블로그 포스팅 토픽 모델링

데이터분석 캡스톤디자인 2017100898 산업경영공학과 박혜원

목차

공통부분

Unsupervised Learning

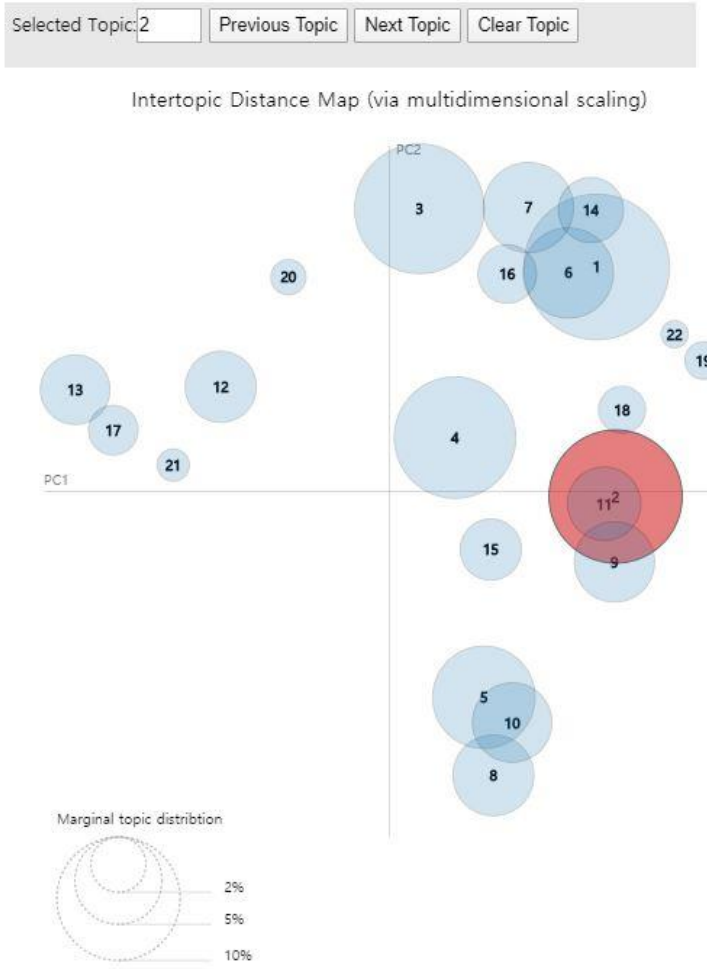
Data

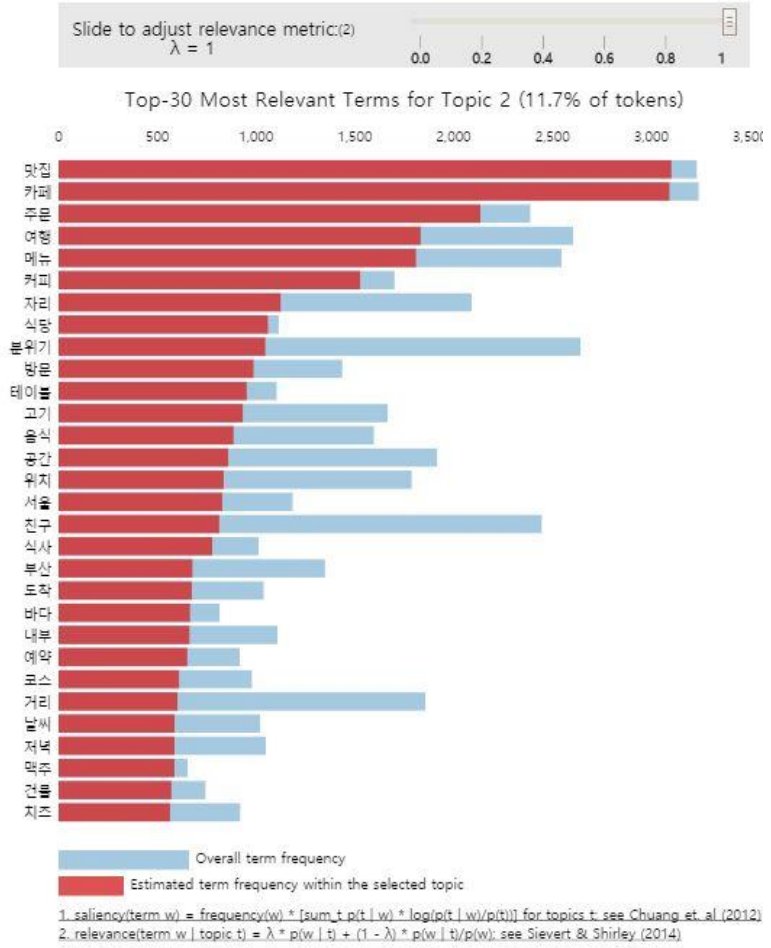
내 블로그 1개 타 블로그 131개
총 **18개 주제**, **132개 블로그**, **4021개 포스팅** 데이터 사용
명사만 사용

