성공하는 것으로 나타남.

3 데이터 분석

3 LDA 분석 결과 - 성공사례



가치있는 DVD를 제작을 주제로 다룬 편딩은 성공하는 것으로 나타남.

3 데이터 분석

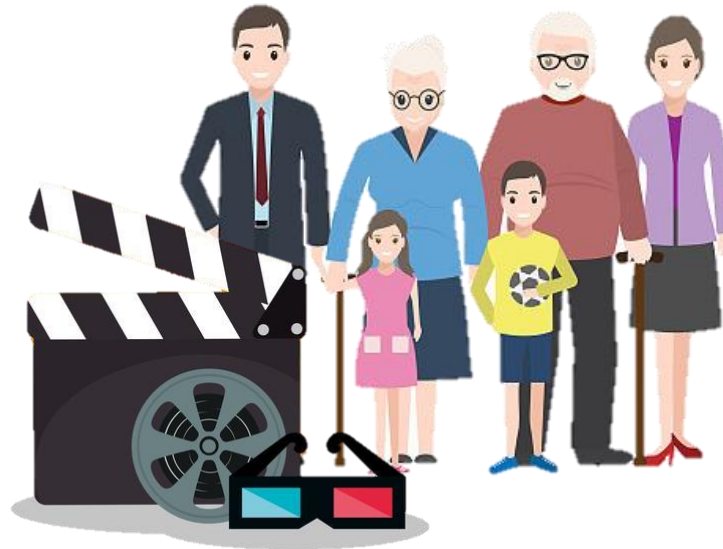
4 LDA 분석 결과 - 실패사례



학생이 제작하는 영화라고 밝힌 편딩은 실패하는 것으로 나타남.

3 데이터 분석

4 LDA 분석 결과 - 실패사례



가족 영화를 다룬 편딩은 실패하는 것으로 나타남.

3 데이터 분석

4 LDA 분석 결과 - 실패사례



인종역사나 인종차별과 같이 민감한 주제를 다루거나 호러 영화와 같이 후원자들의 호와 불호가 극심히 갈리는 주제를 다룬 펀딩은 실패하는 것으로 나타남.

3 데이터 분석

4 LDA 분석 결과 - 실패사례



후원금에 대한 보상관련 내용을 주로 적은 펀딩은 실패하는 것으로 나타남.

3 데이터 분석

5 인공지능 기법을 이용한 펀딩 성공과 실패 예측

소개글

00. 안녕하세요,

저희는 가천대학교에서 단편영화를 만드는 학생들입니다!

심혈을 기울여 준비한 영화 <살인마 숨기기>는 현재 프리프로덕션 막바지 단계로, 메인 프로덕션을 코앞에 두고 있습니다.

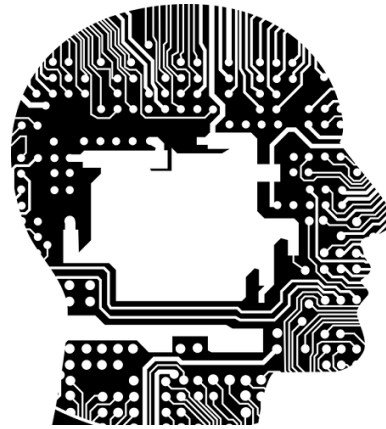
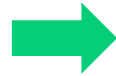
그 과정에서 여러분의 도움을 얻고자 이렇게 글을 쓰게 되었습니다.

<살인마 숨기기>는 대한민국 사회 전반에 뿌리내린 '비난'에 대한 의문에서 시작했습니다. 왜 우리는 비난하지 않아도 될 일로 상대방을 비난하고, 비난받아야 하는 걸까요? 그런 풍토를 욕하면서도 왜 우리는 다시 서로를 비난하게 되는 걸까요?

저희는 '보통에 미치지 못하는 것'이 그 이유가 되는 것에 특히 주목했습니다. 신체적, 정신적 장애, 지력, 심지어는 남들보다 취직을 늦게 했다는 것까지, 보통을 따라가지 못하는 것이 비난의 이유가 되곤 합니다. 하지만 그게 본인의 책임일 순 있어도 남에게 욕먹을 만한 일은 아니잖아요?

주인공 영서는 살인마의 딸입니다. 아빠의 살인이 영서의 죄는 아니지만, 살인마의 딸이라는 낙인이 두려웠던 영서는 아빠의 범죄를 숨기는 데에 집착하게 됩니다. 영서라는 캐릭터는 자

펀딩 소개글



인공지능



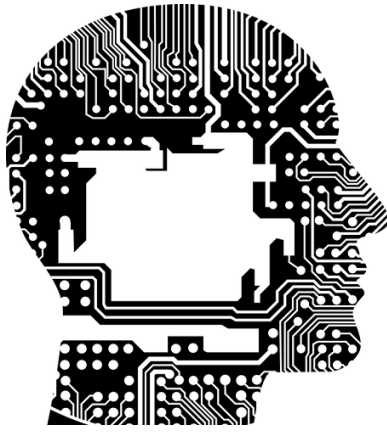
펀딩 성공



펀딩 실패

3 데이터 분석

5 인공지능 기법을 이용한 펀딩 성공과 실패 예측



인공지능



CNN(Convolution Neural Network)



RNN(Recurrent Neural Network)

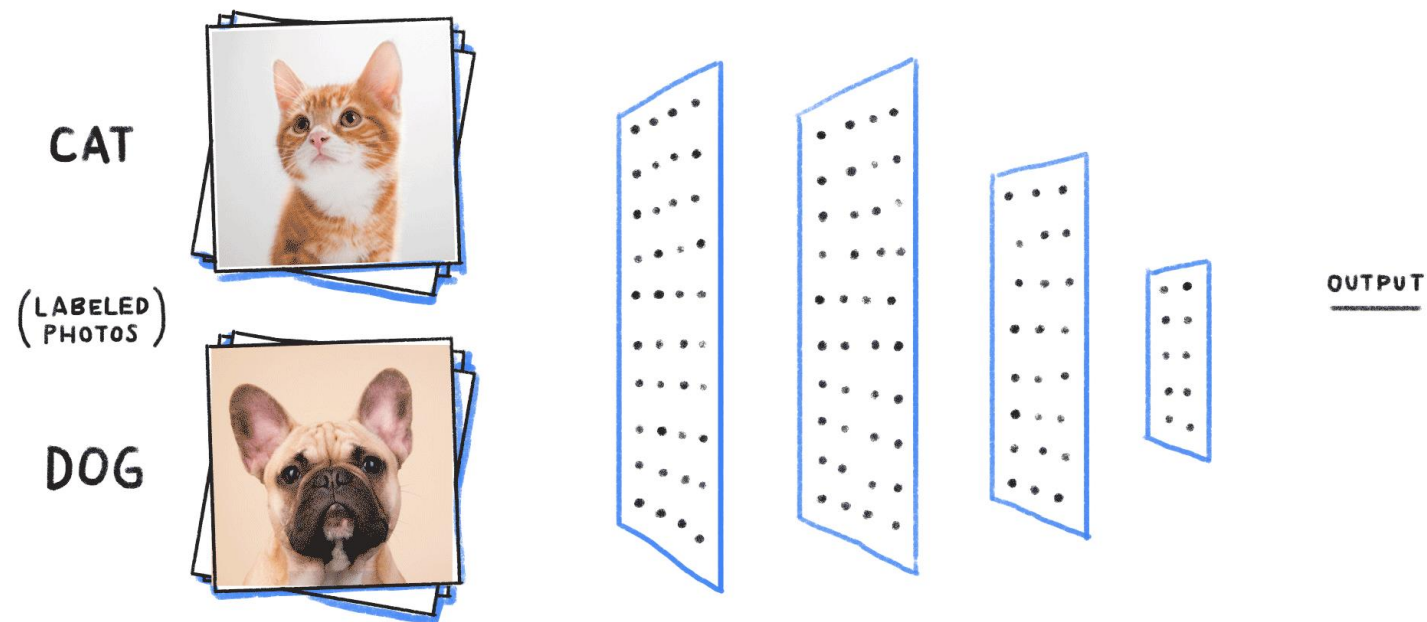


LSTM(Long Short-Term Memory models)

