



낙시성 기사 분류기

BOAZ Advanced Project
9기 방대영 9기 김미성 7기 이지연

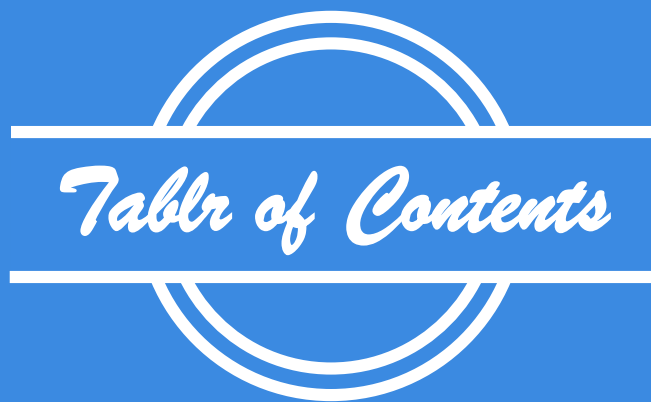
A graphic featuring the text "Table of Contents" in a white, cursive script font, centered between two horizontal white lines. Above and below the text are two concentric white semi-circles, creating a circular frame effect.

Table of Contents

1

개 요

2

데이터 설명

3

Feature 추출

4

모델링

5

결 론

1

개요

1 개요

[뉴스 서비스 문제 대두 -> 낚시성 기사]



네이버 뉴스, 아웃링크라는 끔찍한 선택 2018.05.11.
드루킹 사태를 둘러싸고 작성되는 여러 기사를 보고 있는 심정은, 쓸쓸하다 못해 우습다. 독자는 큰 관심이 없는데 언론만... 아예 네이버 뉴스를 없애버릴까? ...
자그니 | 칼럼 : 한없이 시끄럽고 믿을 수 없게 가까운
♡ 6명이 추천했습니다. | 💬 2



네이버 뉴스배치 조작...그들만의 문제일까? 2017.10.23.
아니 댄 굴뚝에 연기 날까. 그동안 소문만 무성했던 네이버의 뉴스배치 조작 의혹이 사실로... 그런데 네이버 뉴스배치 조작이 과연 그들만의 문제일까. ...
비주얼다이브 | VID 이슈
♡ 28명이 추천했습니다. | 💬 9

today issue : 네이버 뉴스 사람 손길 확 줄었다...뉴스섹션홈... 2018.03.13.
네이버가 AI 기반의 뉴스 추천 기술인 '에어스(AIRS: AI Recommender System)'를 모바일·PC의 '뉴스홈'과 '섹션홈'에 대폭 확대 적용했다. 네이버는 이번...
통플러스 | 오늘의 뉴스 브리핑
♡ 5명이 추천했습니다. | 💬 0

강호동, 동료에 폭행 휘둘러...피해자 증언
'논란' 장현수, 독일전에 결국...충격 예고
"KAL기 폭파 범인, 전두환..." 경악할 토로
독일전 주심, 난리난 결정...이를 어쩌나
류필립·미나, 파혼하나...사돈 갈등으로 '결국'
[속보] 추신수, 극적으로...역대급 경기 상황

1 개요

■ 낚시성 기사 종류

1. 드라마, 예능 내용을 현실 내용인것처럼 표현
2. 자극적인 단어를 사용해서 과장된 제목
3. 아예 제목과 내용이 전혀 다른 기사
4. 클릭하기 전과 클릭한 후에 기사 제목이 다른 것
5. 추측, 예측성 내용을 제목은 확신하듯이 쓰는 기사

강호동, 동료에 폭행 휘둘러...피해자 증언

노사연, 방송 녹화 도중 쓰러져...충격 사태

아이유 "바튼은 좋은 사람이야"

기사입력 2012.09.06 오전 08:47 | 최종수정 2012.09.06 오전 10:27 | 기사원문

18 185

글꼴 + -



수 안드레 아이유(22)는 새로 팀에 합류한 미드필

1 개요

- 기사 조회수 -> 언론사 광고 수익



- 끊임없이 생산되는 낚시성 기사
- 메이저 언론사도 낚시성 기사를 만들고 있는 상황

✓ 낚시성 기사를 사전에 차단할 수 있는 시스템 필요

2

데이터

2 데이터

1. 데이터 수집

[뉴스 기사 데이터 수집]

네이트 뉴스 사이트 : 2010년 1월 ~ 2017년 12월 (8년간)

- 댓글이 달려 있는 기사만 크롤링
- 약 8천 만개 기사를 코드로 훑음

[네이트 뉴스 선정 이유]