Input Data

LDA : CountVectorizer - **Tfidf Vectorizer** (등장 빈도, 중요도 높은 1400개 단어 사용)
K-means : CountVectorizer - **L2 Normalization** (문서-단어 유클리드 거리 정규화, Tfidf Vectorizer 사용했을 때 결과 잘 나오지 않는 문제 발생, 약 30000개 단어 사용)

LDA 토픽모델링





Flash

LDA 토픽모델링

Beauty Fashion	Cash	Car	Drama Movie	Music	Travel
피부	시장	차량	영화	앨범	여행
컬러	투자	디자인	작품	활동	호텔
메이크업	종목	공간	사랑	차트	노래
쿠션	가입	모델	드라마	발매	사랑
핑크	적금	자동차	아이	음원	비행
렌즈	금리	실내	배우	신곡	사진

Sports	IT	Cook	Game	Delicious	Pet Baby
경기	노트북	재료	게임	맛집	아이
시즌	로지텍	요리	이벤트	카페	운동
선수	마우스	레시피	플레이	주문	개월
리그	작업	고기	모바일	사진	엄마
득점	모델	양념	캐릭터	메뉴	아기
기록	보드	소스	서버	커피	강아지

LDA 토픽모델링

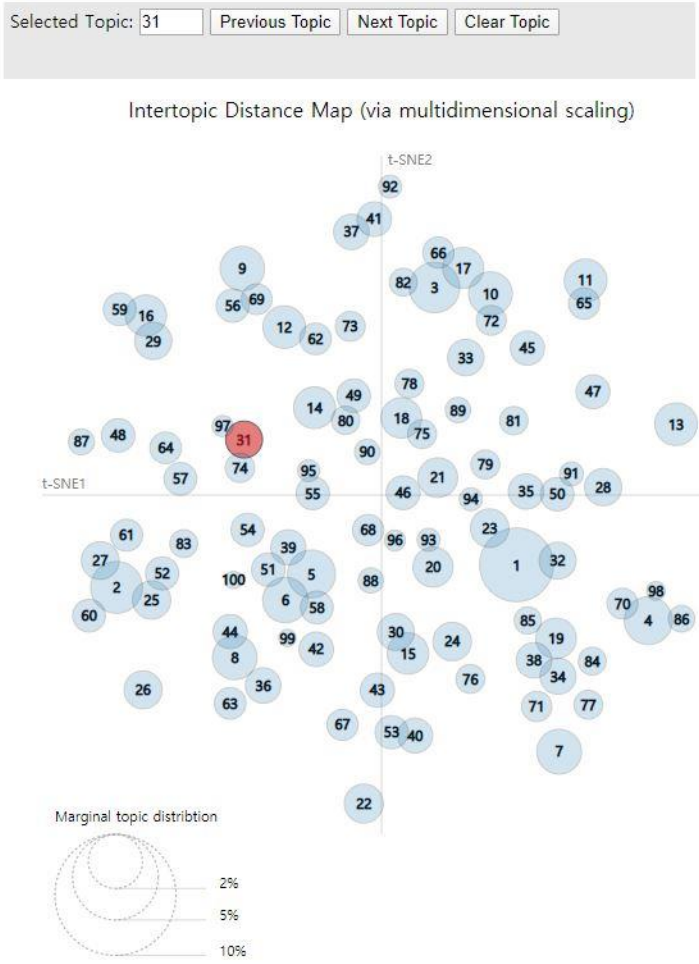
	index	id	title	topic	most common	every topic
105	105	aquatic_moon	나도 모르게 돈 쓰고 나오는 강릉 소품샵 강릉문방구	7.0	0.5690	[(4, 0.07115061), (5, 0.013500582), (6, 0.1144...
106	106	aquatic_moon	예쁜 앤틱 가구가 가득한 강릉시내카페 년임당방앗간	7.0	0.9182	[(0, 0.061039496), (2, 0.012413235), (7, 0.918...
107	107	aquatic_moon	동탄 신도시 홍익돈까스 이게 어떻게 인분인가요	7.0	0.6061	[(0, 0.06834652), (5, 0.30997226), (7, 0.60612...
108	108	aquatic_moon	영통 초밥 맛집 주원초밥 앤 참치	7.0	0.7986	[(1, 0.03012346), (5, 0.15369861), (7, 0.79861...
109	109	aquatic_moon	음악추천 꼭 챙겨 듣는 세븐틴 노래 수록곡 중심으로	13.0	0.7804	[(5, 0.020642232), (8, 0.025509091), (12, 0.14...
110	110	aquatic_moon	대충 살자 야식으로 간장밥 볶아 먹는 나처럼	1.0	0.8518	[(0, 0.033200912), (1, 0.85182023), (7, 0.0308...
111	111	aquatic_moon	권선동미용실 감동을 전하다 에서 클리닉하다	17.0	0.3367	[(2, 0.09771612), (7, 0.30644253), (17, 0.3366...