3 데이터 분석

6 CNN(Convolution Neural Network)

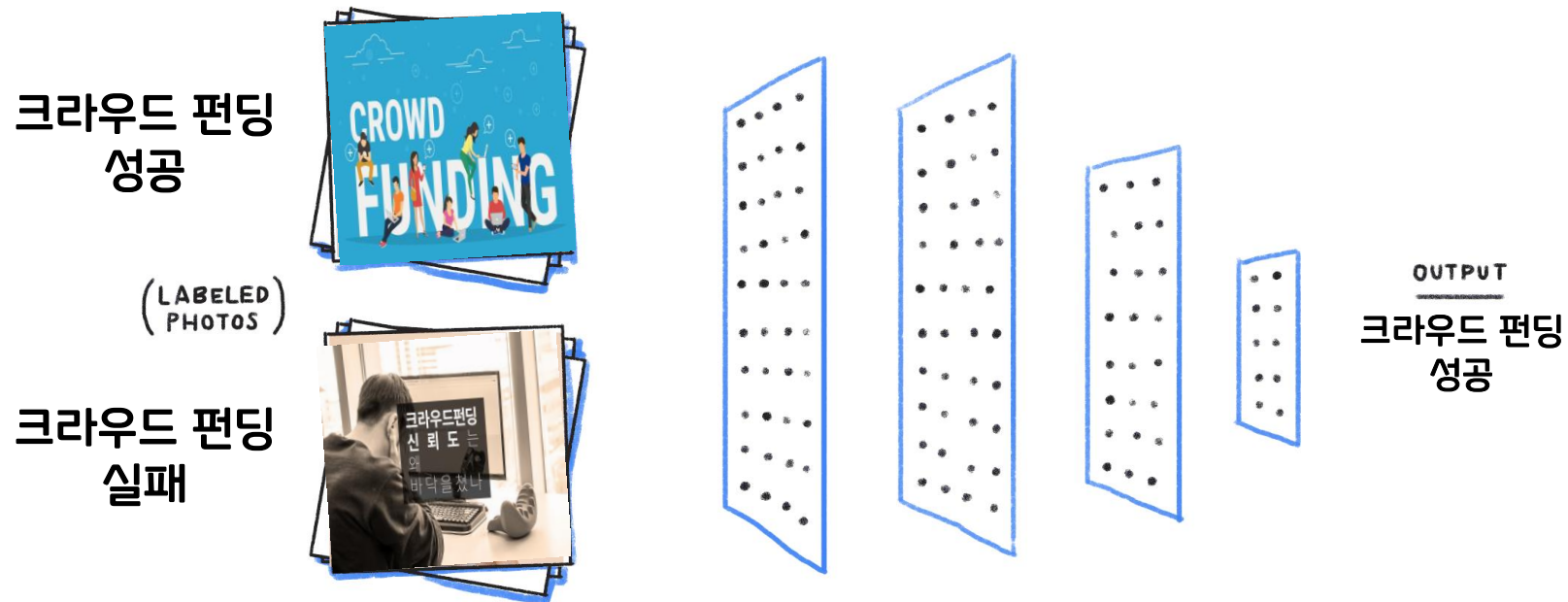
CNN : 주로 이미지에서 특징을 추출하여 이미지를 식별하는데 사용되는 딥러닝 기법.



3 데이터 분석

6 CNN(Convolution Neural Network)

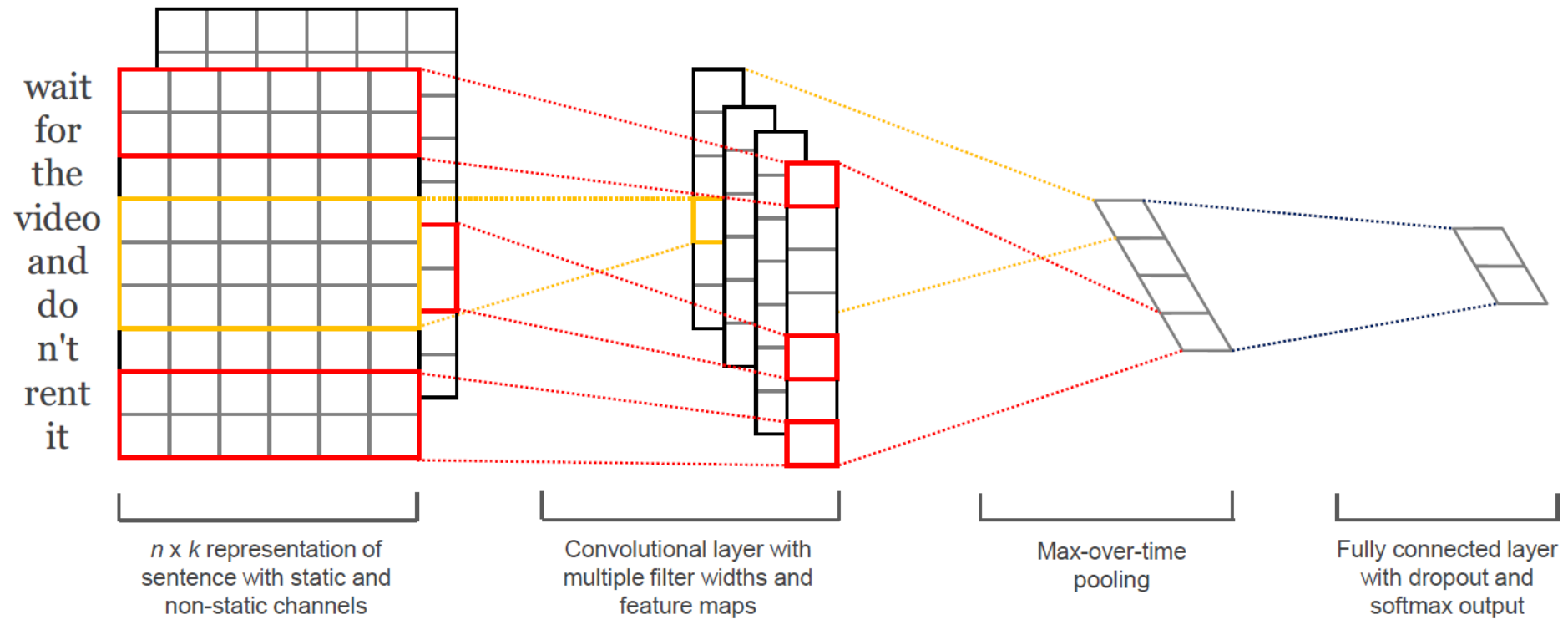
CNN : 텍스트가 가진 정보에서 특징을 추출하여 크라우드 펀딩 성공 예측에 사용함.



3 데이터 분석

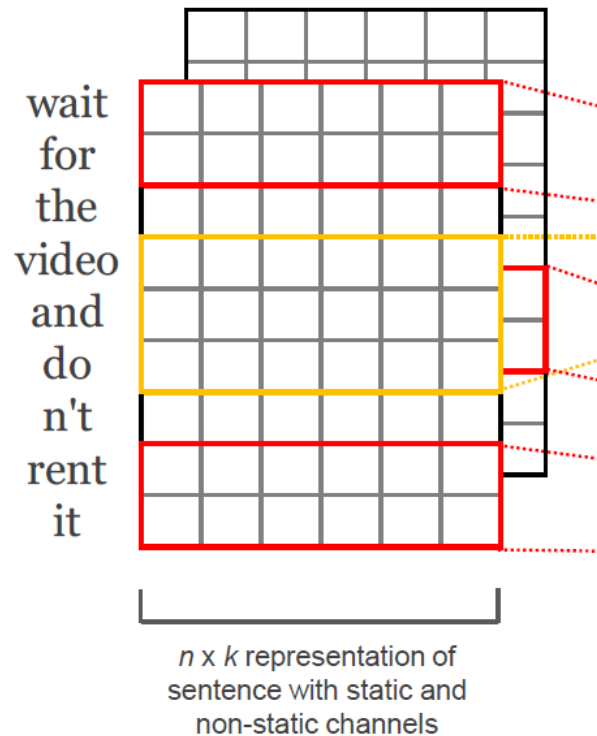
6 CNN(Convolution Neural Network)

Text-CNN의 흐름도



3 데이터 분석

6 CNN(Convolution Neural Network)



CNN 흐름도 중 일부

1. Kernal size

그림의 Kernal size는 2,3으로 이루어져 있으며 이는 bi-gram, tri-gram으로 학습을 진행한다는 뜻.

> 본 연구에서는 Kernal size를 1,2,3,4,5로 구성.

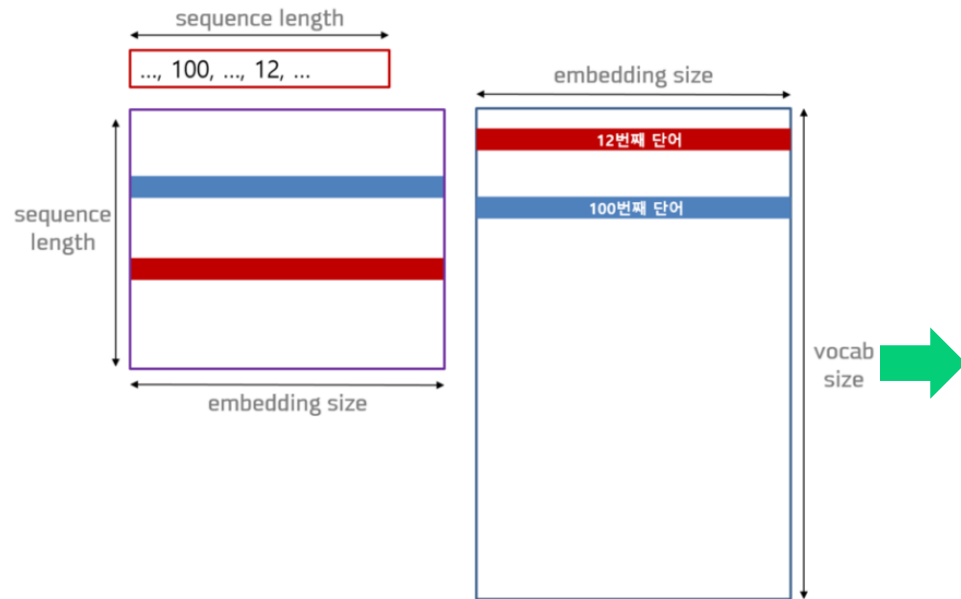
2. Channel size

그림의 Channel size는 2개로 이루어져 있으며 채널은 각기 다른 차원에서 학습을 하기위해 존재.

> 본 연구에서는 channel size를 100개로 구성.

3 데이터 분석

6 CNN(Convolution Neural Network)



Look up 테이블

3. Embedding size

사용자가 지정한 단어벡터 차원수를 나타내며
하나의 단어를 몇 개의 차원으로 보는지 설정.

> 본 연구에서는 embedding size를 30으로 구성.

