단일 토픽 문서에 대한 클러스터링 기반 토픽 모델링

2018 대한산업공학회 춘계공동학술대회

단일 토픽 문서에 대한 클러스터링 기반 토픽 모델링

4 Apr, 2018

SNU Data Mining Center

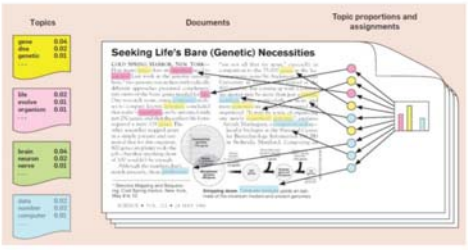
서울대학교 산업공학과

신훈식, 김현중, 조성준

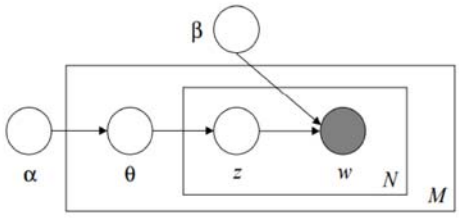
{hunsik, hyunjoong}@dm.snu.ac.kr,
zoon@snu.ac.kr

1. Introduction

- **Topic modeling – Latent Dirichlet Allocation(LDA)***
  - A representative unsupervised generative probabilistic model for text data
  - Topic is modeled as distributions over thousands of terms
  - Document is modeled as mixtures of dozen of topics



<Example of LDA result >



<Graphical model representation>

Word distribution on topics
- *topic* 1 = 0.04 * gene + 0.02 * dna + 0.01 * genetic ...
- *topic* 2 = 0.02 * life + 0.01 * evolve + 0.01 * organism ...
- ...

Topic distribution on documents
- *Document = [topic1, topic2, topic3, ... topic k]*
- *docment1* = [0.13, 0,02, 0.56, ... , 0.09]
- *docment2* = [0.41, 0,12, 0.11, ... , 0.02]
- ...

Parameters
- α = *document − topic density prior*
- β = *topic − word probability matrix*
- θ = *topic distribution*
- z = *latent variable*
- w = *word vector*
- N = *# of words in a document*
- M = *# of documents in a corora*

SNU Data Mining Center

* Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

1

1. Introduction

- **Topic modeling – Latent Dirichlet Allocation(LDA)***
  - A representative unsupervised generative probabilistic model for text data
  - Topic is modeled as distributions over thousands of terms
  - Document is models as mixtures of dozen of topics

- **Topic interpretation**
  - It is well-known that the topics inferred by LDA are **not always easily interpretable** by humans
    - ✓ $topic\ k = 0.02 * \text{'apple'} + 0.01 * \text{'evolve'} + 0.01 * \text{'computer'}$ … → *?*
  - The **problem** with interpreting topics is that **common terms often appear** near the top of multiple topics
    - ✓ $topic\ k = 0.32 *' \text{a}' + 0.11 *' \text{the}' + 0.08 *' \text{do}'$ … → *?*
  - This interpreting way makes it **hard to differentiate** the meanings of these topics.

2. Related works

1 Summarizing topical content with word frequency and exclusivity Bischof and Airoldi (2012)

- LDA에서 "topic"은 단어들의 다항 분포로 표현되기 때문에 단어들의 빈도수(**frequency**)를 기반으로 분포 확률을 학습
- 토픽을 구성하는 단어 중 발생 확률을 기준으로 단어들을 ranking하면 **common word**가 상위에 존재, 토픽 해석은 common word에 취약
- 특정 topic에만 얼마나 등장하였는지 **exclusivity**를 고려하여 빈도수와 함께 word ranking의 기준으로 설정

2 On Estimation and Selection for Topic Models Matthew A. Taddy (2011)

- 토픽 구성 단어를 '**lift**'를 기준으로 ranking하여 해당 토픽을 해석
- Common word의 등장을 줄일 수 있지만 빈도가 매우 낮은 rare word에 취약한 단점이 있음

3 LDAvis: A method for visualizing and interpreting topics Sievert and Shirley (2014

- 기존의 word ranking measure(빈도수, lift)의 선형 조합으로 새로운 measure '**relevance**'를 정의
- Relevance = $r(w,k \mid \lambda) = \lambda \log(\phi_{kw}) + (1-\lambda)\log\left(\frac{\phi_{kw}}{p_w}\right)$
- 기존 ranking 방법에 비해 common word와 rare word에 강건하여 토픽 해석에 용이
- User study를 통해 파라미터 **λ**의 최적값을 0.6으로 설정

4 Topic modeling: beyond bag-of-words Wallach, Hanna M (2011)

- 단어의 순서는 텍스트의 의미를 해석하는데 중요한 역할
- **연어(collocation)**를 추가하여 토픽의 효과적인 해석 시도
- 기존의 graphical model에서 추정해야 할 파라미터가 증가하여 계산 비용이 높아지는 단점