- 한 신문사에만 편향되지 않게
- 네이버보다 훨씬 크롤링 속도가 빠름

[낚시성 기사 기준]

- 베스트 댓글, 일반 댓글에 '낚', '낚시', '기레기', '제목'이 등장하는 기사



✓ 일반기사 2,455,114개

✓ 낚시기사 33,622개

2 데이터

2. 데이터 정보

- 네이트 뉴스 사이트 : 2010년 1월 ~ 2017년 12월 (8년간)
- > 하루에 30000~40000개
- > 8년간 약 8천만~9천 만개의 기사
- > 1초 3~4개 기사(조건부 크롤링)
- > 1일 = 3시간
- > 24시간 = 8일
- 댓글이 달려 있는 기사만 크롤링
- > 24건 중 1건이 댓글이 달리는 기사



✓ 매우 힘들었다.

2 데이터

	A	B	C	D	E	F	G
1	date	cateogory	title	contents	best_comrn	none_best	comment
2							
3	2018-03-02	연예	'해투3' 구하	[헤럴드PO 베스트 댓글	일반 댓글	otr****]	03.0
4	2018-03-02	연예	서우, 지금	사진=온라인 커뮤니티	일반 댓글	anff****]	03.0
5	2018-03-02	사회	이상은 회장	이명박 전 베스트 댓글	일반 댓글	jing****]	03.0
6	2018-03-02	연예	'오늘쉴래요	[엑스포츠뉴스 임지연	일반 댓글	suom****]	03.
7	2018-03-02	사회	소환 조사	(서울=연합 베스트 댓글	일반 댓글	ezpa****]	03.
8	2018-03-02	스포츠	'증축보다	스포츠 기사sCtgr_cd	일반 댓글	sado****]	03.
9	2018-03-02	연예	[TV온에어]	[티브이데일리 장수정	일반 댓글	ssia****]	03.0
10	2018-03-02	연예	'썰전' 유시	'썰전' 유시 베스트 댓글	일반 댓글	puha****]	03.

