WEBTOON
선호도
측정지표 개발

완두콩

홍만호 조희연 이정연

과학기술정보통신부 한국데이터산업진흥원

WEBTOON 선호도 측정지표개발

완두콩
홍만호 조희연 이정연

과학기술정보통신부 한국데이터산업진흥원

팀 소개

조희연

Project Leader

이정연

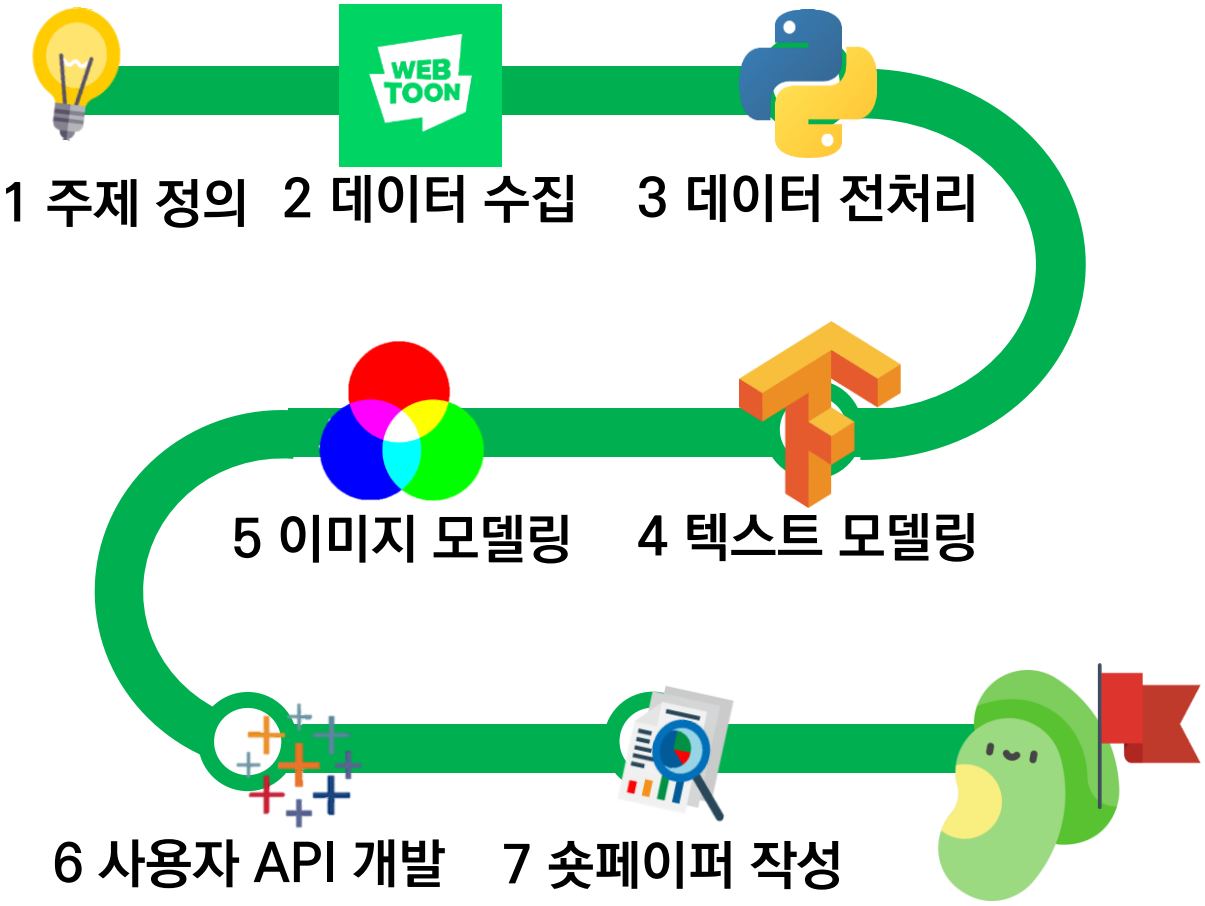
Image Analysis

홍만호

Text Mining

파이썬 기반의 빅데이터 분석 처리 과정

빅데이터 파이썬 프로그래밍	Jupyter, 데이터 유형 및 기초 문법, 데이터 입출력 제어문, Pandas, Numpy
빅데이터 전처리	데이터베이스와 SQL, 데이터 전처리 차원 축소(PCA)
빅데이터 분석 및 활용	빅데이터 분석 기본 방법론 실습 , 빅데이터 시각화 회귀분석, SVM, 결정트리, 랜덤포레스트, xgBoost, 클러스터링 연관규칙, 빈발패턴, 순차패턴 인공신경망, 딥러닝
텍스트마이닝	정규방정식, 텍스트 전처리(nltk, Konlpy) TF-IDF, Clustering Topic Modeling(LSA, LDA, Visualization) Sentiment Analysis(감성분석) Word Embedding(W2V, 딥러닝)
분산 빅데이터 처리	Apache Spark , Spark SQL Spark Streaming, Zepplin , Spark ML



사용자 API 개발 | 논문

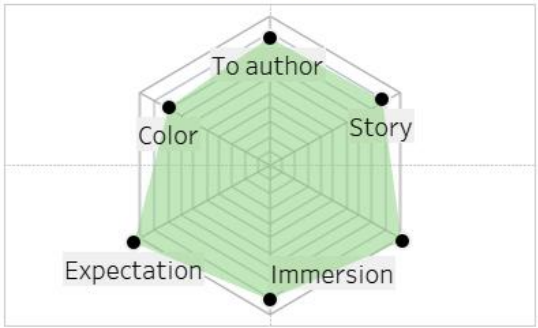
Naver Webtoon
Analyzing 10 Highest Popular Webtoons

과학기술정보통신부 한국데이터산업진흥원

Info

제목 : 신의 탑
작가명 : SIU
장르 : 판타지
연재요일 : 월요일
연재일자 : 2016-06-30
컷 당 글자 수 : 28.4 자
댓글 수 : 122487건

Radar Chart



Main Color

Recommendation

Main Comment Per Topic

기대감 : 우리나라에서 유일하게 윈피스나루토진격의거인이랑 맞먹을 수 있는 웹툰이다

그림체/소재 : 그림체 원래도 이뻤는데 지금 진짜 많이 이뻐졌어

몰입도 : 이 때만 해도 남주가 여주한테 집착하는 집착물인줄 알았지

스토리 : 가면 갈수록 스토리가 전혀 지루하지 않아

To.작가 : 흥미진진하네요 다음 화를 기대하겠습니다

Created By 완두콩

사단법인 한국통계학회 추계학술대회 투고(11월 중)

Preference index development of individual contents based on Naver Webtoon

Jungyeon Lee[a] Hui-yeon Jo[b] Man-ho Hong [c]

Summary: The webtoon is a new genre of cartoon that published through the internet, and is considered a unique case in global cartoon market. A radical increasement of internet users has established conditions for fostering the platform for webtoons and their production. In addition, the popularity of Korean dramas and K-pop has cultivated the spread of webtoons in the countries receiving dramas and K-pop. Compared to the globally growing market and contents, feedback that consumers can deliver to originator is limited to star ratings and comments. Star ratings are a general indicator, and originator can't figure out the strengths and weaknesses of your content. In this study, we develops preference index of personal content based on Naver webtoon. This is expected to help smooth communication between creators, consumers and the platform.

Keywords: Webtoon, Preference Index, Topic Modeling, Sentiment Analysis, Color Similarity

[a]Undergraduate student, Department of Statistics and Information Science, Dongduk Women's University, Seoul 02748, Republic of Korea.