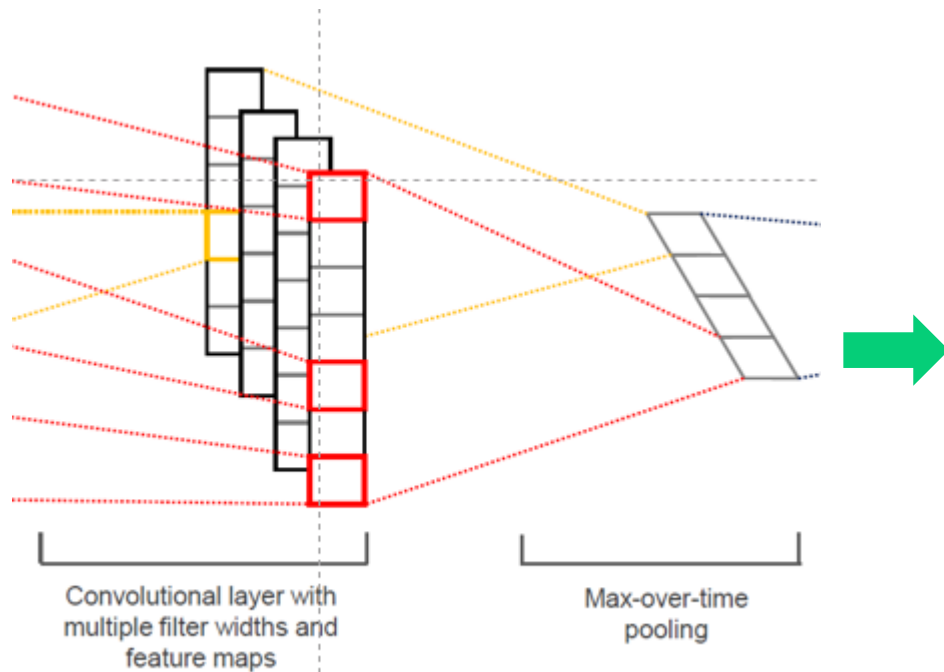
4. vocab size

데이터에 존재하는 단어의 종류 수를 나타내며
모델 학습 전 사용자가 단어를 지정.

> 본 연구에서는 단어 종류가 67,450개로 존재.

3 데이터 분석

6 CNN(Convolution Neural Network)



CNN 흐름도 중 일부

5. Pooling layer

Pooling layer는 convolution layer의 출력 데이터를 입력으로 받아서 출력 데이터의 크기를 줄이거나 특정 데이터를 강조하는 용도로 사용.

> 본 연구에서는 Max-pooling을 사용.

Single depth slice

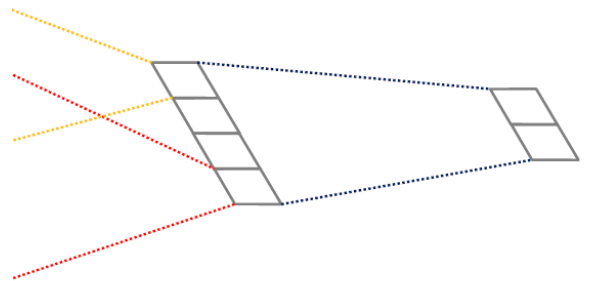
1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

max pool with 2x2 filters and stride 2

6	8
3	4

3 데이터 분석

6 CNN(Convolution Neural Network)

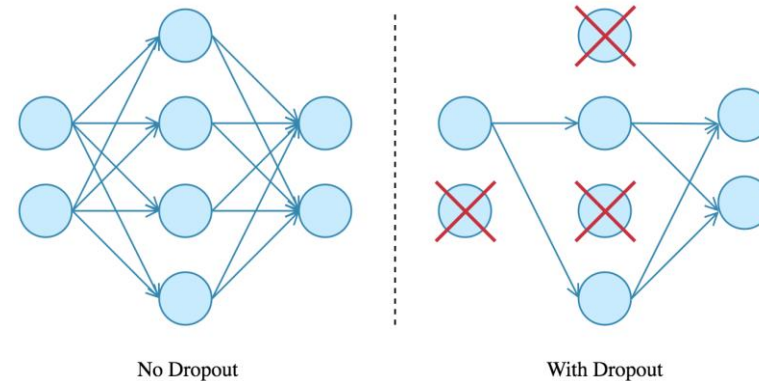


CNN 흐름도 중 일부

6. Dropout layer

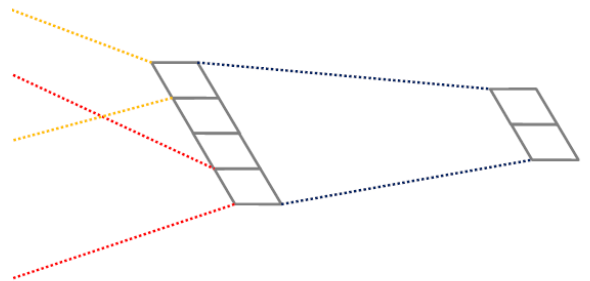
Dropout layer는 모델이 과적합 되지 않도록 계층 내 뉴런의 일부를 비활성화 시키기 위한 layer

> 본 연구에서는 Dropout 비율을 0.5로 설정



3 데이터 분석

6 CNN(Convolution Neural Network)



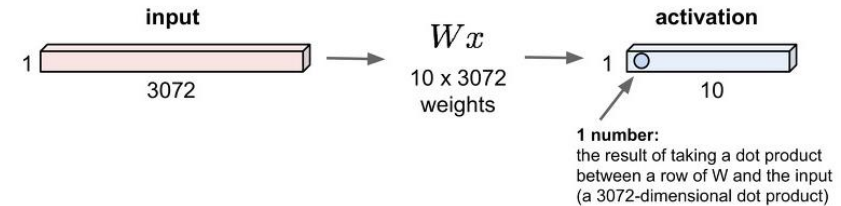
CNN 흐름도 중 일부

7. Fully connected layer

Fully connected layer는 max-pooling과정 및 dropout layer를 거쳐 나온 데이터(input)에 적당한 가중치(w)를 곱하여 클래스 갯수(성공,실패 2개) 만큼 output을 생성

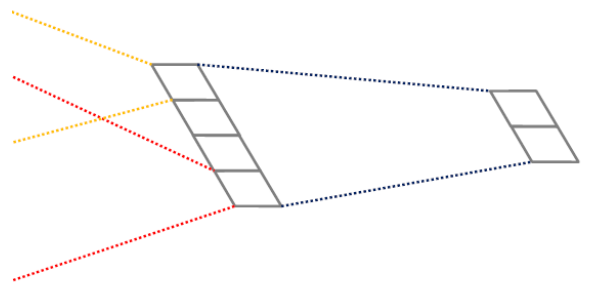
Fully Connected Layer

32x32x3 image -> stretch to 3072 x 1



3 데이터 분석

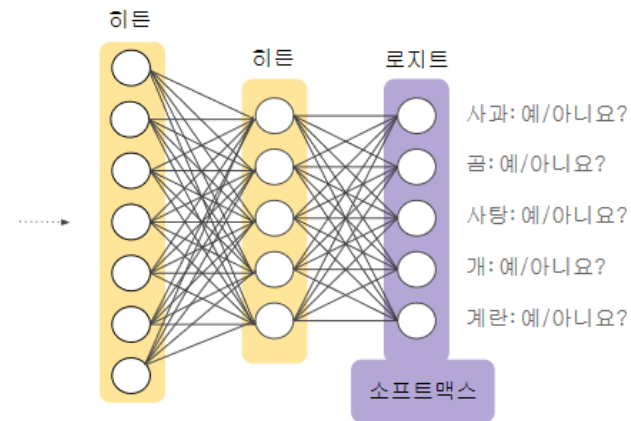
6 CNN(Convolution Neural Network)



CNN 흐름도 중 일부

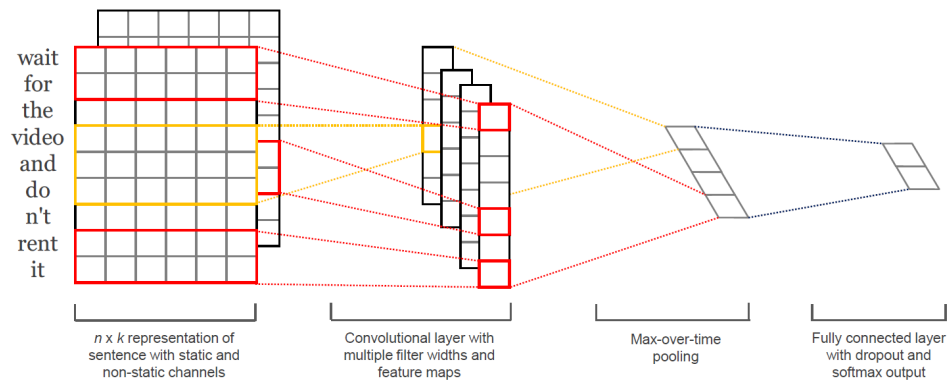
8. Softmax layer

Softmax layer는 최종적인 결과물이 종속변수 중 어디에 속하는지 분류하기 위한 layer



3 데이터 분석

6 CNN(Convolution Neural Network)



9. 기타 세부옵션

optimizer : rmsprop

epoch : 15

batch size : 50

learning rate : 0.0005

10. 정확도

총 학습에 걸린 시간 : 61분 52초

Training Test 분할 비율 : 7 대 3

Training data : 89.268

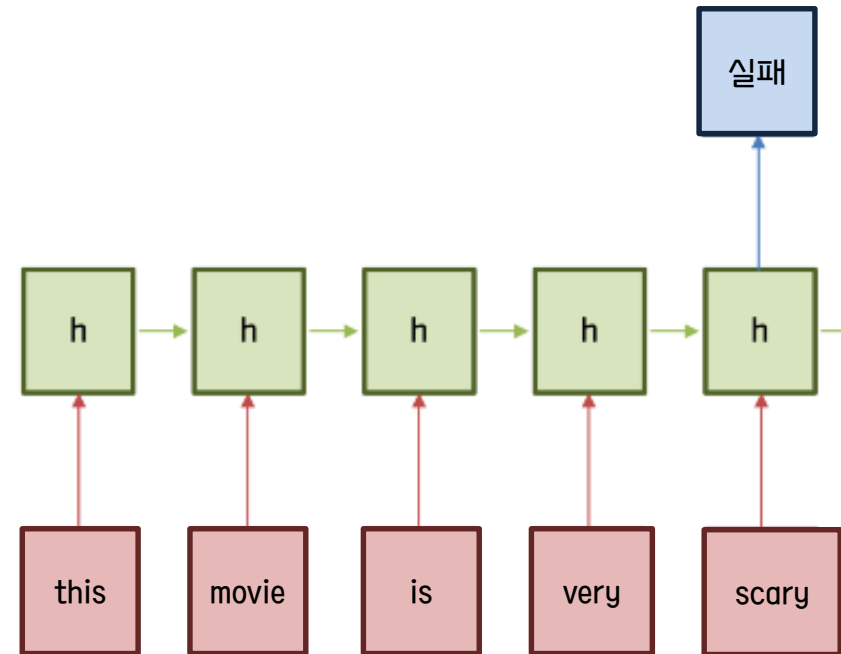
Testing data : 71.467

3 데이터 분석

7 RNN(Recurrent Neural Network)

RNN : 텍스트가 가진 정보의 시퀀스를 고려하여 크라우드 펀딩 성공 예측에 사용함.