3. Data

- **Data :** 2015. 01 ~ 2015. 12 기간에 Reuters sports 뉴스기사 **(# of news = 7,950)**
  - 다른 텍스트 데이터에 비해 길이가 짧고 문서 내에 **하나의 요약된 주제**를 포함하고 있는 특징
  - 구성된 단어들로 해석하기 위해 토픽의 의미가 보다 친숙하고 명확한 **스포츠 뉴스를** 주 데이터로 사용

	headline	content
0	IAAF appeals to CAS over Russian doping bans	LONDON The world governing body for athletics ...
1	NBA player accuses Milwaukee-area jeweler of r...	MILWAUKEE Milwaukee Bucks forward John Henson ...
2	Viennese waltz in the desert for Dubai leader ...	DUBAI Jan 29 A dynamic sequence of five birdie...
3	Autograph hunter Hoffman signs for early Maste...	AUGUSTA, Georgia Charley Hoffman began his day...
4	F1 teams must agree on calendar and older engines	LONDON Formula One teams will have to reach un...
5	Mercedes apologise to stunned Hamilton	MONACO Mercedes bosses queued up to apologise ...
6	Monaco just does not suit our car, say Williams	MONACO Williams said the tight and twisty Mona...
7	Jaguars back in London for fourth year running...	LONDON The Jacksonville Jaguars will return to...
8	Pats center out, Seattle tackle questionable f...	The New England Patriots will be without start...
9	Blackhawks tame Wild to move 2-0 clear in series	(The Sports Xchange) - Right winger Patrick Ka...
10	Blatter faces bigger challenge at divided FIFA	ZURICH It was like time had stood still.\nAs S...
11	Platini, romantic or pragmatist?	BERNE One of the most exquisitely gifted socce...
12	Luis Enrique will stay, says Barcelona president	BARCELONA Barcelona president Josep Maria Bart...
13	U.S. name German Vogts as technical advisor	Former Germany coach Berti Vogts has been appo...
14	FIFA officials booed before World Cup trophy c...	VANCOUVER FIFA officials were booed by the cro...

< 2015 reuters sports 뉴스기사 >

3. Data

- **Data representation : Bag-of-words(BoW) & Term frequency inverse document frequency(TF-IDF)**
  - BoW : 텍스트 데이터를 등장하는 단어의 빈도수를 사용하여 순서와 상관없이 수치형 벡터로 표현
  - TF-IDF : BoW로 표현된 빈도수에 document frequency(df)를 고려하여 반비례하게 수치를 보정