[b]Graduate student, Department of Statistics and Information Science, Chungbuk National University. 1, Chungdae-ro, Seowon-gu, Cheongju-si, Chungcheongbuk-do, Republic of Korea. volume893@gmail.com

[c]Undergraduate student, Department of Applied Statistics, Gachon University. 1332, Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Republic of Korea. ghdaksgh123@naver.com

1 Introduction

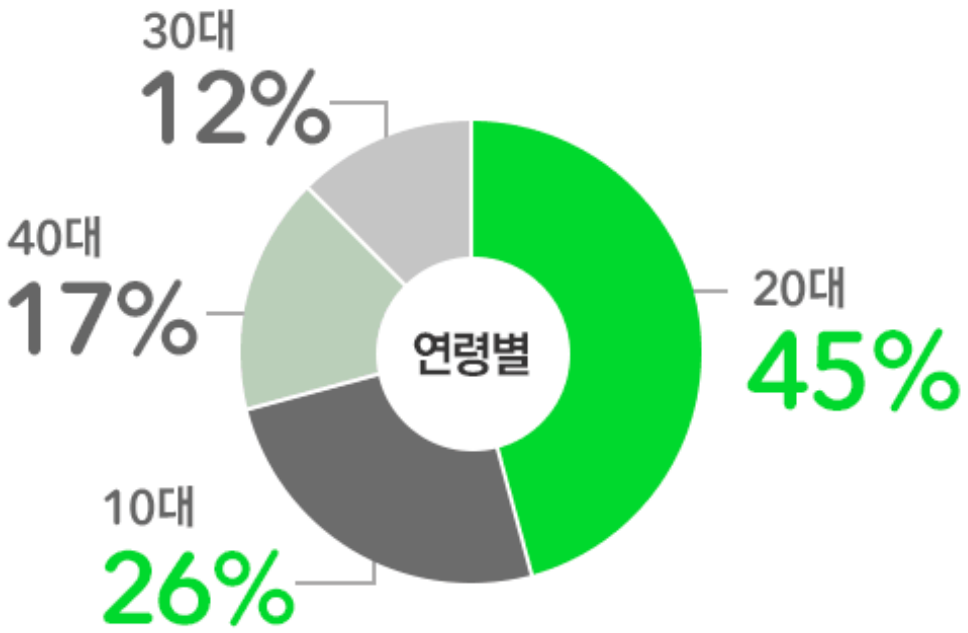
웹툰은 인터넷을 뜻하는 '웹(web)'과 만화를 뜻하는 '카툰(cartoon)'의 합성어로, 스마트 기기의 보급이 시작되면서 이용자 수가 빠르게 증가했다. 지하철, 버스와 같은 대중교통에서 스마트폰 등으로 웹툰을 보는 사람들이 많은 것만 봐도 그렇다. 2018년 12월 26일 기준, 웹툰 분석 서비스(WAS)에 의하면 2018년 신규 등록된 웹툰 수는 2,254편이다. 기존에 연재 중인 작품을 제외하고 매일 신작 웹툰이 6.17편 이상 공개된 것이어서 시장의 활성과 기대감이 반영된 수치라 할 수 있다. 이에 따라 만화, 애니메이션 관련 학과의 인기가 높아지고 있으며, 이런 환경은 수준 높은 작품으로 이어지고, 좋은 작가들을 더 많이 만들어내는 구조를 형성하고 있다. 국내에서는 웹툰의 OSMU(One Source Multi Use; 하나의 소재를 서로 다른 장르에 적용하여 파급효과를 노리는 마케팅 전략) 산업이 활성화되고 있다. 이미 검증된 스토리와 대중성, 영상으로 만들기 용이한 시각적 특성을 가졌기 때문에 웹툰의 영상화가 증가하는 것이다. 이와 같이 국내 웹툰 플랫폼들은 광고, 유료결제, 굿즈, 2차 콘텐츠와 같은 수익모델로 매출을 증가시켰고, 플랫폼 간의 경쟁으로 국내 시장을 키웠다. 현재는 국내 주요 플랫폼이 해외 시장 진출을 중요하게 여기고 있으며, 실제로 몇몇 작품들의 진출이 성공하는 등의 성장을 이루고 있다. 웹툰의 평가 지표로 별점, 조회수가 있지만, 별점이 높다고 해서 가장 인기 있는 웹툰인 것도 아니며, 이러한 총평가 지표로는 해당 웹툰의 세부 강·약점을 파악할 수 없다. 웹툰 작가는 댓글 전체를 확인하기 어렵기 때문에 독자들의 전반적인 피드백을 받는 데에 한계가 있다. 본 논문에서는 독자들의 다양한 평가를 효과적으로 전달할 수 있는 지표를 개발하려 한다. 먼저 독자들의 주관적인 의견이 들어간 댓글들을 이용한다. 토픽 모델링을 활용하여 주요한 주제를 뽑아내고, 토픽별로 감성분석을 이용하여 해당 댓글의 점수를 부여한다. 이후 기존에 연재 중인 웹툰과의 장르별 색상 유사도를 계산할 것이며, 그 안에서도 각 웹툰의 별점을 이용해 더 인기 있는 웹툰에 가중치를 둘 것이다. 또, 글이 지나치게 많은 웹툰을 선호하지 않는 독자를 위해 OCR(Optical Character Reader/Recognition) 기능을 적용하여 컷당 평균 글자 수를 계산해 제공할 것이다. 본 연구의 공헌은 다음과 같다. 댓글과 웹툰 이미지를 이용해 선호도 측정 점수를 부여하여 작가는 자신의 콘텐츠에 대한 객관적 접근을, 독자는 자신의 취향에 맞는 선택을, 플랫폼은 투자할 가치가 있는 콘텐츠를 파악할 수 있게 한다.