“this movie is very scary”

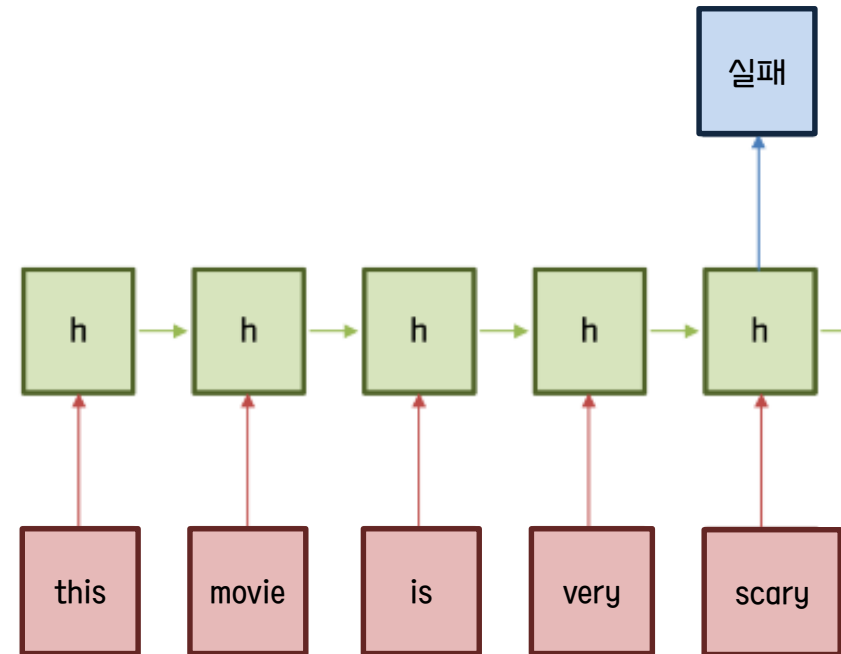


3 데이터 분석

7 RNN(Recurrent Neural Network)

RNN : 텍스트가 가진 정보의 시퀀스를 고려하여 크라우드 펀딩 성공 예측에 사용함.

“this movie is very scary”

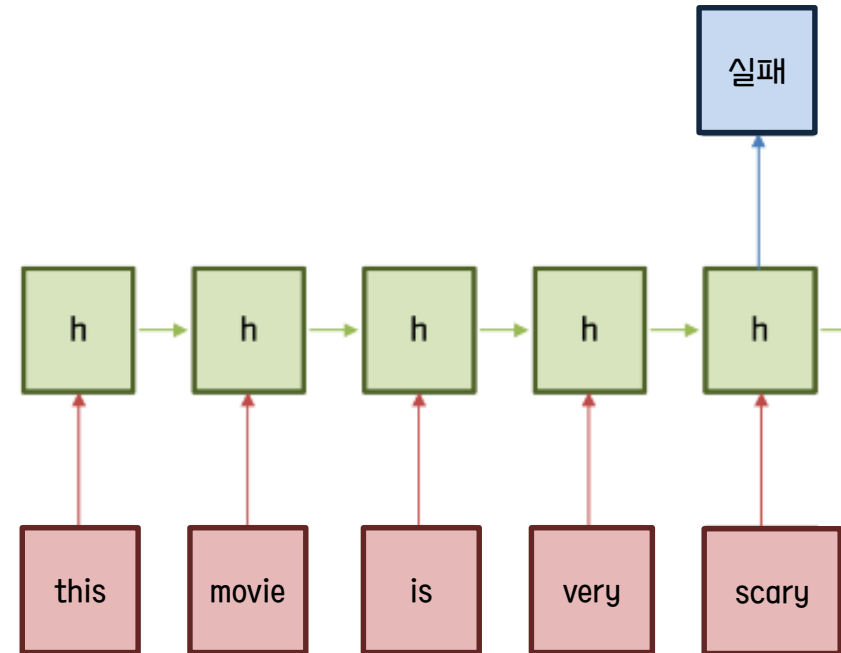


3 데이터 분석

5 RNN(Recurrent Neural Network)

RNN : 텍스트가 가진 정보의 시퀀스를 고려하여 크라우드 펀딩 성공 예측에 사용함.

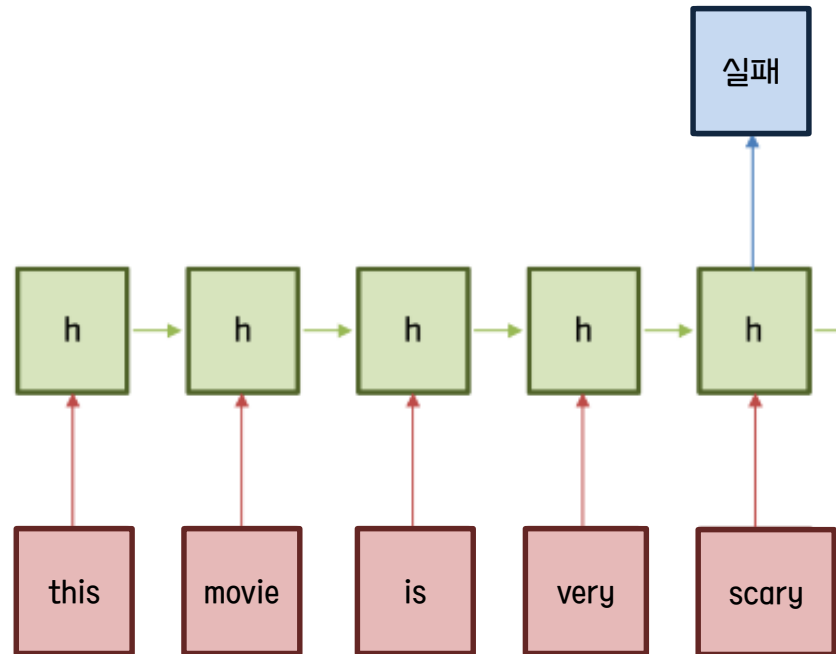
“this movie is very scary”



3 데이터 분석

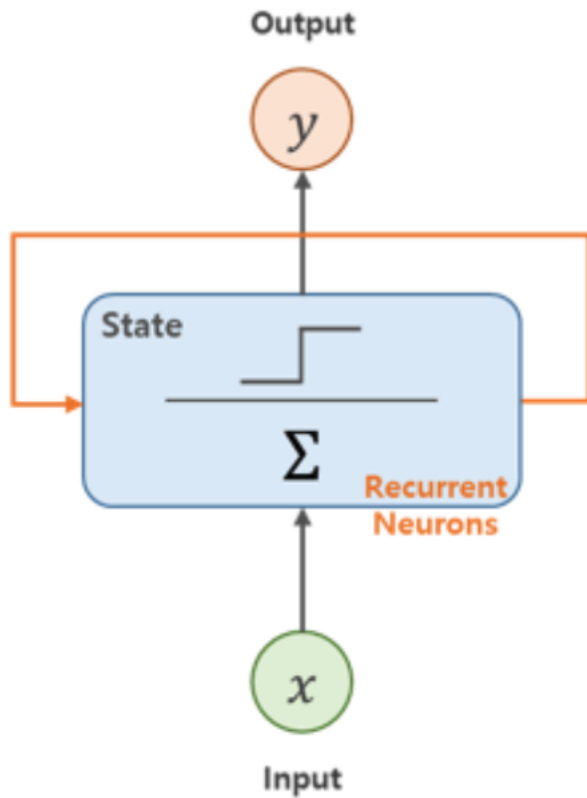
5 RNN(Recurrent Neural Network)