112	112	aquatic_moon	수원 인계동 시 맛집 수원시청역 북촌손만두 피냉면 먹었어요	7.0	0.6713	[(1, 0.13237222), (7, 0.6712791), (10, 0.02238...
113	113	aquatic_moon	일상 월의 기억 홍콩 마카오 여행 정리하기 센트럴 구경하기	7.0	0.5224	[(0, 0.34678847), (4, 0.036284294), (6, 0.0446...
114	114	aquatic_moon	아이패드 블루투스 키보드 퓨전에프앤씨 아이노트 후기	21.0	0.3280	[(4, 0.04934391), (7, 0.016988032), (9, 0.2172...
115	115	aquatic_moon	엔드라이브에서 발견한 것들	9.0	0.4835	[(5, 0.16412263), (9, 0.48348778), (11, 0.0689...
116	116	aquatic_moon	신상 미니곡물크림샌드 먹어 보기	18.0	0.4190	[(5, 0.31070375), (17, 0.25879407), (18, 0.418...
117	117	aquatic_moon	리뷰 내가 제일 좋아하는 수원 영통역 배달 떡볶이 떡깨비	7.0	0.4411	[(1, 0.25122476), (5, 0.23776017), (7, 0.44112...
118	118	aquatic_moon	일상 월의 기억 홍콩 마카오 여행 정리하기	7.0	0.3739	[(0, 0.34805942), (2, 0.111977346), (7, 0.3738...
119	119	aquatic_moon	일상 월의 기억 홍콩 마카오 여행 정리하기	7.0	0.5932	[(0, 0.2953831), (2, 0.0337281), (7, 0.5932247...
120	120	aquatic_moon	일상 요즘 사는 방법	7.0	0.4742	[(2, 0.26963365), (5, 0.16250327), (6, 0.01221...
121	121	aquatic_moon	음악추천 지극히 내 취향인 여자 아이돌 여돌 노래 추천	13.0	0.7694	[(6, 0.09701217), (12, 0.06941996), (13, 0.769...
122	122	aquatic_moon	제주 뷰가 예쁜 서귀포 흑돼지 맛집 난드르바당	7.0	0.7239	[(1, 0.20382965), (7, 0.7239402), (12, 0.06755...
123	123	aquatic_moon	음악추천 지극히 내 취향인 남자 아이돌 남돌 노래 추천	13.0	0.8180	[(5, 0.06897298), (12, 0.10596629), (13, 0.817...
124	124	aquatic_moon	천안 신부동 야우리 술집 펍 추천 다운언더	7.0	0.5394	[(5, 0.30600187), (6, 0.094300255), (7, 0.5394...