1 문제 정의
- 주제 선정 배경
- 목표
- 기대효과

2 분석 과정
- 프로젝트 구성도
- 텍스트 분석
- 이미지 분석
- 스코어링
- 타당도 검증

3 활용 방안
- API 구현
- 확장성

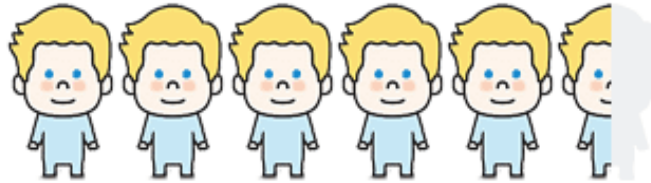
20대가 즐겨찾는 문화 콘텐츠, 웹툰



네이버 웹툰 이용자 연령대

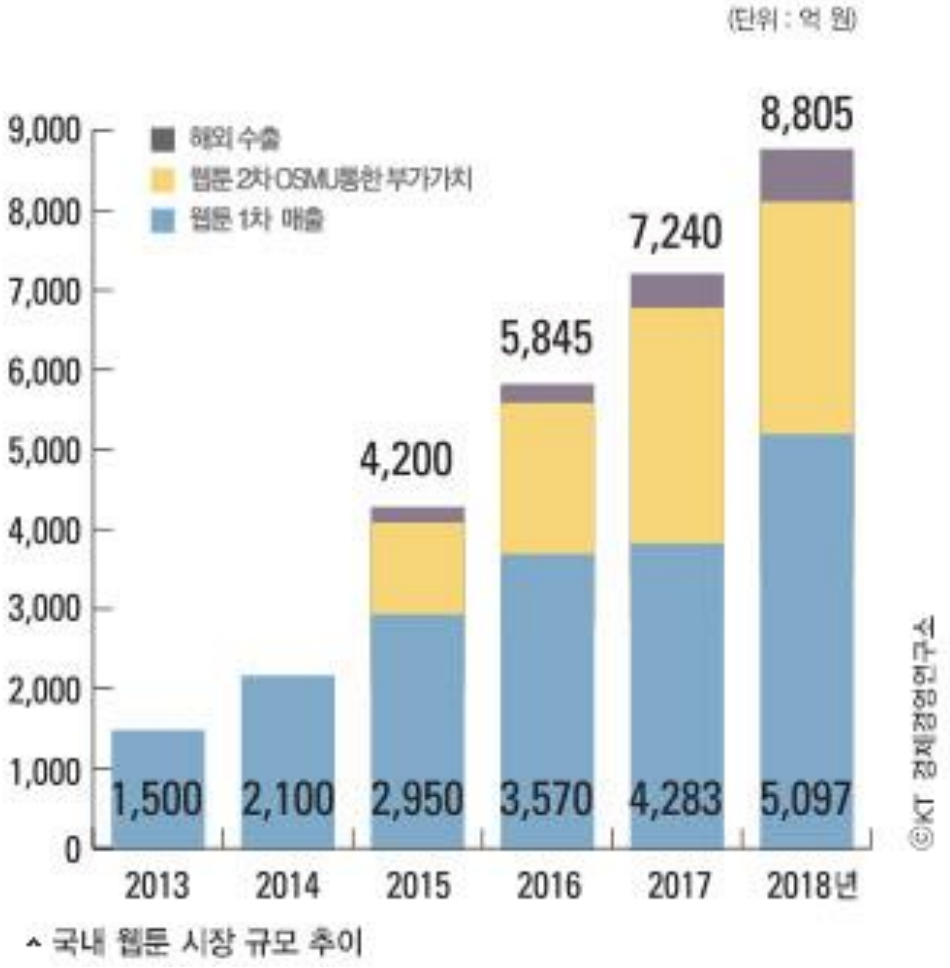
웹툰 시장의 성장 : 웹툰 이용자 수 및 시장 규모 증가

하루 평균 웹툰 이용자 약 620만 명



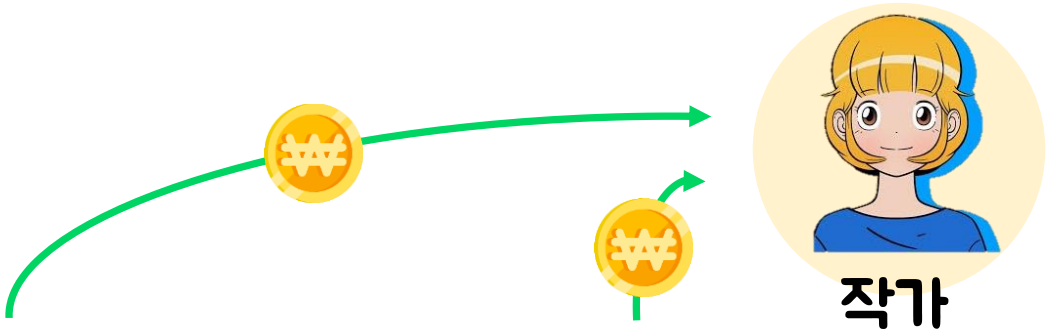
55%(남)

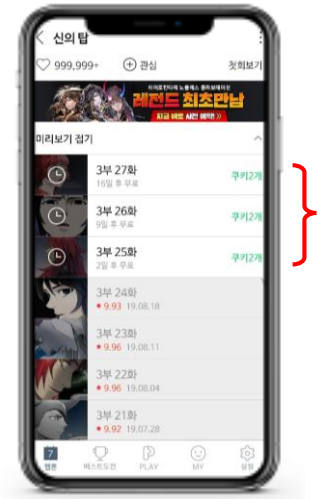
45%(여)



^ 국내 웹툰 시장 규모 추이

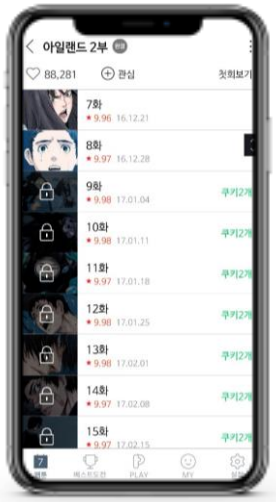
웹툰 시장의 성장 : 웹툰 유료화 수익 모델 확립





한 화당
200원

(1) 다음 화 미리보기



한 화당
200원

(2) 완결작품 다시보기

(3) 웹툰 내 광고

(4) OSMU

문제 제기

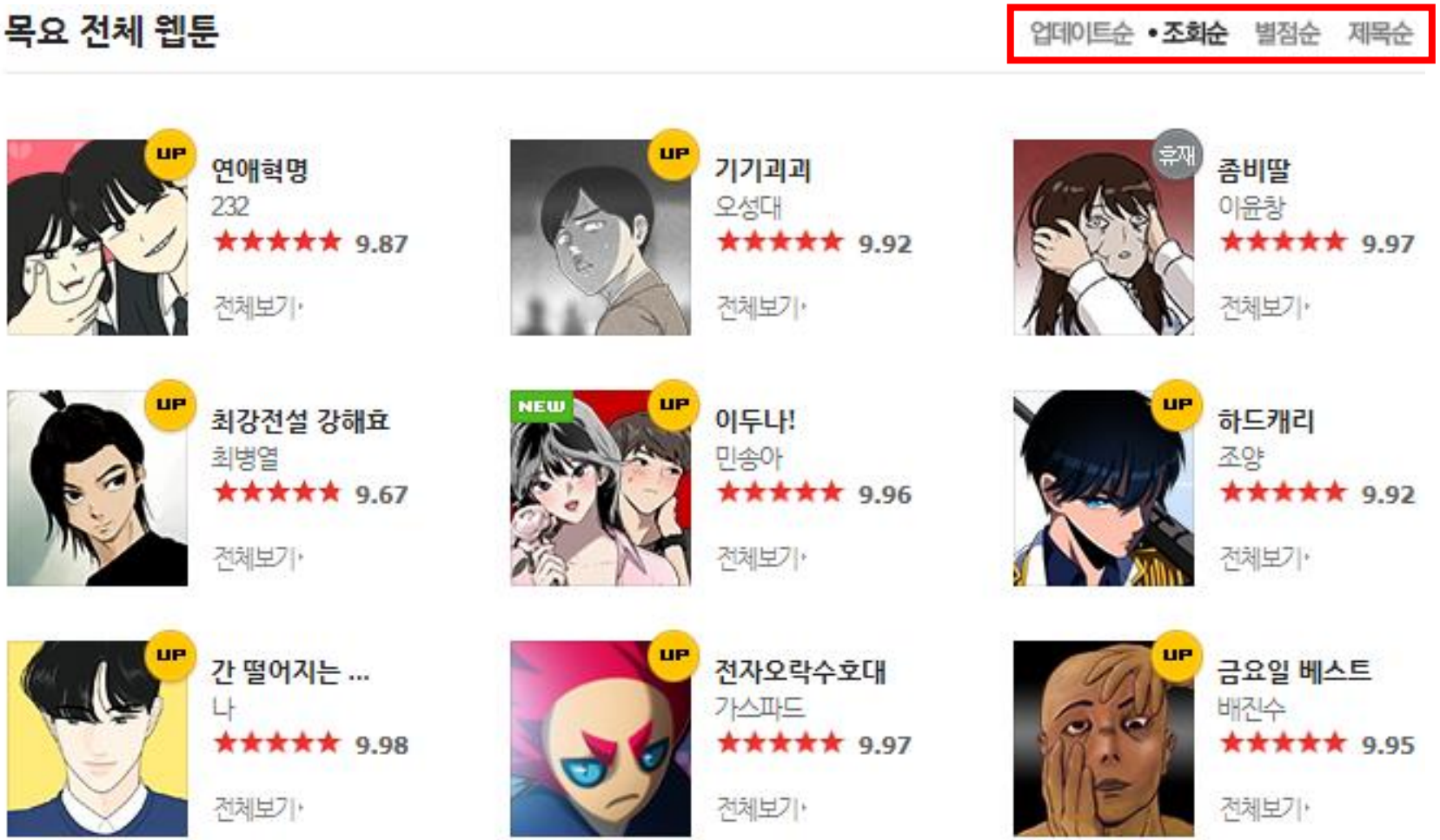
작가에게 독자들의 피드백이 효과적으로 전달되나요?