RNN의 전체 흐름도



3 데이터 분석

5 RNN(Recurrent Neural Network)

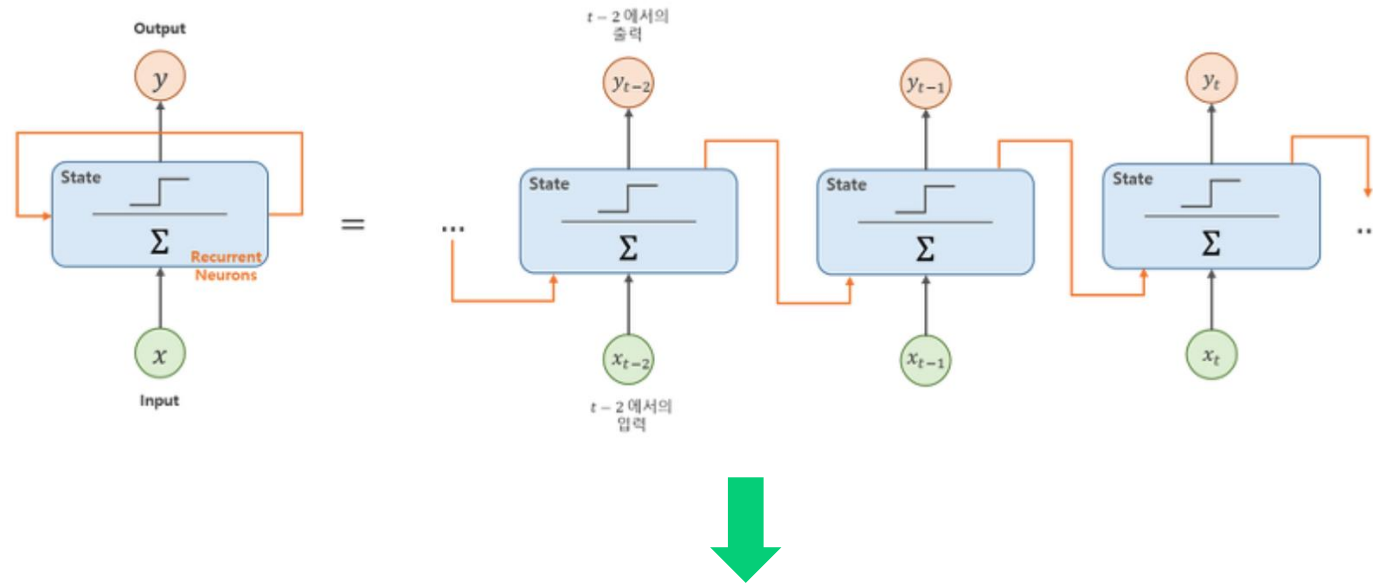


RNN은 일반적인 신경망과 비슷하지만, 출력이 다시 입력으로 받는 부분이 있다.

RNN은 입력(x)을 받아 출력(y)를 만들고, 이 출력을 다시 입력으로 받는다.

3 데이터 분석

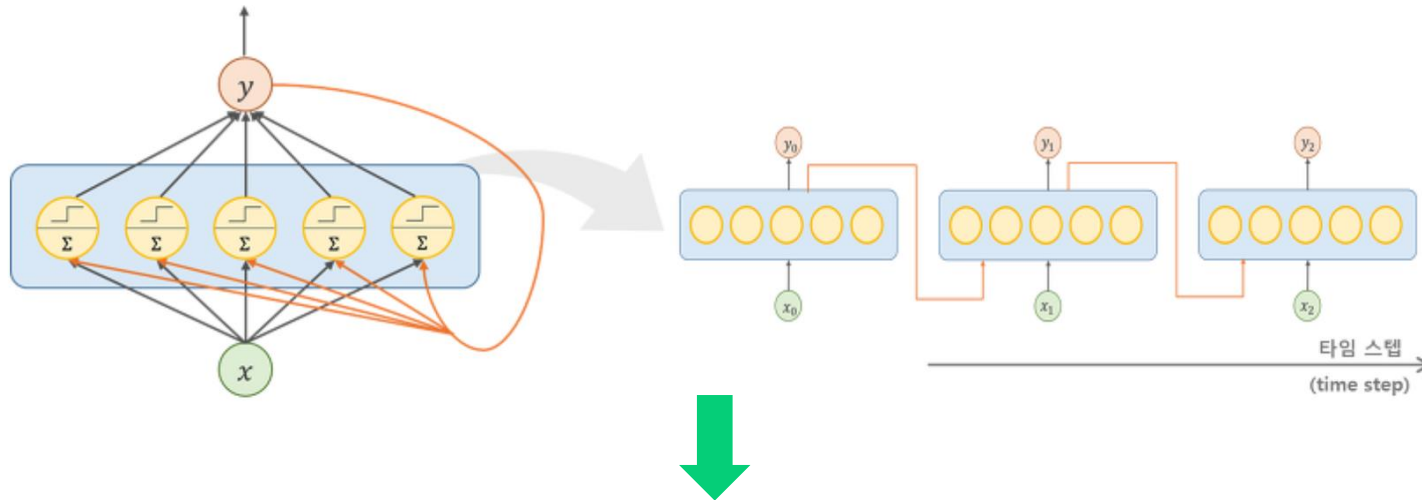
5 RNN(Recurrent Neural Network)



일반적으로 RNN을 그림으로 나타낼 때는 각 타임 스텝(time step) t 마다 순환 뉴런을 펼쳐서 타임스텝별 입력(X_t)과 출력(Y_t)을 나타낸다.

3 데이터 분석

5 RNN(Recurrent Neural Network)

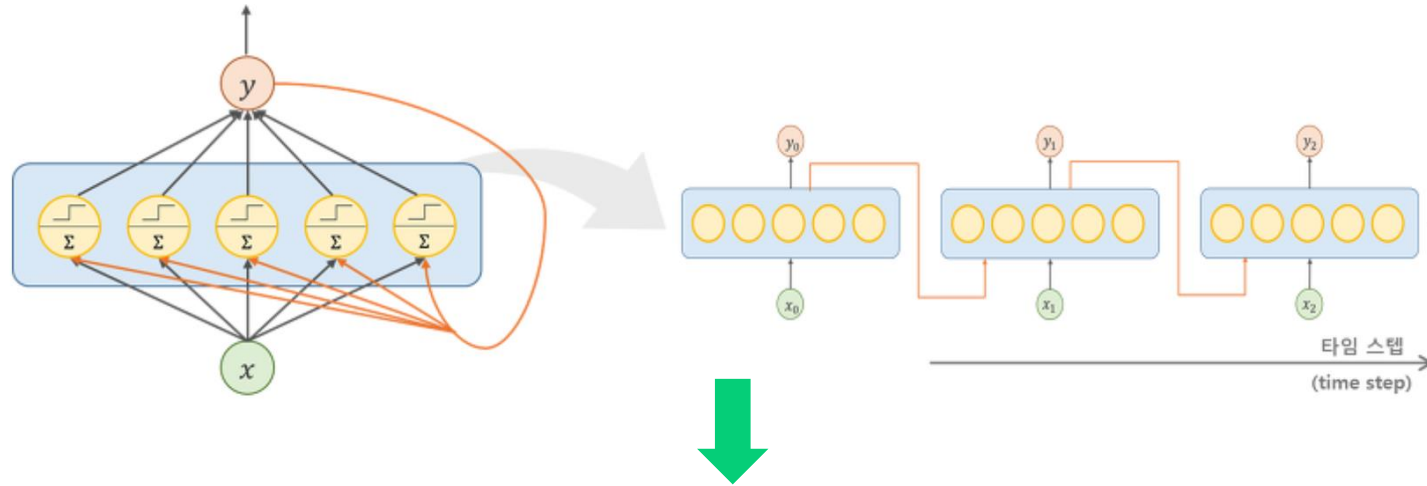


순환 뉴런으로 구성된 층(layer)은 타임 스텝 t 마다 모든 뉴런은 입력 벡터 X_t 와 이전 타임 스텝의 출력 벡터 Y_{t-1} 을 입력 받는다.

각 순환 뉴런은 두 개의 가중치 W_x 와 W_y 를 가지는데, W_x 는 X_t 를 위한 것이고 W_y 는 이전 타임 스텝의 출력 Y_{t-1} 을 위한 것이다. 이것을 순환 층(layer) 전체로 생각하면 가중치 벡터 W_t 와 W_y 를 행렬 W_x 와 W_y 로 나타낼 수 있으며 다음의 식과 같이 표현할 수 있다.

3 데이터 분석

5 RNN(Recurrent Neural Network)



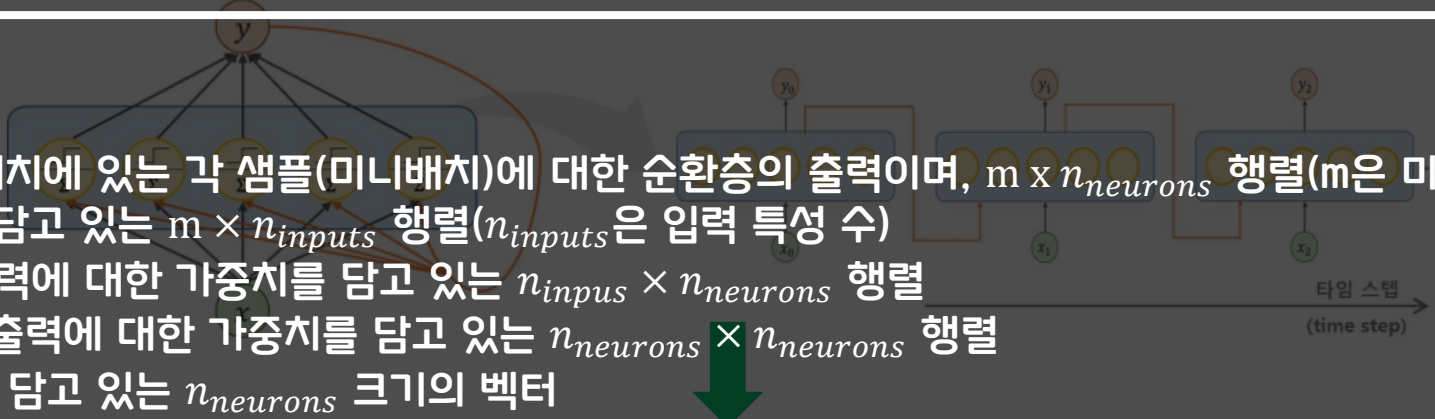
그리고 타임 스텝 t 에서의 미니배치(mini-batch)의 입력을 행렬 X_t 로 나타내어 순환층의 출력을 한번에 계산할 수 있다.