2 데이터

- 카테고리 별 낚시성 기사 수

IT/과학	경제	사회	세계	스포츠	연예	정치	최신뉴스	칼럼
668	2492	4356	1475	6331	11755	1873	3145	38



스포츠, 연예 카테고리만 선택 -> 20000건



기사를 직접 보면서 낚시성 기사 체크

2 데이터

- 카테고리 별 낚시성 기사 수



실제 낚시성 기사 => 10000건



일반기사 10000건, 낚시성 기사 10000건 데이터 셋 확보

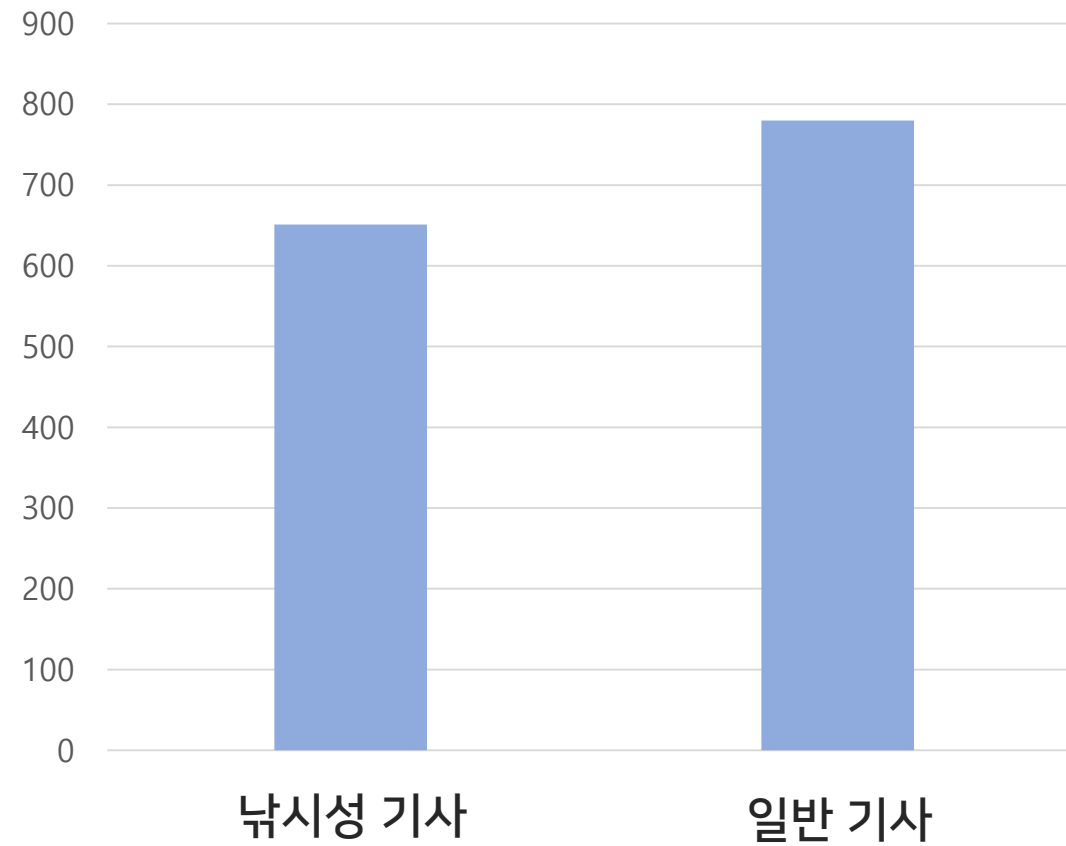


전처리 과정(불필요한 텍스트 제거, 형태소 분석, 불필요한 품사 제거, 스타밍 등등)