1) 조회수, 별점 등 콘텐츠에 대한 총 평가 지표만 존재

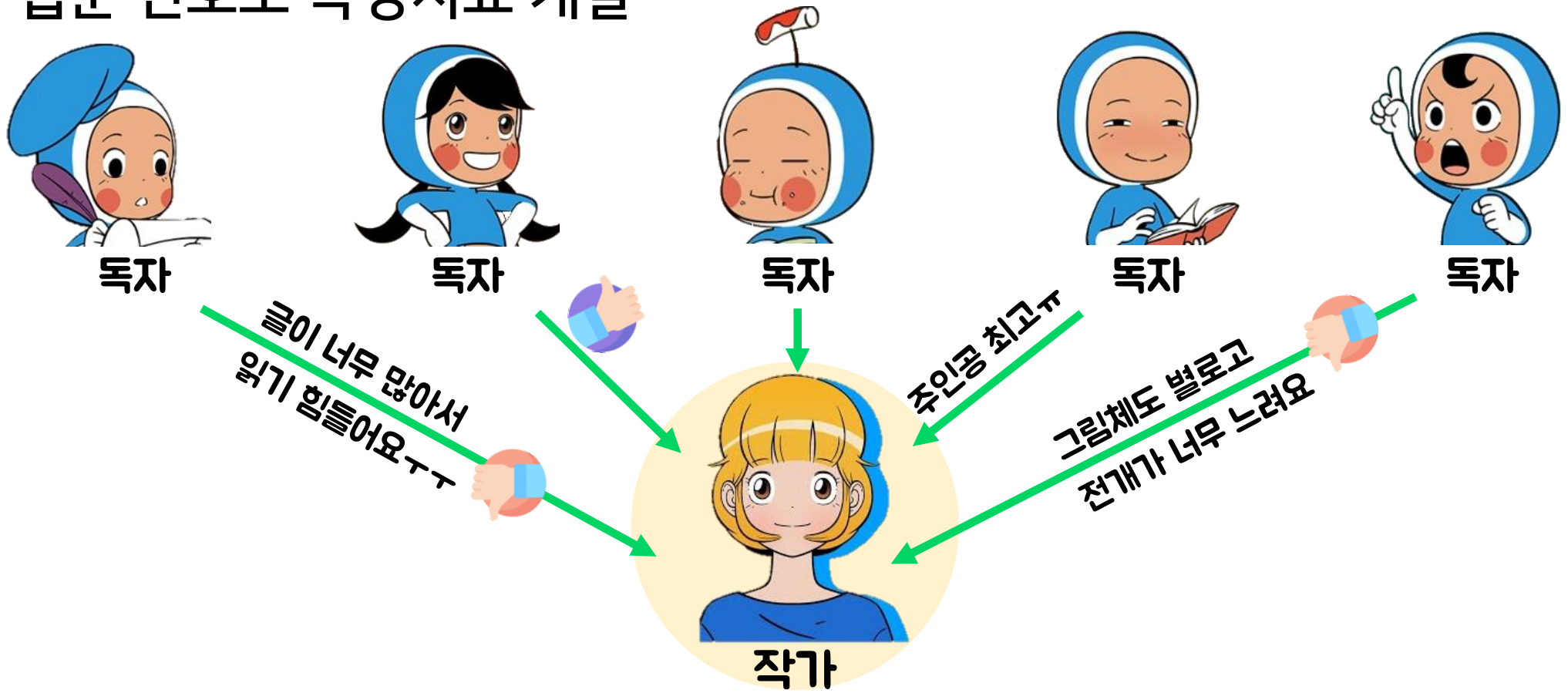
2) 댓글 전수 확인 불가능

2) 컨텐츠의 세부 강·약점 한눈에 파악하기 어려움

문제 제기


목요 전체 웹툰
업데이트순 • 조회순 별점순 제목순
UP
연애혁명
232
9.87
전체보기
UP
기기괴괴
오성대
9.92
전체보기
휴재
좀비딸
이윤창
9.97
전체보기
UP
최강전설 강해효
최병열
9.67
전체보기
NEW
UP
이두나!
민송아
9.96
전체보기
UP
하드캐리
조양
9.92
전체보기
UP
간 떨어지는 ...
나
9.98
전체보기
UP
전자오락수호대
가스파드
9.97
전체보기
UP
금요일 베스트
배진수
9.95
전체보기

목표 : 웹툰 선호도 측정지표 개발



독자들의 다양한 평가를 효과적으로 전달할 수 있는 지표 개발

기대효과 : 작가, 독자, 플랫폼

작가

콘텐츠에 대한
객관적인 접근

독자

자신의 취향에 맞는
웹툰 선택의 폭 넓어짐

플랫폼

수익성 좋은 웹툰
사전에 인지하여 투자

2 분석 과정

분석과정 개요

1 데이터 수집
네이버 웹툰 내
댓글 및 이미지 크롤링

2 텍스트 분석
- 토픽 모델링
- 감성 분석

3 이미지 분석
- 색상 유사도
+ 가독성

4 스코어링

5 타당성 평가
2019 네이버 웹툰 최강자전
본선 진출작 예측

토너먼트 진행 방식

- 접수 작품 중 예선 진출작 105작품 선정
- 예선 105작품 중 독자투표수가 많은 상위 32작품을 대상으로 토너먼트 진행
- 32강 〉 16강 〉 8강 〉 4강 〉 결승 순서로 독자투표수를 기준으로 한 토너먼트 진행

1) 데이터 수집

*2019-08-15기준

Train data

현재 연재중인
네이버 웹툰 댓글

4,782,553건

현재 연재중인
네이버 웹툰 최신화 ~ -5화

242 작품

Test data

2019 네이버 웹툰 최강자전
예선작 댓글

16,692건

2019 네이버 웹툰 최강자전
예선작 1화

105 작품

1) 텍스트 분석과정 개요

1 댓글 필터링

- 자/모음, 영어, 숫자, 특수문자 제거
- 6글자 이상 필터링
- 웹툰 별 주인공 이름 레이블링

댓글 4,782,553건 -〉 3,819,687건

2 전처리

- PyKoSpacing으로 띄어쓰기 수정 *
- 원형 복원(Lemmatization)
- Okt와 MeCab 활용 형태소 분석*
- 불용어 처리

3 토픽 모델링

- LDA
- 정확도를 위해 Okt와 Mecab에서 중복되는 명사만 사용

4 감성분석

- 감성사전 구축
- 댓글(좋아요, 싫어요) 가중치로 사용

'신의 탑' 밤 -〉 주인공

2) 주요 전처리 : 띄어쓰기 보정

PyKoSpacing 모듈 활용하여 띄어쓰기 보정

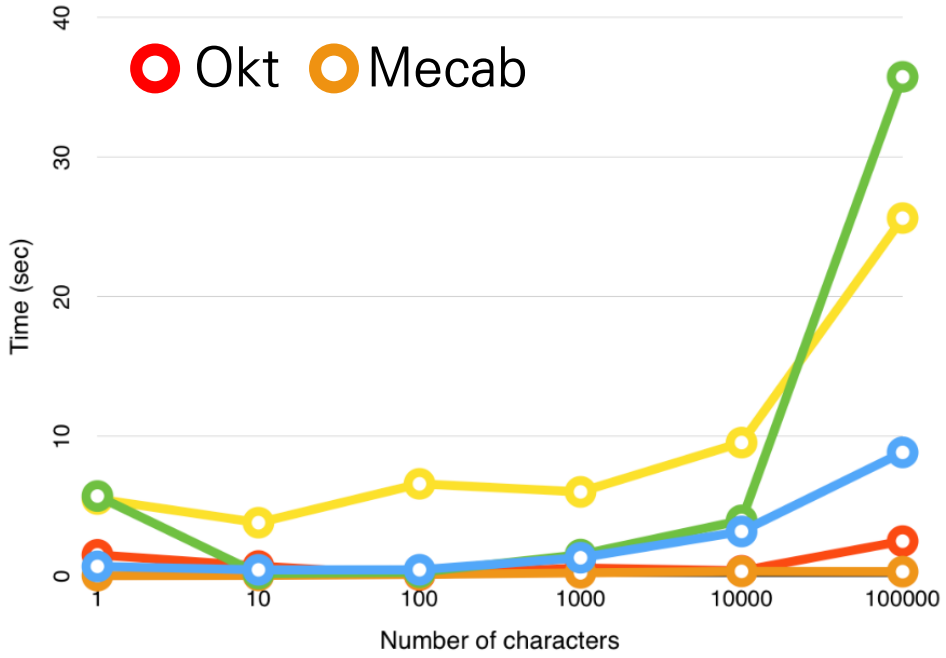
: 뉴스 기사로 띄어쓰기 규칙을 학습시킨 딥 러닝 모듈

```
In [2]: spacing('아 버 지 가 방 에 들 어 가 신 다 ')
Out[2]: '아 버 지 가  방 에  들 어 가 신 다 '
```

띄어쓰기 되어있지 않은 댓글 처리 예시

```
In [2]: spacing('만 호 는 귀 염 뽀 짝 한 연 애 가 하 고 싶 어 요 ')
Out[2]: '만 호 는  귀 염  뽀 짝 한  연 애 가  하 고  싶 어 요 '
```