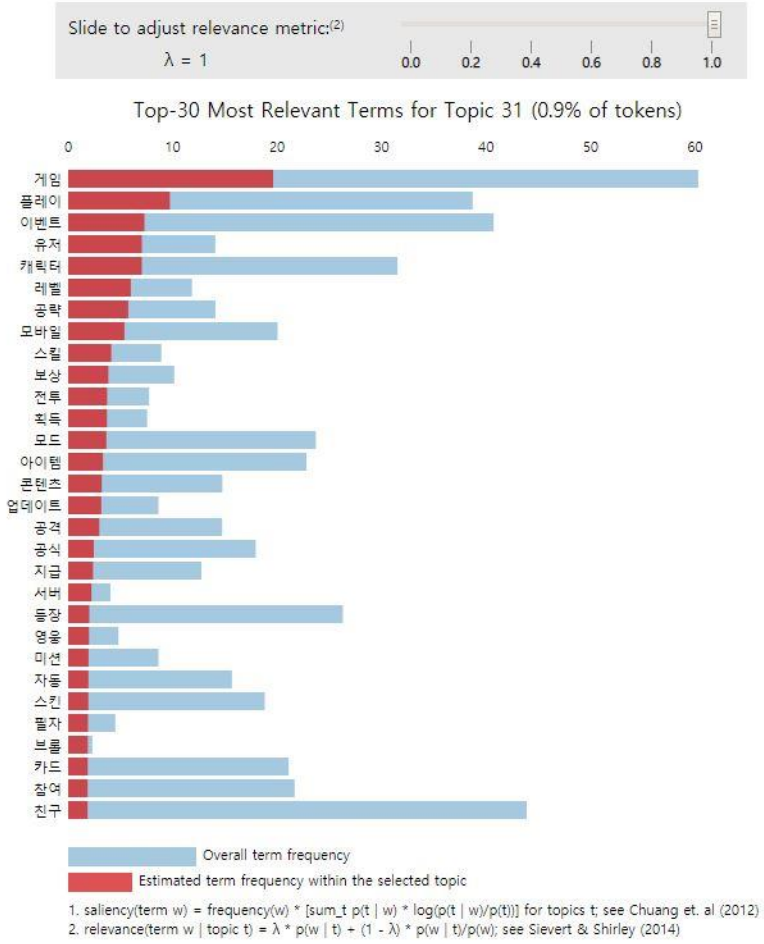
LDA로 추측한 각 포스팅의 주제

	blogger	topic_num	topic	real_topic	check
80	[redacted]	21.0	[pet, baby]	IT	False
81	[redacted]	17.0	[beauty, fashion]	beauty	True
82	[redacted]	19.0	game	game	True
83	[redacted]	18.0	[beauty, fashion]	fashion	True
84	[redacted]	18.0	[beauty, fashion]	DIY	False
85	[redacted]	2.0	[drama, movie]	drama	True
86	[redacted]	17.0	[beauty, fashion]	beauty	True
87	[redacted]	1.0	cook	cook	True
88	[redacted]	7.0	delicious	travel	False
89	[redacted]	21.0	[pet, baby]	IT	False
90	[redacted]	21.0	[pet, baby]	sports	False
91	[redacted]	7.0	delicious	car	False
92	[redacted]	16.0	[pet, baby]	pet	True
93	[redacted]	17.0	[beauty, fashion]	beauty	True
94	[redacted]	21.0	[pet, baby]	IT	False
95	[redacted]	7.0	delicious	delicious	True
96	[redacted]	8.0	[pet, baby]	pet	True
97	[redacted]	7.0	delicious	travel	False
98	[redacted]	15.0	sports	sports	True
99	[redacted]	2.0	[drama, movie]	movie	True

포스팅 주제비율로 추측한 각 블로그의 주제

K-means 토픽모델링





Flash

K-means 토픽모델링

Pet	Sports	Cash	Beauty	Music	Animation
강아지	운동	종목	컬러	노래	애니
반려견	자세	투자	렌즈	음악	애니메이션
훈련	시합	시장	피부	아이돌	칼날
사료	대회	적금	메이크업	사랑	작품
견종	경기	만원	발색	보컬	스토리
급여	선수	주식	성분	녹음	만화

Delicious	Game	Movie	Book	IT	Travel
맛집	게임	영화	작품	아이폰	비행
음식	플레이	개봉	사건	갤럭시	여행
주문	사전예약	작품	로맨스소설	노트북	투어
메뉴	모바일	배우	소설	카메라	호텔
고기	이벤트	감독	작가	로지텍	바다
떡볶이	유저	관객	여주	이어폰	리조트