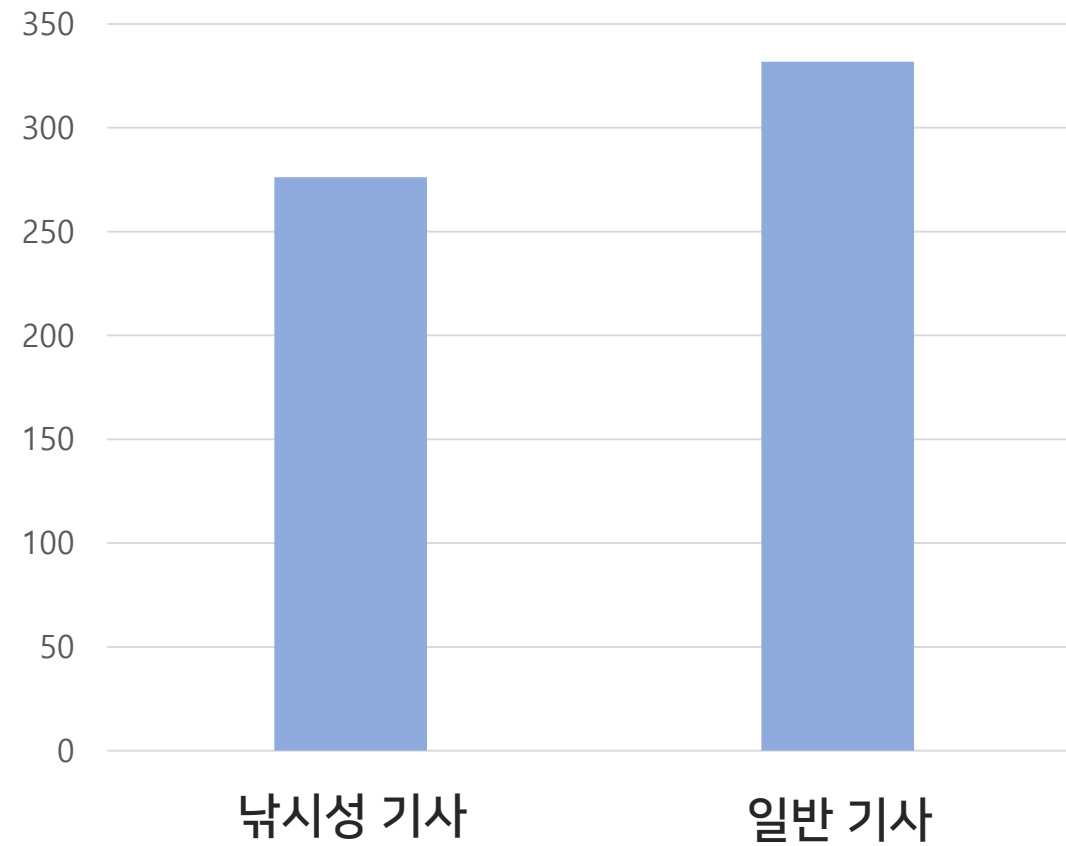
3

Feature 추출

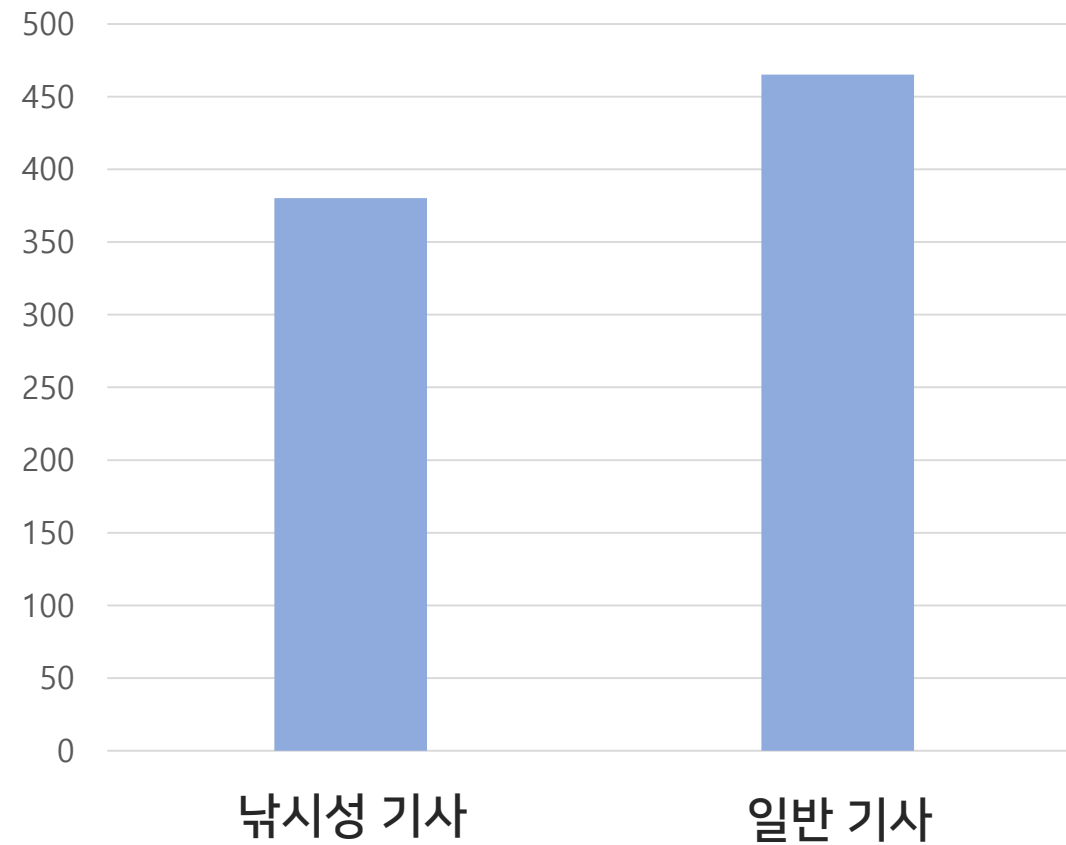
3 Feature 1-1 제목과 본문의 일치도



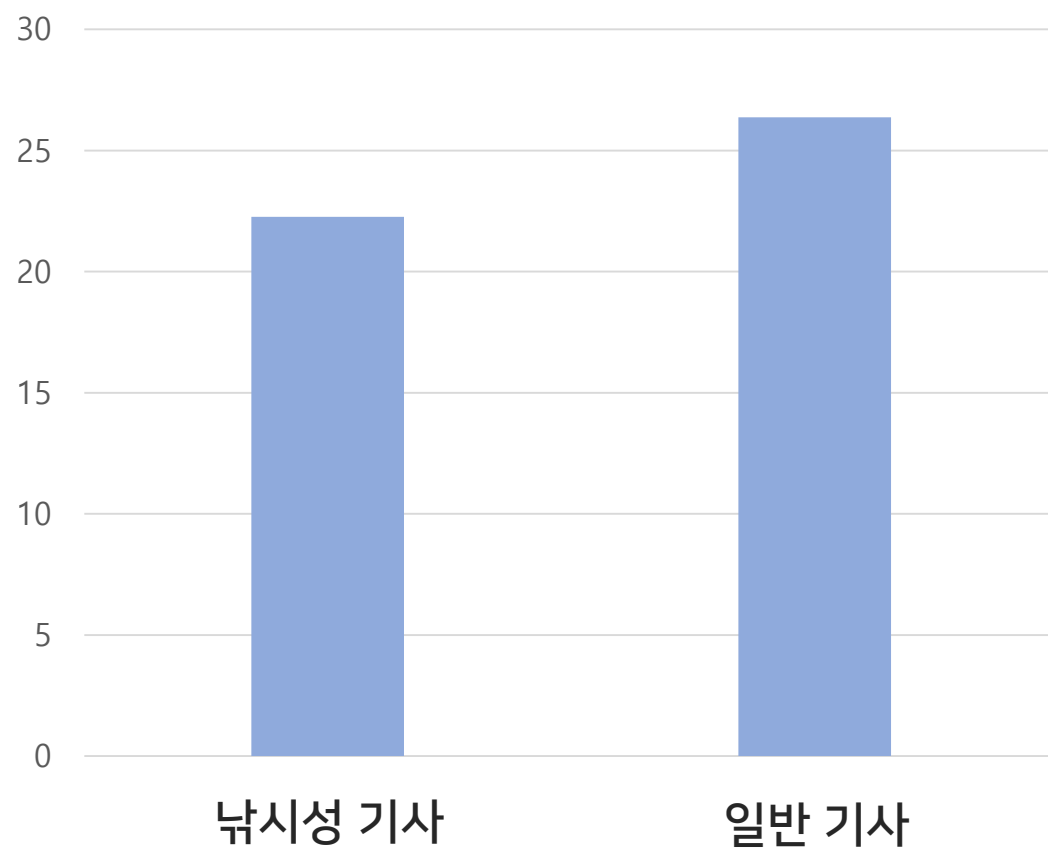
3 Feature 1-2 Bigram을 사용한 제목과 본문의 일치도