3 데이터 분석

5 RNN(Recurrent Neural Network)

Y_t 는 X_t 와 Y_{t-1} 의 함수이므로, 타임 스텝 $t=0$ 에서부터 모든 입력에 대한 함수가 된다. 첫 번째 타임 스텝인 $t=0$ 에서는 이전의 출력이 없기 때문에 일반적으로 0으로 초기화 된다.

Y_t : 타임 스텝 t 에서 미니배치에 있는 각 샘플(미니배치)에 대한 순환층의 출력이며, $m \times n_{neurons}$ 행렬(m 은 미니배치, $n_{neurons}$ 은 뉴런 수)
 X_t : 모든 샘플의 입력값을 담고 있는 $m \times n_{inputs}$ 행렬(n_{inputs} 은 입력 특성 수)
 W_x : 현재 타임 스텝 t 의 입력에 대한 가중치를 담고 있는 $n_{inputs} \times n_{neurons}$ 행렬
 W_y : 이전 타임 스텝 $t-1$ 의 출력에 대한 가중치를 담고 있는 $n_{neurons} \times n_{neurons}$ 행렬
 b : 각 뉴런의 편향(bias)을 담고 있는 $n_{neurons}$ 크기의 벡터



그리고 타임 스텝 t 에서의 미니배치(mini-batch)의 입력을 행렬 X_t 로 나타내어 순환층의 출력을 한번에 계산할 수 있다.

$$\begin{aligned} Y_t &= \phi(X_t \cdot W_x + Y_{t-1} \cdot W_y + b) \\ &= \phi([X_t \ Y_{t-1}] \begin{bmatrix} W_x \\ W_y \end{bmatrix} + b) \end{aligned}$$

3 데이터 분석

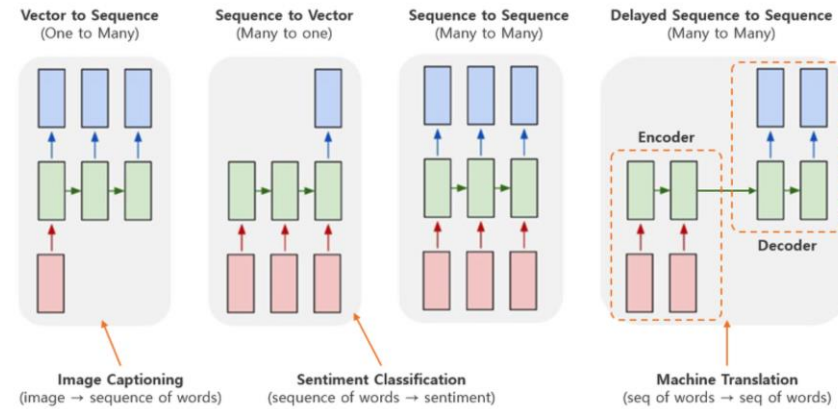
5 RNN(Recurrent Neural Network)

타입 스텝 t 에서 순환 뉴런의 출력은 이전 타입 스텝의 모든 입력에 대한 함수이기 때문에 이것을 메모리라고 볼 수 있다. 이렇게 타입 스텝에 걸쳐 어떠한 상태를 보존하는 신경망의 구성 요소를 메모리 셀(memory cell) 또는 셀(cell)이라고 한다. 일반적으로 타입 스텝 t 에서 셀의 상태 H_t (H = hidden)는 아래의 식과 같이 타입 스텝에서의 입력과 이전 타입 스텝의 상태에 대한 함수이다.

$$h_t = f(h_{t-1}, X_t)$$

3 데이터 분석

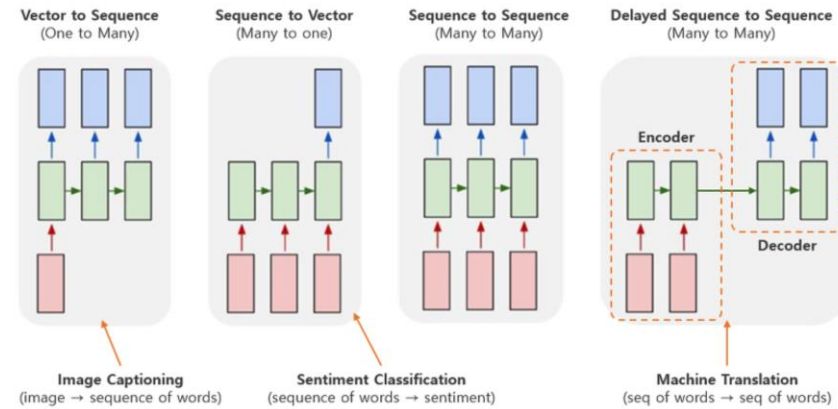
5 RNN(Recurrent Neural Network)



RNN은 그림과 같이 다양한 입력 시퀀스를 받아 출력 시퀀스를 만들 수 있다. 위의 그림에서 Vector-to-Sequence는 첫 번째 타임 스텝에서 하나의 입력만(다른 모든 타임 스텝에서는 0)을 입력 받아 시퀀스를 출력하는 네트워크이며, 이러한 모델은 Image Captioning에 사용할 수 있다. Sequence-to-Vector는 Vector-to-Sequence와의 반대로 입력으로 시퀀스를 받아 하나의 벡터를 출력하는 네트워크로, Sentiment Classification에 사용할 수 있다.

3 데이터 분석

5 RNN(Recurrent Neural Network)



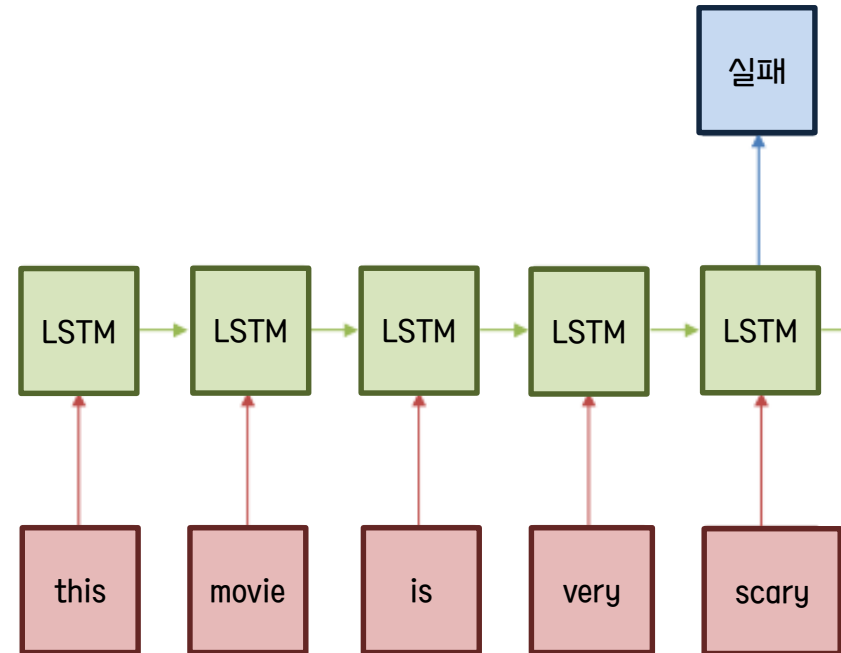
위의 그림의 오른쪽에서 세 번째 Sequence-to-Sequence는 시퀀스를 입력받아 시퀀스를 출력하는 네트워크이며, 주식가격과 같은 시계열 데이터를 예측하는 데 사용할 수 있다. 마지막으로 Delayed Sequence-to-Sequence는 인코더(encoder)에는 seq-to-vec 네트워크를 디코더(decoder)에는 vec-to-seq 네트워크를 연결하는 것으로, 기계 번역에 사용된다.

3 데이터 분석

8 LSTM(Long Short-Term Memory models)

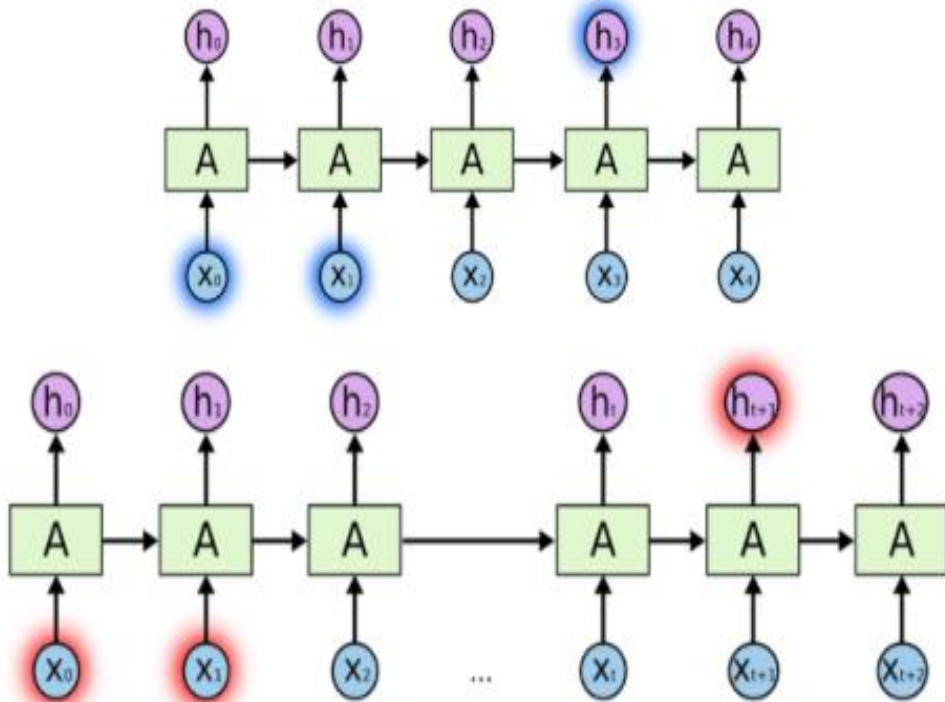
LSTM : RNN에서 발생하는 기울기 소실문제를 방지하기 위해 다음 층으로 값을 넘길지 넘기지 않을지 관리하는 단계를 기존 RNN에 추가한 개선된 모형.