K-means 토픽모델링

Drama
드라마
캐릭터
시즌
넷플릭스
미드
캐릭터

Fashion
원피스
가방
컬러
아이
디자인
브랜드

DIY
가죽
지갑
카드
바느질
가방
만들기

Car
차량
디자인
모델
출시
실내
자동차

Baby
아이
엄마
아기
개월
생후
이유식

Cook
요리
레시피
재료
만들기
간장
소스

Other
서울
자막
출처
코로나
마지막
목록

K-means 토픽모델링

	id	title	result	topic
105	aquatic_moon	나도 모르게 돈 쓰고 나오는 강릉 소품샵 강릉문방구	13	delicious
106	aquatic_moon	예쁜 앤틱 가구가 가득한 강릉시내카페 년임당방앗간	13	delicious
107	aquatic_moon	동탄 신도시 홍익돈까스 이게 어떻게 인분인가요	33	delicious
108	aquatic_moon	영통 초밥 맛집 주원초밥 앤 참치	18	delicious
109	aquatic_moon	음악추천 꼭 챙겨 듣는 세븐틴 노래 수록곡 중심으로	52	music
110	aquatic_moon	대충 살자 야식으로 간장밥 볶아 먹는 나처럼	97	cook
111	aquatic_moon	권선동미용실 감동을 전하다 에서 클리닉하다	32	beauty
112	aquatic_moon	수원 인계동 시 맛집 수원시청역 북촌손만두 피냉면 먹었어요	18	delicious
113	aquatic_moon	일상 월의 기억 홍콩 마카오 여행 정리하기 센트럴 구경하기	74	travel
114	aquatic_moon	아이패드 블루투스 키보드 퓨전에프앤씨 아이노트 후기	38	IT
115	aquatic_moon	엔드라이브에서 발견한 것들	20	other
116	aquatic_moon	신상 미니곡물크림샌드 먹어 보기	20	other
117	aquatic_moon	리뷰 내가 제일 좋아하는 수원 영통역 배달 떡볶이 떡깨비	76	delicious
118	aquatic_moon	일상 월의 기억 홍콩 마카오 여행 정리하기	74	travel
119	aquatic_moon	일상 월의 기억 홍콩 마카오 여행 정리하기	74	travel

K-means로 추측한 각 포스팅의 주제

	blogger	topic	real_topic	check
65	[redacted]	cash	cash	True
66	[redacted]	movie	movie	True
67	[redacted]	beauty	beauty	True
68	[redacted]	sports	sports	True
69	[redacted]	IT	IT	True
70	[redacted]	delicious	delicious	True
71	[redacted]	beauty	beauty	True
72	[redacted]	cash	cash	True
73	[redacted]	other	drama	False
74	[redacted]	movie	movie	True
75	[redacted]	book	book	True
76	[redacted]	delicious	delicious	True
77	[redacted]	IT	IT	True
78	[redacted]	sports	sports	True
79	[redacted]	game	game	True
80	[redacted]	IT	IT	True
81	[redacted]	beauty	beauty	True
82	[redacted]	game	game	True
83	[redacted]	fashion	fashion	True
84	[redacted]	fashion	DIY	False

포스팅 주제비율로 추측한 각 블로그의 주제

LDA vs K-Means 비교

 # 평가

LDA

주제 일관성: 0.6198

K-means

주제 일관성 평가모델 LDA에만 지원

Word2Vec 이용해 각 Cluster 내 단어끼리의 Similarity 구한 뒤 평균 점수 구함

Average Similarity: 0.4298

LDA vs K-Means 비교

 # 실제 블로그 주제와 비교

LDA

```
>>> other 추가 정답률:86 / 132  ================ 65.15151515151516%
>>> other 제외 정답률:86 / 117  ================ 73.50427350427351%
```

K-means

```
>>> other 추가 정답률:98 / 132  ================ 74.24242424242425%
```

LDA vs K-Means 비교

 # 특징

LDA

4min 51s

한 문서에 여러 종류 토픽 존재 가능하다고 가정
단어 수 1400개로 고정

K-means

15min 49s

한 문서에 한 토픽 존재 가능하다고 가정
대규모 데이터에 적용 가능
(단어 수 30000개 이상)

LDA vs K-Means 비교

 # 한계

LDA

없는 주제 많음
Pet + Baby / Drama + Movie 등
주제가 합쳐져서 나옴

반복적인 주제할당으로 Topic Modeling이
일어나므로 **매우 많은 문서 및 단어 환경에서는**
적합하지 않음

K-means

일관성 평가 기법 문제

Clustering 주제 많을 때 **K 개수 선정**의 어려움
K 매우 많은 수로 고정한다면
후처리/군집 병합 필수

후처리 할 수 없을 땐 **일일이**
후처리 및 라벨링 해야 됨
(좋지 않은 방법, 나의 역량)

LDA vs K-Means 비교

 # 내 블로그 주제 선정

LDA

맛집 (Delicious): 71
육아&애완동물 (Baby & Pet): 35
뷰티(Beauty & Fashion): 11
음악 (Music):10

주제: Delicious (맛집)

K-means

Delicious (맛집): 52
전자제품(IT):21
여행(Travel):10
음악(Music): 8
뷰티(Beauty): 7

주제: Delicious (맛집)

LDA vs K-Means 비교

 # 어떤 모델이 더 적합할까?

LDA

한 문서당 주제가 여러 개인 경우 (블로그, 트위터)
소량의 문서 / 적은 주제를 군집화 할 경우
기타(Other) 문서가 적을 경우

K-means

한 문서당 주제가 하나인 경우 (신문 기사)
다량의 문서 / 많은 주제를 군집화 할 경우
기타(Other) 문서가 많을 경우

감사합니다