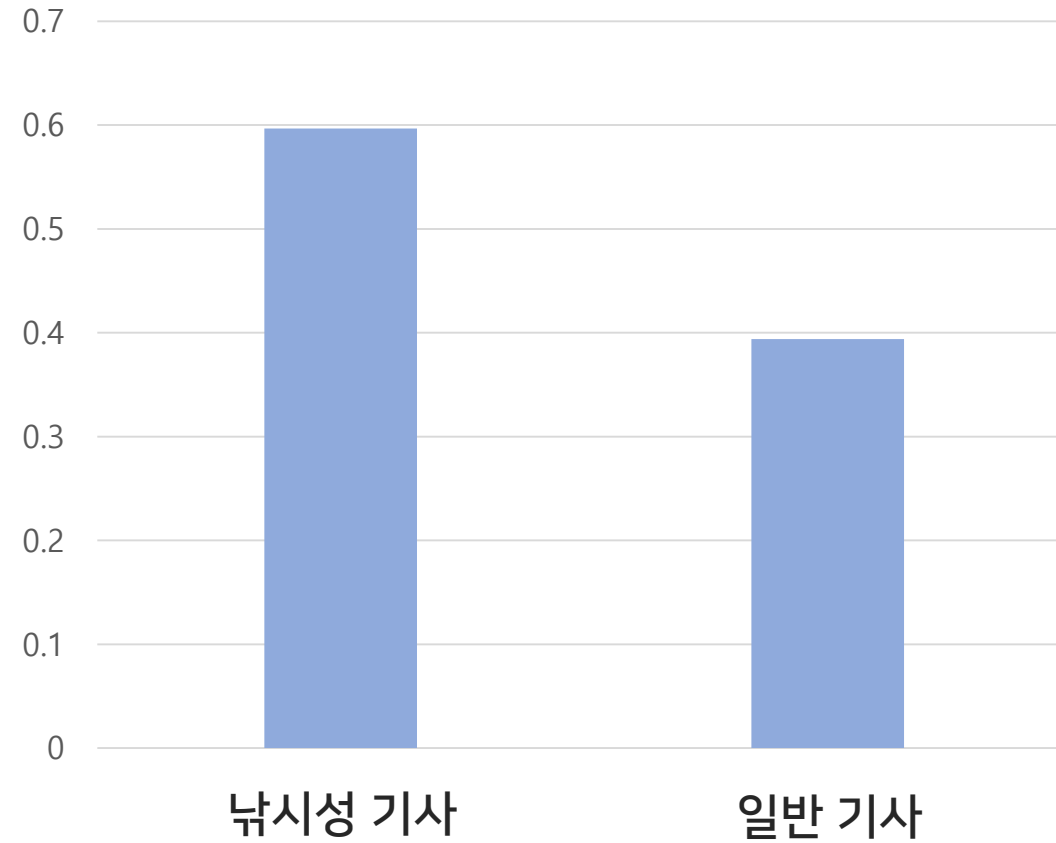
3 Feature 2-1 본문의 단어 수 비교



3 Feature 2-2 본문의 문장 수 비교

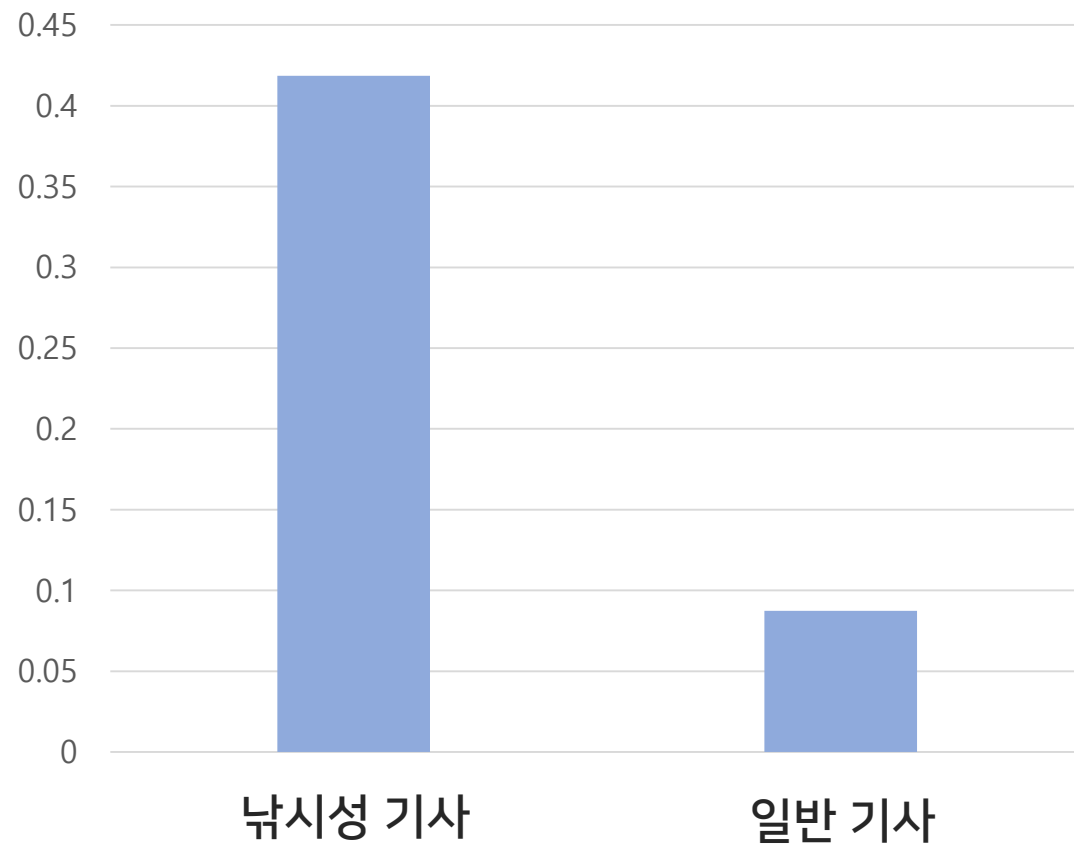


3 Feature 3 말 줄임표 (…), 물음표, 느낌표 수 비교