2) 주요 전처리 : Okt, Mecab 활용한 형태소 분석



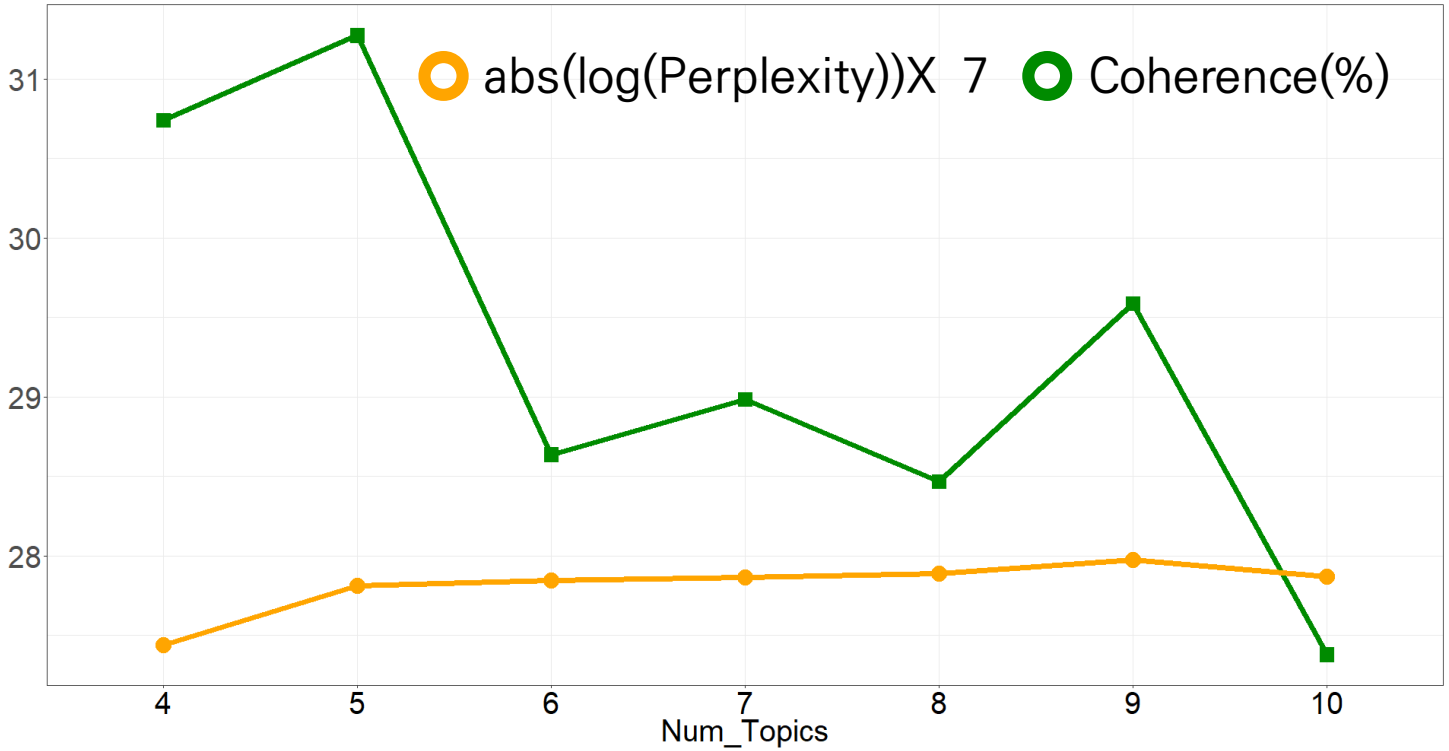

```
In [9]: Okt().pos(sentence,
   ...: stem = True,
   ...: norm = True)
Out[9]:
[('만 호 ', 'Noun'),
 ('는 ', 'Josa'),
 ('귀 염 뽀 짝 ', 'Noun'),
 ('한 ', 'Josa'),
 ('연 애 ', 'Noun'),
 ('가 ', 'Josa'),
 ('하 다 ', 'Verb'),
 ('싶 다 ', 'Verb')]
```

```
In [10]: Mecab().pos(sentence)
Out[10]:
[('만 호 ', 'NNP'),
 ('는 ', 'JX'),
 ('귀 염 ', 'NNG'),
 ('뽀 짝 ', 'MAG'),
 ('한 ', 'XSA+ETM'),
 ('연 애 ', 'NNG'),
 ('가 ', 'JKS'),
 ('하 ', 'VV'),
 ('고 ', 'EC'),
 ('싶 ', 'VX'),
 ('어 요 ', 'EF')]
```

정확성을 위해 Okt와 Mecab의 교차 명사만 사용

만호 연애

3) LDA : 토픽 개수 선정



Perplexity

: 값이 작을수록 학습이 잘 되었다고 평가.
하지만, 낮은 값이 항상 토픽모델링 해석에 적절한 결과를 의미하지는 X

Topic Coherence

: 상위 단어 간의 유사도를 계산.
해당 주제가 의미적으로 일치하는 단어들끼리 모여 있는지 알 수 있는 지표

3) LDA : Topic & Keywords

Topic 1 (10.6%)

"또 대작 타는 냄새 드립이 난무하겠지만 이건 대작 같아요 다음 내용이 궁금해요"

기대감

Topic 2 (54.7%)

"작가님 제발 그림작가 구하시면 안될까요 그림작가만 있으면 웹툰 세계 정복 가능 "

그림체, 소재

3) LDA : Topic & Keywords

Topic 3 (6.9%)

"저 누렁머리 남자 여주 좋아하면 양다리 되던가
저 파랑머리가 여주 좋아하면
그럼 걍 사귀는 거지 후후후"

몰입도

Topic 4 (18.3%)

"스토리 좋고 그림체 좋고
분량도 적당한데 왜 별점이 정도지
신작이라고 별점테러 하지 좀 말자
진짜 꼴보기 싫다 "

스토리

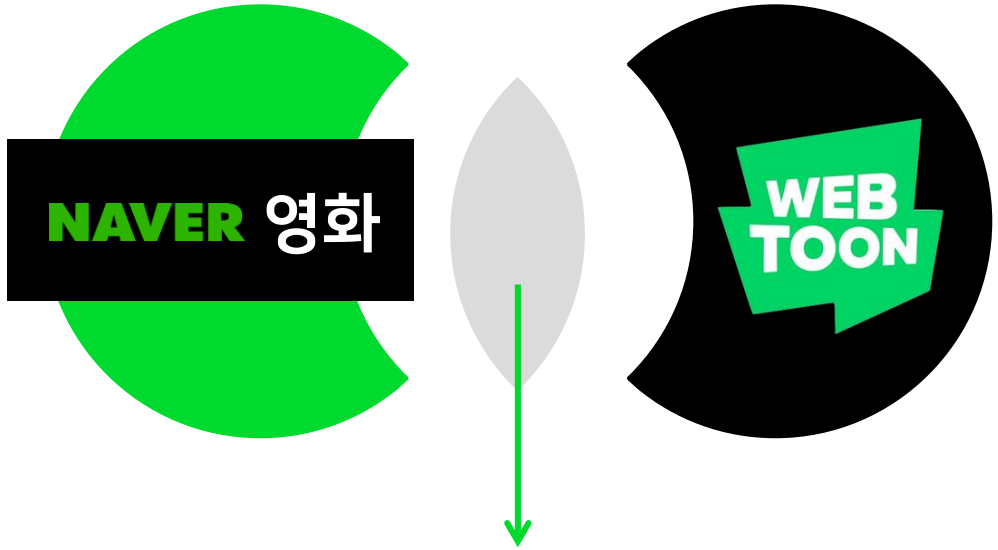
Topic 5 (9.5%)

"이거 웹툰 공모전 출전했던 작품이네
계속 이 작품만 투표했었는데
네이버 정식연재로 또 뵙게 되네요
축하드려요 얼른 위로 올라가세요 "