“this movie is very scary”



3 데이터 분석

8 LSTM(Long Short-Term Memory models)



〈관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 RNN학습 능력 저하〉

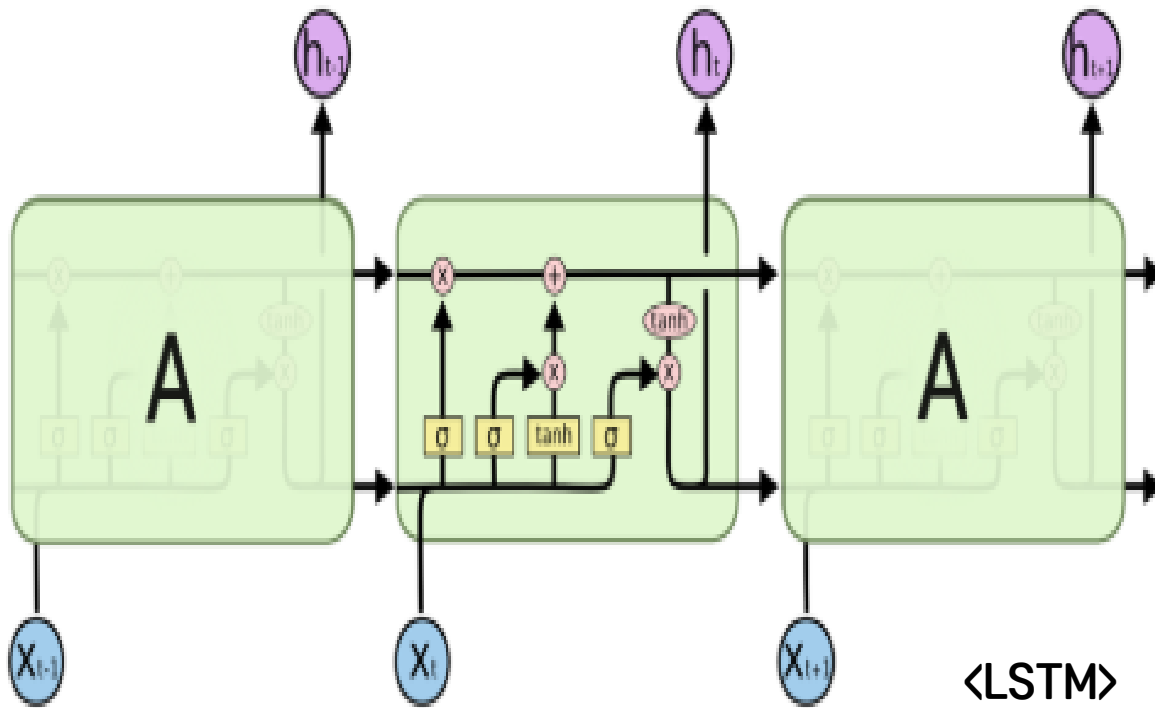
RNN은 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 역전파시에 그래디언트가 점차 줄어 학습 능력이 저하되는 것으로 알려져 있음

이를 Vanishing gradient problem이라고 하고, 이 문제를 극복하기 위해서 고안된 것이 LSTM 임

3 데이터 분석

8 LSTM(Long Short-Term Memory models)

LSTM의 흐름도



$$f_t = \sigma(W_{xh_f}x_t + W_{hh_f}h_{t-1} + b_{h_f})$$

$$i_t = \sigma(W_{xh_i}x_t + W_{hh_i}h_{t-1} + b_{h_i})$$

$$o_t = \sigma(W_{xh_o}x_t + W_{hh_o}h_{t-1} + b_{h_o})$$

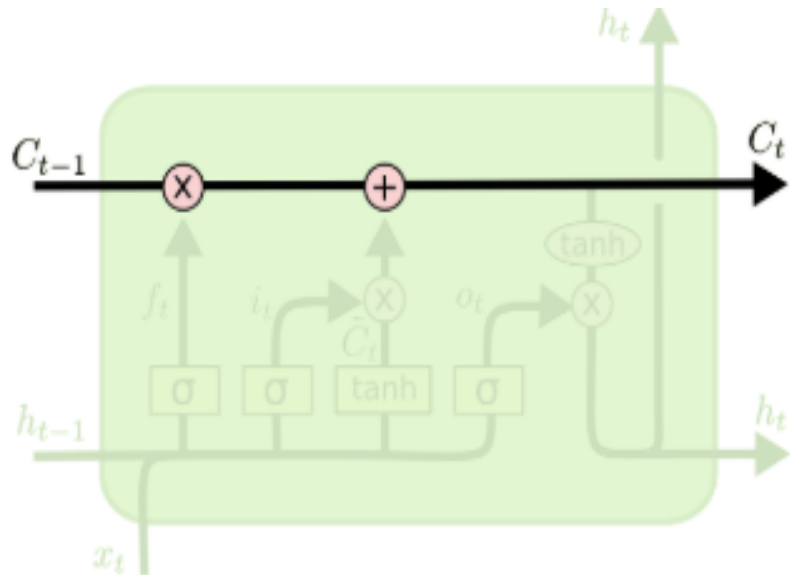
$$g_t = \tanh(W_{xh_g}x_t + W_{hh_g}h_{t-1} + b_{h_g})$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$h_t = o_t \odot \tanh(c_t)$$

3 데이터 분석

8 LSTM(Long Short-Term Memory models)



LSTM의 cell state

LSTM의 cell state

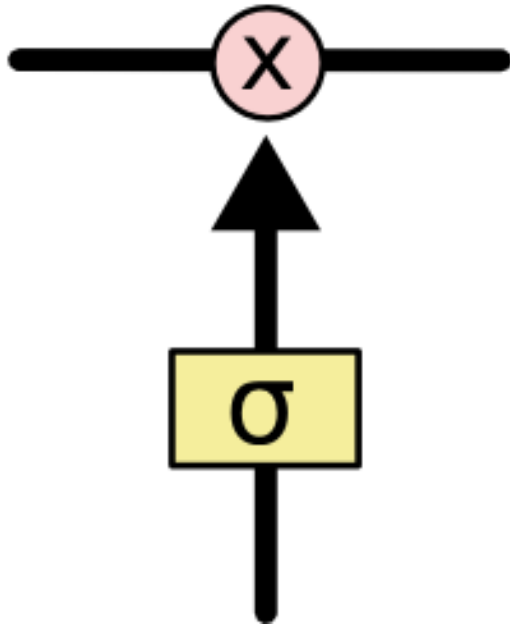
LSTM의 핵심은 cell state이며, 모듈 그림에서 수평으로 그려진 위의 선에 해당함

Cell state는 컨베이어 벨트와 같아서, 작은 linear interaction만을 적용시키면서 전체 체인을 계속 구동시킴

LSTM의 cell state에 정보를 더하거나 없앨 수 있는 능력이 있는데, 이 능력은 gate라고 불리는 구조에 의해 조심스럽게 제어됨

3 데이터 분석

8 LSTM(Long Short-Term Memory models)



LSTM의 첫 번째 gate

LSTM의 gate

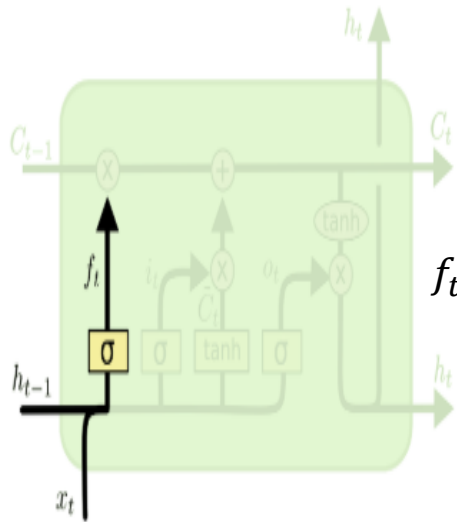
Gate는 정보가 전달될 수 있는 추가적인 방법으로, Sigmoid layer와 pointwise 곱셈으로 이루어져 있음