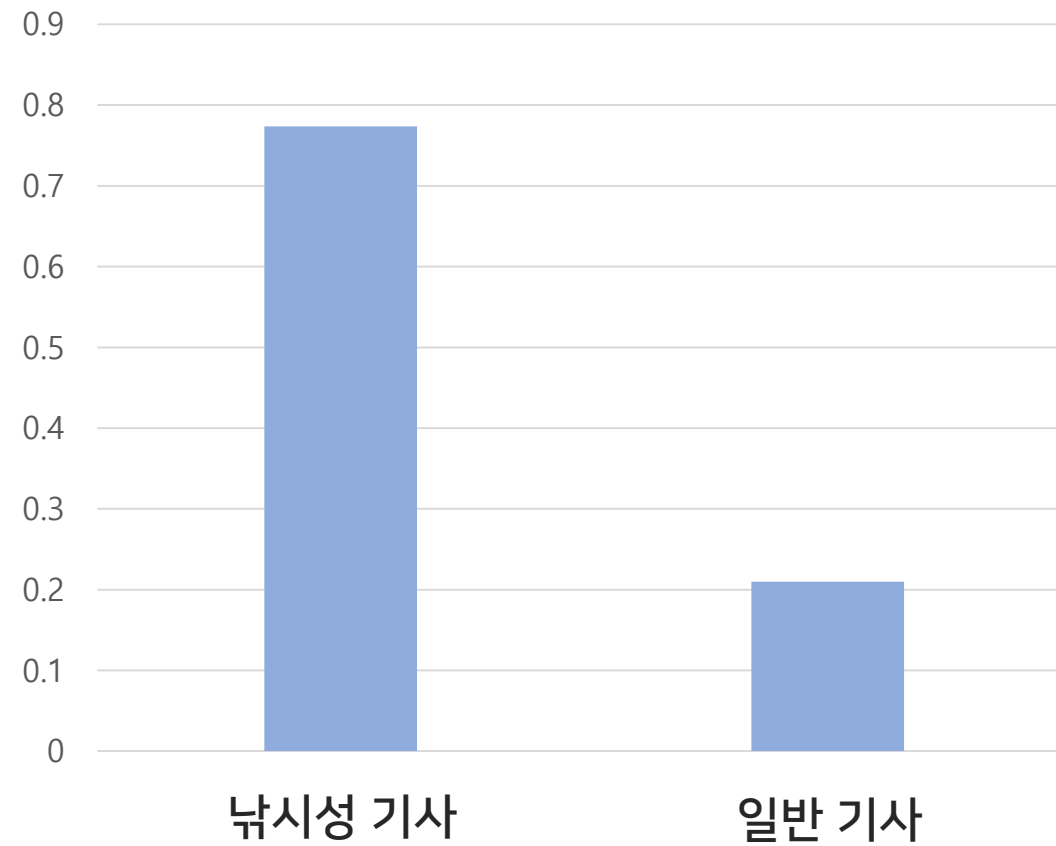
3

Feature 4 선정적 단어 사용 빈도



* 선정적 단어 및 단순 감탄사 목록 : 충격, 속보, 고백, 깜짝, 경악, 파격, 노출, 폭발, 폭로, 진짜, 근황, 분노, 위기, 반전, 비키니, 초토화

3 Feature 5 기자 이름



3

Feature 6