To. 작가

4) 토픽 별 감성분석

네이버 영화리뷰 15만건 기준으로 제작된 감성사전 사용



소속된 포털사이트 동일

대부분이 구어체인 리뷰성 댓글

긍•부정 판별 예시

```
In [3]: Grade('흥  별 로  재 밌 네 요 ')
긍 정 입 니 다
```

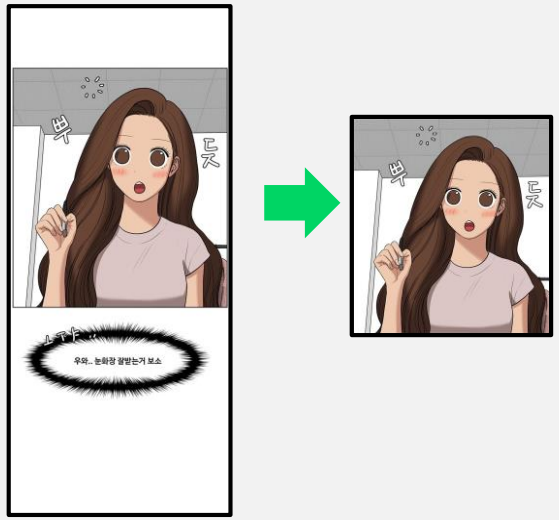
좋아요 싫어요 개수 가중치로 사용

1) 이미지 분석과정 개요

1 전처리

- 배경색 제외

: Image Segmentation



야옹이 〈여신강림〉

2 색상 유사도 측정

- RGB값 추출

: 각 웹툰별로 상위 6개 색상 추출

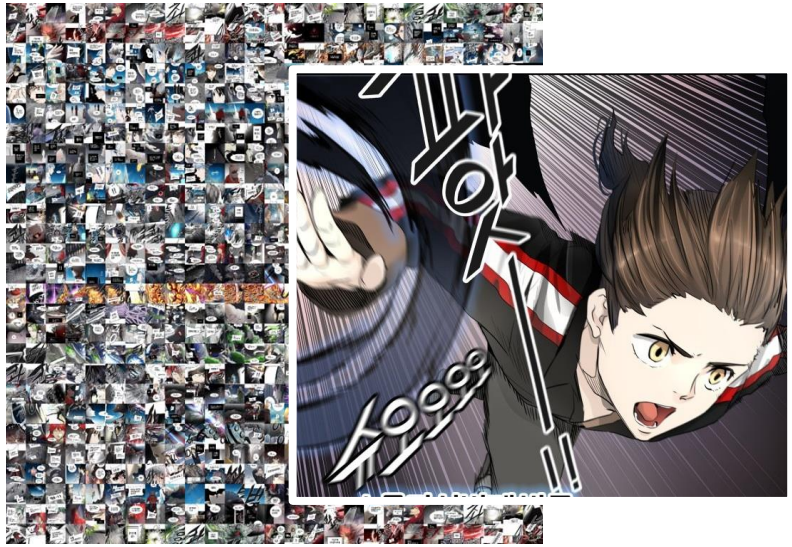
- 별점(5.01 ~ 9.99)을 가중치로 한 장르별 색상 유사도 계산
- min-max scailling 정규화

+ 가독성 측정 +

- OCR로 텍스트 추출
- 웹툰 단위로 컷 당 평균 글자 수 측정

2) 색상 유사도 측정 – RGB값 추출

: 웹툰 단위로 6개의 대표 색상 추출



SIU 〈신의 탑〉

순끼 〈치즈 인더 트랩〉

2) 색상 유사도 측정 – 장르별 색상 유사도 계산

장르별 네이버 웹툰 별점을 가중치로 장르별 색상 유사도 계산

네이버 웹툰 : SF / 판타지

Doll 체인지
★★★★★ 9.97

2인용 인간
★★★★★ 9.96

간 떨어지는 동거
★★★★★ 9.98

히어로메이커
★★★★★ 9.96

최강자전

SF / 판타지
간택되었다냥!

Similarity 1 Similarity 2 Similarity 3 Similarity 4

Weighted Average L2 Distance

+) 가독성 측정

강호진 〈 호랭총각 〉

구글드라이브 OCR

– 크롤링된 이미지 단위로 OCR 수행

호랭총각 장가보내기 추진위원회 공동연합

MOU

호랭 호림

림

MOU

짝짝에

그럼 우리 너구리 연맹은 모두 힘을 합쳐 호랭총각 장가보내기에 온 힘을 쏟도록 합시다!

장가!

가즈아!

가독성 측정

– 비율 = 글자수 / 컷 수

양영순 〈덴마〉

스코어링 : 텍스트 점수, 이미지 점수

Title	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Color_Score	Score
신의 탑	97	95	84	81	79	72	508
복학왕	51	65	58	57	55	66	352
연애혁명	81	82	61	55	82	81	442
용이산다	97	81	84	81	79	55	477
열렙전사	88	80	84	81	79	33	445
뷰티풀 군바리	81	88	76	77	91	78	491
마음의 소리	97	95	84	81	79	69	505
유미의 세포들	97	95	84	81	79	50	486
기기괴괴	97	95	84	81	79	26	462

2019 네이버 웹툰 최강자전 본선 진출작 예측

name	실제 순위	지표 순위
오늘 죽는 너에게	1	1
오로지 오로라	2	2
하루만 네가 되고싶어	5	3
아침을 지나 밤으로	7	4
하늘은 왜 파랄까	4	5
왕년엔 용사님	9	6
⋮	⋮	⋮
붓꽃 예술 고등학교	10	15
아르테미스 신드롬	19	16
⋮	⋮	⋮
빤타스틱 스위퍼	40	32

16강 진출작품 $\frac{15}{16}$ 일치

32강 진출작품 $\frac{25}{32}$ 일치

순위상관계수
0.759

3 활용방안

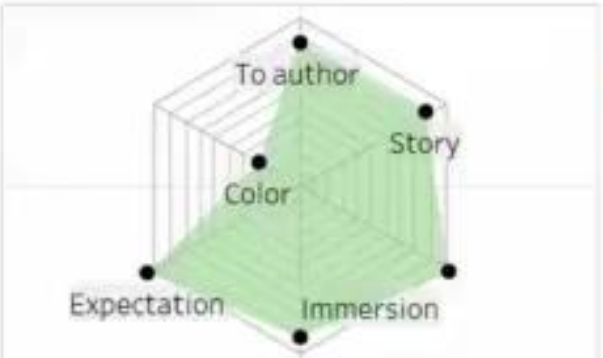
Naver Webtoon
Analyzing 10 Highest Popular Webtoons

과학기술정보통신부 한국데이터산업진흥원

Info

제목 : **갓 오브 하이스쿨**
작가명 : **박용제**
장르 : **액션**
연재요일 : **금요일**
연재일자 : **2011-04-08**
컷 당 글자 수 : **15.7 자**
댓글 수 : **53879건**

Radar Chart



Main Color

Recommendation

Main Comment Per Topic

기대감 : **우리나라에서 유일하게 원피스나루토진격의거인이랑 맞먹을 수 있는 웹툰이다**

그림체/소재 : **그림체 원래도 이뻤는데 지금 진짜 많이 이뻐졌어**

몰입도 : **이 때만 해도 남주가 여주한테 집착하는 집착물인줄 알았지**

스토리 : **가면 갈수록 스토리가 전혀 지루하지 않아**

To.작가 : **흥미진진하네요 다음 화를 기대하겠습니다**

Created By 완두콩

해외시장

일본 디지털 코믹스 시장

라인망가	NHN 〉 LINE	·라인메신저의 성공으로 일본 3위 디지털 코믹스
코미코	NHN 〉 NHN Play Art	·한국형 웹툰의 일본시장 진출 성공케이스, 시장 1위
픽코마	Daum Kakao 〉 Kakao Japan	·카카오재팬의 일본 웹툰/웹소설 유료 서비스, 2017년 공격적인 마케팅과 서비스로 일본시장 2위 점유