Sigmoid layer는 0과 1 사이의 숫자를 내보내며, 이 값은 각 컴포넌트가 얼마나 정보를 전달해야 하는지에 대한 척도를 나타냄

그 값이 0이라면 '아무 것도 넘기지 말아라'가 되고 값이 1이라면 '모든 것을 넘겨라'가 됨

3 데이터 분석

8 LSTM(Long Short-Term Memory models)



LSTM의 forget gate layer

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_t) \rightarrow$$

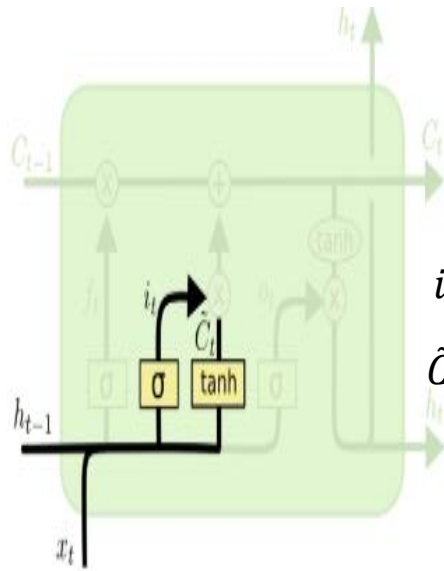
1. Forget gate layer

Cell state로부터 어떤 정보를 버릴 것인지를 정하는 것으로, sigmoid layer에 의해 결정됨

Forget gate layer라고 부름

3 데이터 분석

8 LSTM(Long Short-Term Memory models)



LSTM의 input gate layer

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$



2. Input gate layer

앞으로 들어오는 새로운 정보 중 어떤 것을
Cell state에 저장할 것인지를 정함

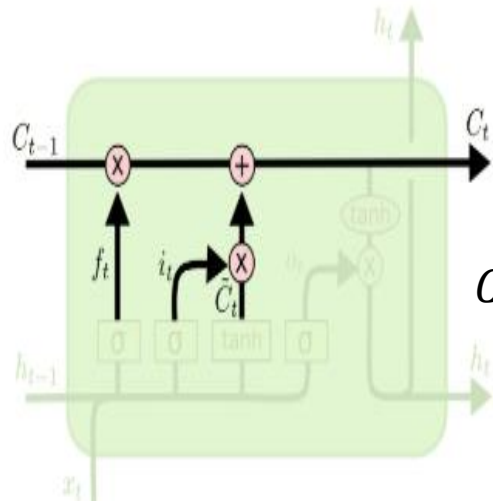
먼저 sigmoid layer가 어떤 값을 업데이트할 지
정함

다음 tanh layer가 새로운 후보 값 벡터를 만들고,
Cell state에 더할 준비를 함

이 두 단계에서 나온 정보를 합쳐서 state에 업데이트
할 재료를 만듦

3 데이터 분석

8 LSTM(Long Short-Term Memory models)



LSTM의 cell state 업데이트

$$C_t = f_t * C_{t-1} + i_t * \check{C}_t$$



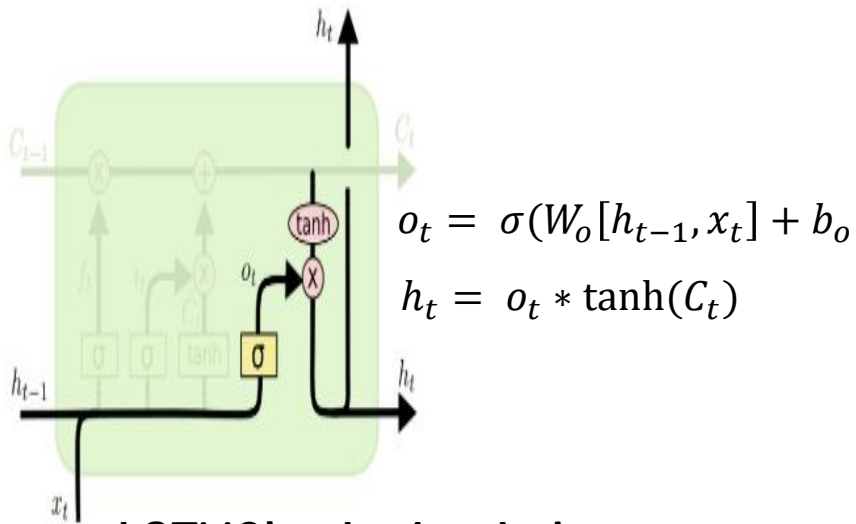
3. Cell state update

과거 state를 업데이트해서 새로운 cell state를 만듦

이미 이전 단계에서 어떤 값을 얼마나 업데이트해야 할지 정했기에 여기서는 그 일을 실천하는 단계

3 데이터 분석

8 LSTM(Long Short-Term Memory models)



LSTM의 output gate layer

4. Output gate layer

무엇을 output으로 내보낼지 정하는 단계

먼저, sigmoid layer에 input데이터를 넣어서 cell state의 어느 부분을 output으로 내보낼지를 정함

그리고 cell state를 tanh layer에 넣어 -1과 1사이의 값을 받은 뒤 방금 계산한 sigmoid gate의 output과 곱함

그렇게 되면 output으로 보내고자 하는 부분만 내보낼 수 있음

3 데이터 분석

9 정확도 비교

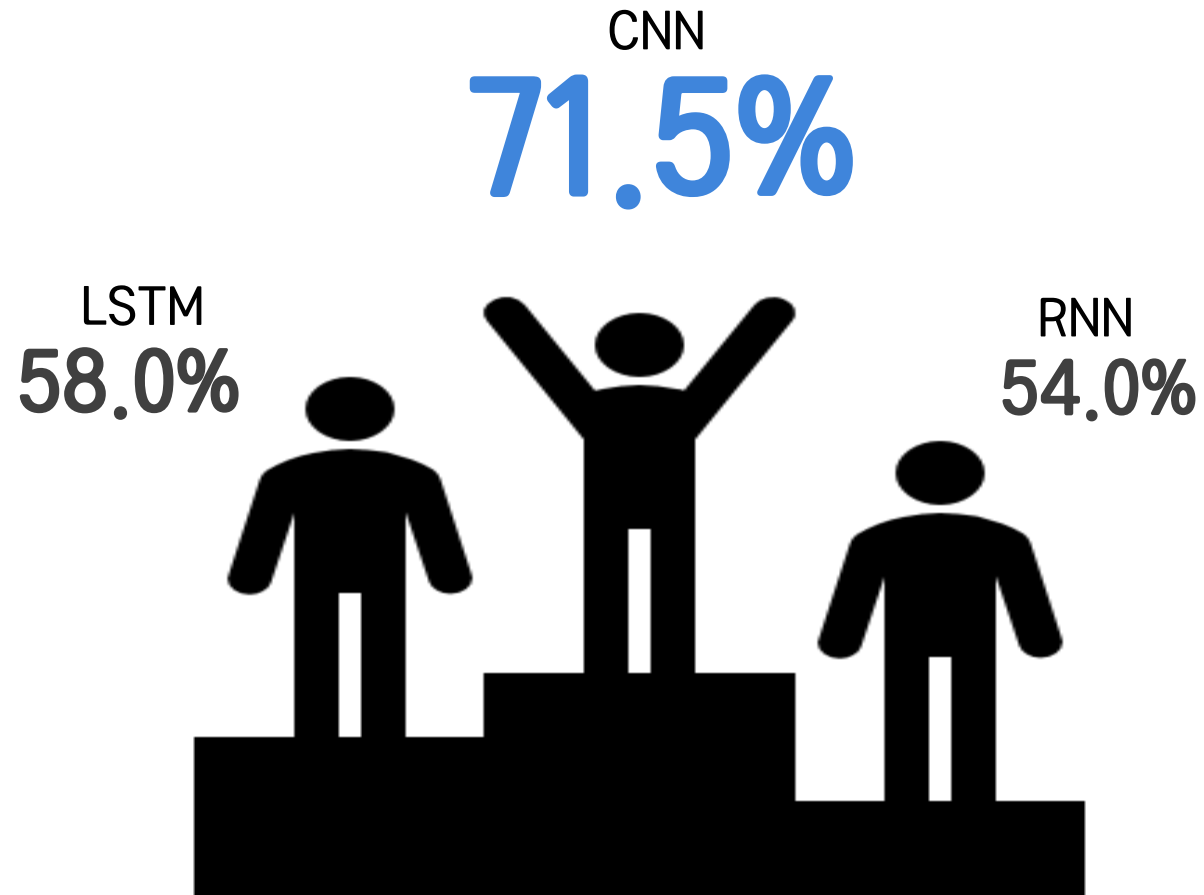


이 때, 훈련 데이터와 평가 데이터는 7:3의 비율로 랜덤하게 분할.

$$\text{※ 정확도} = \frac{\text{성공 또는 실패로 맞춘 사례}}{\text{모든 크라우드펀딩 사례}} \times 100(\%)$$

3 데이터 분석

9 정확도 비교



각 정확도는 평가 데이터 기준의 정확도.

4 결론

1 LDA 분석을 통한 결론



인종역사나 인종차별과 같이 민감한 주제를 다루거나 호러 영화와 같이 후원자들의 호와 불호가 극심히 갈리는 주제를 다룬 펀딩이 실패하는 것으로 나타남.

4 결 론

1 LDA 분석을 통한 결론



청소년 문화나 대중음악과 같이 대중적이고 쉽게 접할 수 있는 주제를
다룬 펀딩이 성공하는 것으로 나타남.

4 결론

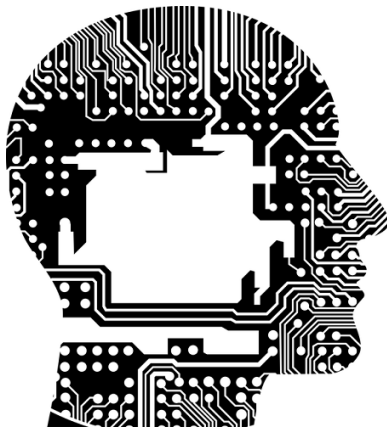
1 LDA 분석을 통한 결론



영화예술분야에서 펀딩을 성공시키려면 후원자들이 이해하기 쉽고 대중적으로 많은 공감을 불러 일으킬 수 있는 주제로 펀딩의 소개글을 구성하는 것이 좋을 것으로 판단됨.

4 결론

2 인공지능 기법을 이용한 펀딩 성공과 실패 예측에 대한 결론



CNN 모델

- CNN 모델의 크라우드펀딩 성공과 실패에 대한 판별 정확도는
Training data : 89.3%
Test data : 71.5%
로 나타나 어느정도 판별력이 있는 것으로 보임.
- 해당 분석 결과는 영화 크라우드 펀딩 소개글이 펀딩 성공
에 영향을 미친다는 것을 알려줌.

➤ 선행연구에서 사용한 정량적인 데이터에 텍스트 정보가 담긴 정성적 데이터를 추가하여 분석한다면 기존 모델의 성능을 더 높일 수 있을 것으로 기대됨.