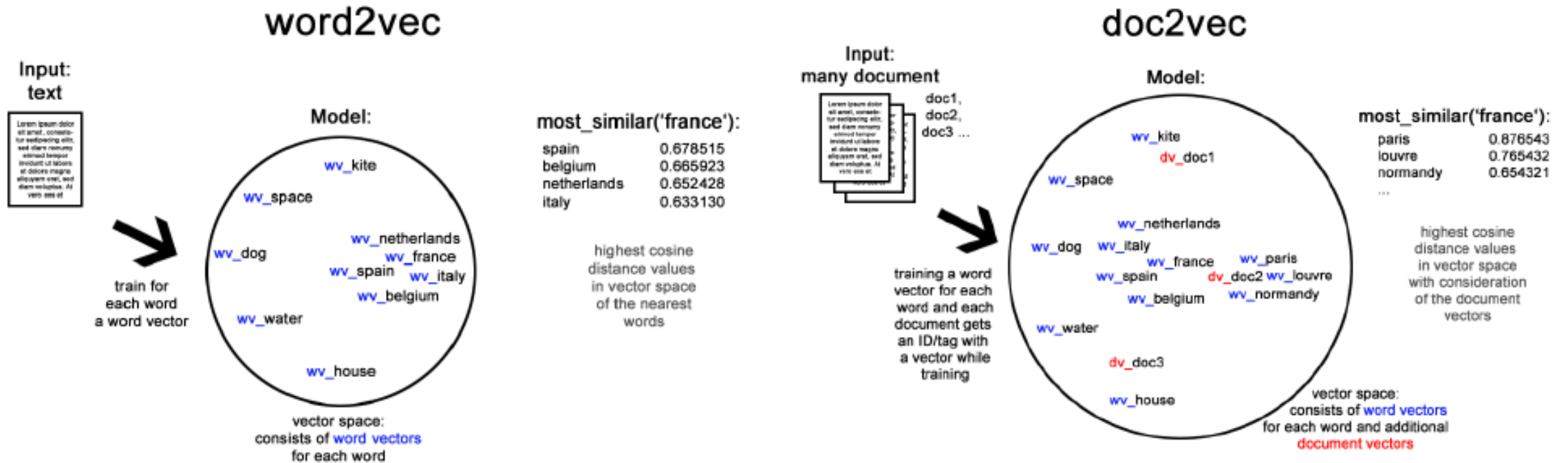
Doc2vec(워드임베딩)을 활용한 제목과 본문의 유사도 수치 비교

- ✓ 워드임베딩 = 텍스트 -> 벡터화(수치화)
- ✓ 학습하는 데이터 => 사전

Ex) 단어 간 유사도 분석 시
=> 사과, 풋사과(기존 사전)
=> 사과, 맛없다(Doc2vec 사전)