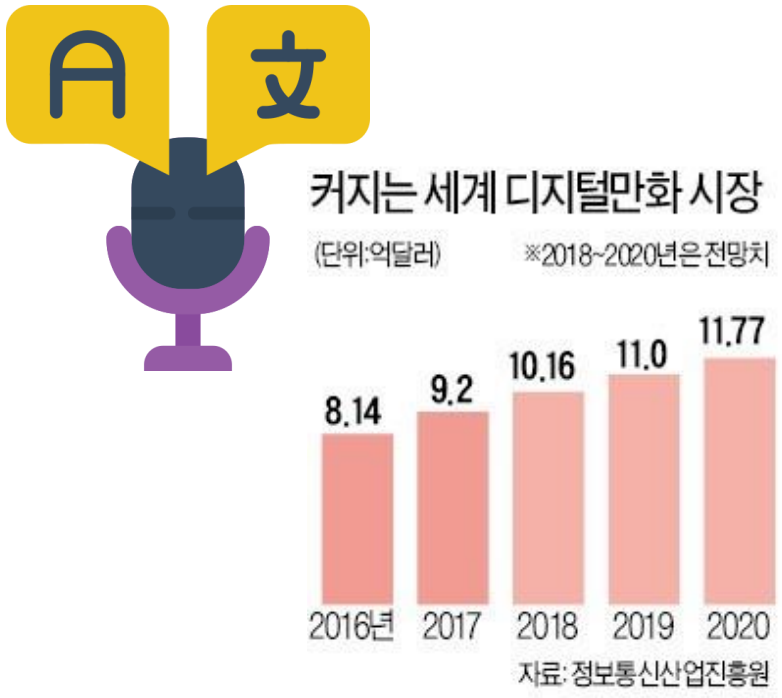
인도네시아/태국/대만 디지털 코믹스 시장

인도네시아	라인웹툰	네이버	·1위 웹툰 서비스업체
	코미카	파노라마 〉 코미카	·2위 웹툰 서비스 업체
태국	라인웹툰	–	·1위 디지털 코믹스 서비스 업체
	Ookbee 코믹스	–	·현지 2위 디지털 코믹스 서비스 업체
	코미코	NHN 〉 NHN Play Art	·코미코 재팬이 본사, 한국, 대만, 태국, 2
대만	라인웹툰	네이버	·현지 1위 웹툰 업체
	탑툰	–	·한국 탑코의 대만 서비스, 현지 2위
	코미코	NHN 〉 NHN Play Art	·코미코 재팬이 본사, 한국, 대만, 태국, 2
	투믹스	–	·대만진출 준비 완료, 시기 조율 중

북미 디지털 코믹스 시장

DC코믹스	타임워너	·슈퍼맨, 배트맨 등의 Justice League, 종이책과 동일한 디지털만화 판매, 영화로 확장 진행 중
마블코믹스	디즈니 〉 Marvel	·아이언맨, 토르, 헐크, 캡틴아메리카 등의 Marvel Universe, 종이책과 동일한 디지털만화 판매, 영화로 성공적인 확장
이미지코믹스	–	·마블출신의 작가 7명이 1992년 창업한 출판사, 워킹데드로 유명, 종이책과 디지털 동시 취급
코믹솔로지	Amazon	·마블, DC, 이미지코믹스, IDW등의 출판사 만화서비스 2014 아마존에 인수됨
타파스	–	·김창원 대표, 한국계 미국 최초 웹툰업체(초기 다음과 지속적인 협업, 현재는 로컬화 완료)
라인웹툰	네이버 〉 네이웹툰	·전세계 대상 서비스, 북미 300만 MAU 확보로 지속적인 확대 진행 중
레진코믹스	–	·2015년 12월 북미 서비스 시작, 팬덤 확보하며 지속적인 성장
태피툰	–	·앱기반, 국내우수 웹툰 번역 서비스 진행 중(호평 속에 지속 성장 중)

해외시장



국가별 반응 키워드별로 확인 가능

언어 번역해서 동일한 지표로 활용

다른 도메인으로의 확장: YouTube

영상 썸네일과 댓글을 데이터로 사용하여 선호도 측정지표 동일하게 사용

인기 급상승 동영상 #1

※기내 🩸류 반입 금지※ 약 빤 장성규의 꿈의 직장★ 상큼터지는 항공사 직업 리뷰 | 워크맨 ep.15

조회수 3,914,562회 8.3만

양춘자 20시간 전
장성규 진짜 제정신인가 안 웃긴 편이 없넼ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
111 답글

이해수 21시간 전(수정됨)
JTBC에서 아나운서체험 어떰?ㅋㅋㅋㅋㅋ
464 답글
답글 8개 보기

무민 22시간 전(수정됨)
7:37 머리 젖어요
발 젖어요 다리 젖어요 다아젖어요
궁댕 방실방실 킬포 ㅠㅠㅠ
1천 답글
답글 5개 보기

References

[1] 네이버 웹툰(https://comic.naver.com/index.nhn)

[2] 이동주, 연종흠, 이상구 (2011). 한국어 문장의 띄어 쓰기 오류 교정과 최적 형태소 분석을 위한 통합 확률 모델, 한국정보과학회 학술발표논문집, 38(1A), 237-240

[3] https://github.com/lokesh/color-thief/

[4] https://sosal.kr/1067

감사합니다

APPENDIX :

01 사용 데이터

02 텍스트 분석 상세

03 이미지 분석 상세

Raw data

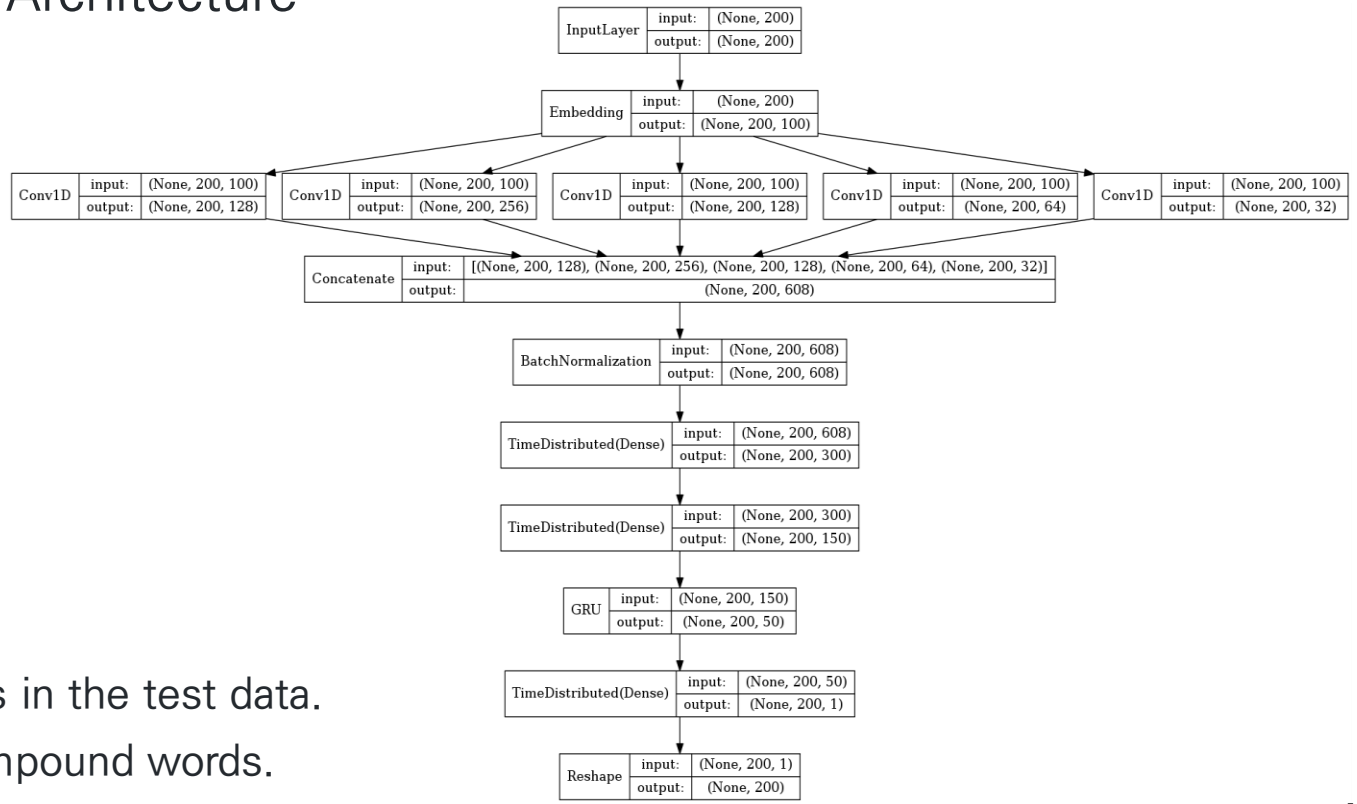
lchi****,lchinin3,2019-08-14,신의 탑,1,1,진짜 그저 감탄만 번째정주행,2,0
icec****,응아니야,2019-08-14,신의 탑,1,1,프롤로그 의문점하츠명예아낙복수쿤권력라
gimg****,김규리,2019-08-14,신의 탑,1,1,아 정주행 번째 마약웹툰,1,0
ycr1****,BLDoo,2019-08-14,신의 탑,1,1,-,0,0
ckd0****,창준,2019-08-14,신의 탑,1,1,핸드폰 바꾸니깐 봤다는 표시가 없어져서 다시
mmal****,이레미,2019-08-14,신의 탑,1,1,정주행 일 질아신탑은 시작하면 정독은 기본
sanl****,이산,2019-08-14,신의 탑,1,1,또오오오오보러오떠,2,0
wiw6****,wiw6****,2019-08-13,신의 탑,1,1,첫번째 베댓 마춛뻡 많이 거슬리네요,8,0
jsh0****,아아므,2019-08-13,신의 탑,1,1,내 여친도 거깄냐,5,0
hw10****,최현웅,2019-08-13,신의 탑,1,1,번째다시봐도 갓작이다,4,0
sm91****,sm91****,2019-08-13,신의 탑,1,1,베댓 맞춤법 거슬리네,3,0
ldw2****,dlehdnjs,2019-08-13,신의 탑,1,1,정주행 한번하는데 일은 걸리는듯,0,2
kimj****,김준서,2019-08-13,신의 탑,1,1,아 이제 이거 정주행해볼까,1,0
2113****,KON,2019-08-13,신의 탑,1,1,와 댓글 만개 ,2,0
qkrd****,박이레,2019-08-13,신의 탑,1,1,번째 정주행,0,0
rkdj****,단짝그리고워너블,2019-08-13,신의 탑,1,2,벳댓 주인공도 믿지마,0,0
ciwo****,김시원,2019-08-13,신의 탑,1,2,난 년,0,0
yhk2****,연히,2019-08-12,신의 탑,1,2,년덩안,2,0
nica**** 박세완 2019-08-12 신의 탑 1 2 신의탑 나왔을때부터 봤는데 너무 조아요 0 0