Doc2vec 성능 => 파라미터 조절 + **데이터의 양**

3 Feature 6 Doc2vec(워드임베딩)을 활용한 제목과 본문의 유사도 수치 비교



3 Feature 6 Doc2vec(워드임베딩)을 활용한 제목과 본문의 유사도 수치 비교

```
TaggedDocument(words=["유승옥", "대학교", "장학금", "받아", "쌈겨풀", "수술", "성형", "고백"], tags=['sent56']),
```

```
TaggedDocument(words=["뉴스", "기사", "02", "▲", "해피투게더", "엑스포츠", "뉴스", "원", "민준", "기자", "유승옥", "이", "대학교", "때",  
"장학금", "을", "받아", "쌈겨풀", "수술", "을", "했", "밝혔", "2", "일", "방송", "된", "2", "해피투게더", "3", "는", "수지", "서우", "제시",  
"유승옥", "최현석", "이", "출연해", "추천", "특집", "으로", "꾸며", "이", "날", "유승옥", "은", "들이", "성형수술", "에", "대해서", "물어보  
자", "눈", "을", "했", "고", "솔직하게", "수술", "한", "사실", "을", "공개", "했", "유승옥", "은", "눈", "을", "보시", "면", "지금", " 짹  
이", "다", "라고", "수술", "부작용", "을", "털어", "이제", "는", "진짜", "수술", "안", "하고", "싶다", "고", "개인", "인", "바람", "을", "드  
러냈", "이", "에", "김신영", "은", "수술", "을", "한", "시기", "가", "언제", "인지", "물어봤", "유승옥", "은", "대학교", "1", "학년", "때",  
"장학금", "받아", "수술", "을", "했", "고", "말해", "현장", "을", "웃음", "바다로", "만들었", "대중문화", "부", "enter@portsnews.com", "사  
진", "해피투게더", "유승옥", "㊂", "방송", "화면", "▶", "하니", "가슴", "만진", "몬유", "에", "당황", "진담", "▶", "길건", "김태우", "나  
를", "동물원", "원숭이", "처럼", "만들", "협박", "▶", "이병헌", "이민정", "특남", "우리", "아이", "에게", "만큼은", "제발", "...", "▶", "허  
지웅", "예원", "욕설", "한마디", "로", "갈냐", "와", "비슷해", "▶", "임창정", "열애설", "상대", "임은경", "에", "들이", "대고", "싶", "만  
큼", "...", "㊂", "엑스포츠", "뉴스", "무단", "전재", "및", "재", "배포", "금지", "0", "기사", "이미지", "0"], tags=['sent10066'])
```

✓ 문장 간 유사도 분석이 가능하다

3 Feature 6 Doc2vec(워드임베딩)을 활용한 제목과 본문의 유사도 수치 비교

2만개 데이터를 가지고 Doc2vec 학습 -> 정확도가 현저히 낮음

2만개 데이터 Doc2vec 학습 -> 2시간

200만개 정도의 데이터 학습이 필요함 -> 장비나 시간적으로 한계



피처를 뽑아내지 못함

3 Feature

[illegible]

4

모델링

4 모델링 결과

- ✓ 일반기사 10000개
- ✓ 낚시기사 10000개
- ✓ Train 80%, Test 20%

Naïve Bayes	Gaussian
정확도	0.81
재현율	0.76

4 모델링 결과

- ✓ 일반기사 10000개
- ✓ 낚시기사 10000개
- ✓ Train 80%, Test 20%

SVM	cost=1, gamma=1/6	cost=0.1, gamma=10	cost=0.5, gamma=10
정확도	0.86	0.75	0.82
재현율	0.84	0.94	0.87

4 모델링 결과

- ✓ 일반기사 10000개
- ✓ 낚시기사 10000개
- ✓ Train 80%, Test 20%

Neural Network	Hidden layer=1	Hidden layer=10
정확도	0.83	0.84
재현율	0.87	0.84

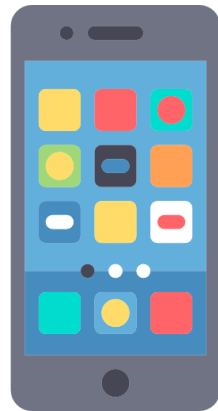
5

결론 및 활용방안

5 결론 및 활용방안

✓ 결론 및 보완점

1. Excellent 모델은 아니지만 Good 모델
2. Pyqt or App의 형태로 GUI화하면 좋을 것 같음
3. Doc2vec을 구현할 수 있는 환경이 되면 훨씬 정밀한 모델을 만들 수 있을 것
4. Feature 5 때문에 모델 성능 유지가 어려움



5 결론 및 활용방안

✓ 활용방안

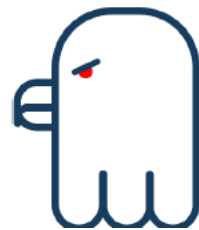
1. 기존 포털 사이트 뉴스서비스에 적용
-> 뉴스 게시 전 사전차단

2. 프리미엄 서비스 제공(포털사이트)

- > 월 정액제
- > 아이디로 로그인
- > 악성 댓글 및 낚시성 기사, 게시물들이
블라인드 or 차단되는 서비스

You**Tube**Red

매의 눈 _
악성 댓글 분류 시스템



감사합니다