PyKoSpacing 모듈

Performance

Test Set	Accuracy
Sejong(colloquial style) Corpus(1M)	97.1%
OOOO(literary style) Corpus(3M)	94.3%

Model Architecture



- Accuracy = # correctly spaced characters/# characters in the test data.
  - Might be increased performance if normalize compound words.

감성사전 구축

1.정답이 있는 네이버 영화 리뷰 데이터 15만건에 대해서 **품사 태깅**

2.품사 태깅한 단어들에 대해 **Word2Vec**을 이용해 학습시킨 임베딩 벡터로 변환

3.단어 벡터들을 **BiLSTM**에 넣어서 양쪽 끝 state들에 대해서
fully connected layer와 **Softmax**함수를 이용해 분류

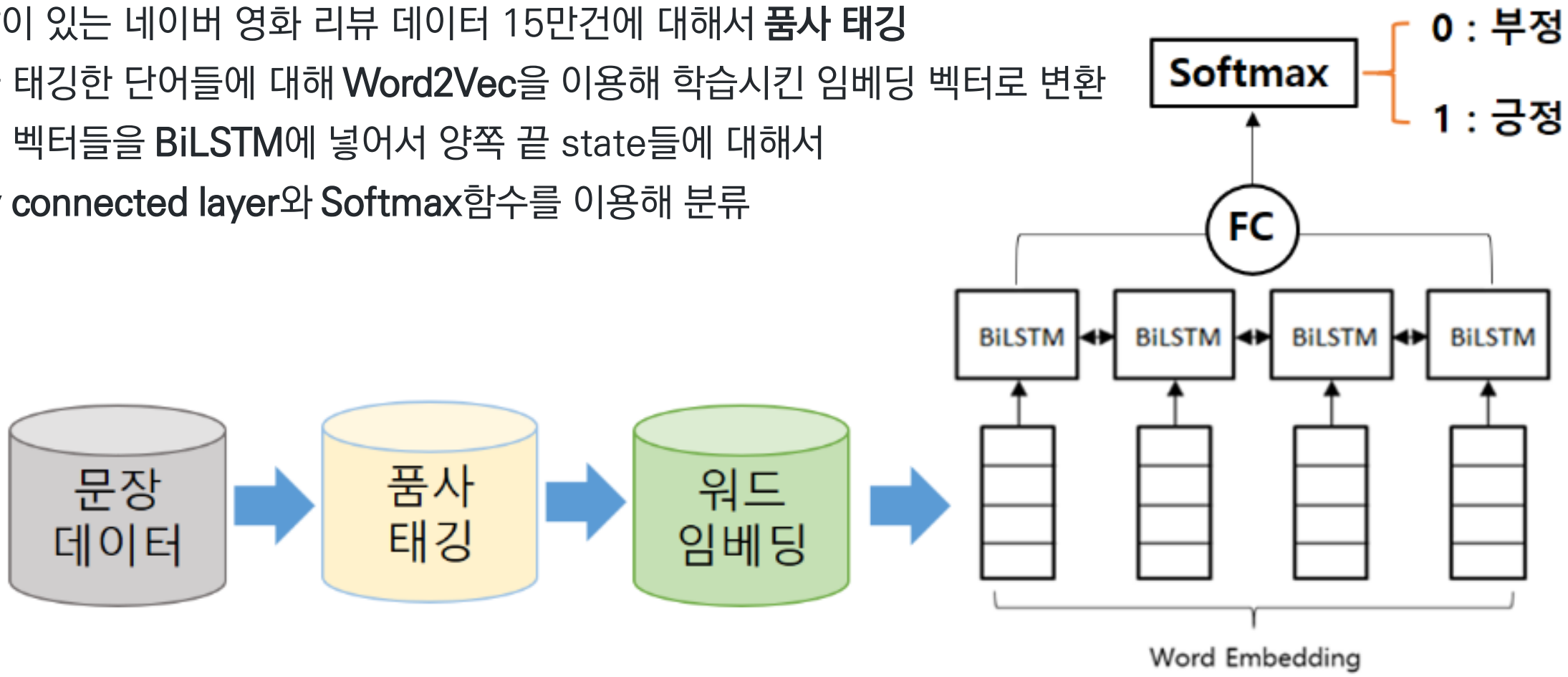
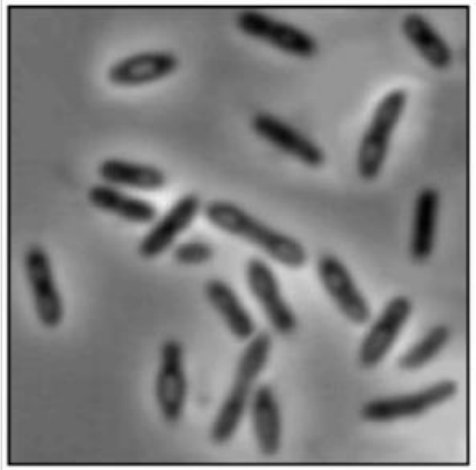


Image Segmentation

픽셀 기반 방법 이용한 Image Segmentation

: thresholding에 기반한 방식(특정 임계값을 정하고 그보다 작으면 검은색으로, 그보다 같거나 크면 흰색으로 표시하는 방법)

픽셀들의 분포를 확인한 후 적절한 threshold를 설정하고, 픽셀 단위 연산을 통해 픽셀 별로 나누는 방식이며, 이진화에 많이 사용이 된다.



이미지를 흑백처리 후

1. 흰 배경의 웹툰일 경우

: 흰 배경에서 검정색을 인식해서 Image segmentation

2. 검정 배경의 웹툰일 경우

: 검정 배경에서 흰색을 인식해서 Image segmentation

OCR(Optical Character Recognition, 광학 문자 인식)

사람이 쓰거나 기계로 인쇄한 문자의 영상을 이미지 스캐너로 획득하여

기계가 읽을 수 있는 문자로 변환하는 것

감사합니다