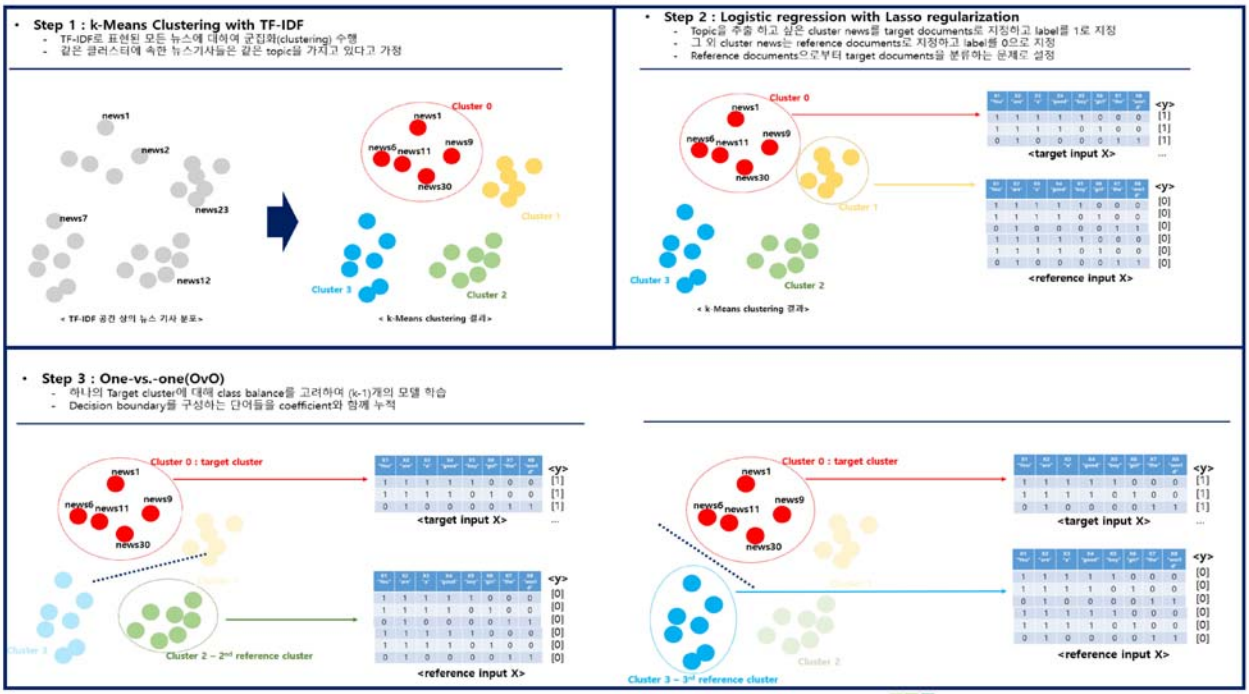
- **Bow 예시**

Headline : IAAF appeals to CAS over Russian **doping** bans
Content:
LONDON The world governing body for athletics has appealed to the Court of Arbitration for Sport (CAS) over punishments handed out by Russia's anti **doping** agency (RUSADA) to six of its athletes. The athletes, including 2012 Olympic 3,000 metres steeplechase champion Yuliya Zaripova and Olympic 50km walking champion Sergey Kirdyapkin, have all been banned by RUSADA. However, the International Association of Athletics Federations (IAAF) disagrees with the "selective" disqualification of results.

X1 "world "	X2 "doping"	X3 "have"	X4 "the"	X5 "hope"	X6 "Olympic"	...	Xn "growth"
1	2	1	2	0	2	...	0

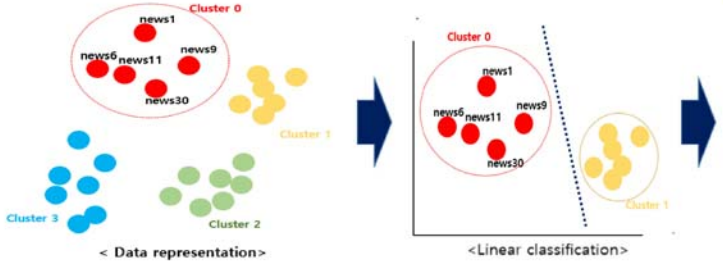
- **News** → * Vector representation of document [1, **2**, 1, 2, 0, 2, ... 0] (BoW) → [2.7, **0.9**, 1.1, 0, 0.4, 0.5, ... 0] (TF-IDF)

4. Proposed Method



4. Proposed Method

- **Experiment Setting**
  1. **Data format**
     - Bag-of-words(all tokens, tokens without stopwords, only nouns) + TF-IDF transformation
     - Target documents = each news cluster
     - Reference documents = all other news clusters
  2. **Clustering :** k-means clustering(n_cluster=11, max_iteration=300, random_stated=1234)
  3. **Classification algorithm** : Logistic regression(tolerance = 0.01) + Lasso penalty + One-vs.-one
- **Baseline** : LDA-vis model with λ = 0.6
- **Process of proposed method**



index	words	coefficient
0	open	0.4575
1	set	0.3112
2	seed	0.3101
3	djokovic	0.2051
4	slam	0.2003
5	number	0.1716
6	murray	0.1153
7	serena	0.0813
8	federer	0.0505
9	wimbledon	0.0012
10	finals	0.0009

<Features with β != 0>

5. Result

- **Topic words(all tokens used)**

topic id	topic words
1	.. .. the, ", I, to, ", he, it, and, that, of, is, in ,s
2	the, ..., in, her, a, to, she, match, and, set, s, of, on, was
3	the, ..., of, to, and, in, said, that, by, ", olympic, ", for, on
4	the, .. a, ..at, golf, to, on, tour, and, I , pga, of, round, in
5	fifa, the, .. of, .. soccer, blatter, to, and, in, presidnet, and, committee, said, by
6	the, .. in, team, and, .. league, of, and, will, players, cup, to, for, $
7	the, .. a, in, game, .. to, yards, quaterback, and, with, goal, on, sattle, touchdown
8	points, quarter, guard, nba, the, rebounds, warriors, cleveland, assists, james, .. scored, curry, forward, game
9	-, :, hernandez, .. •, police, lloyd,), (, murder, his, was, trial, died, university
10	france, froome, tour,), de, stage, (, rider, spain, team, riders, race, germany, rossi, netherland
11	formula, mercedes, race, ferrari, car, driver, mclaren, bull, vettel, cars, red engine, drivers, races, baldwin

<LDA-vis : Topic words>

topic id	topic words
1	fifa, soccer, blatter, president, officials, 2022, marketing, to, switzerland,
2	formula, prix, race, mercedes, mclaren, car, ferrari, hamilton, it, season, to,
3	doping, olympic, games, international, olympics, iaaf, ioc, wada, committee,
4	golf, tour, lot, this, he, his, masters, it, week, back, champion, you, woods,
5	nfl, bowl, patriots, brady, super, fantasy, quarterback, the, draftkings,
6	said, team, his, football, national, reporting, statement, reported, race,
7	open, his, set, seed, nadal, court, federer, mathc, slam, serve, djokovic, wimbledon, murray, tennis, grand
8	par, birdies, under, hole, round, tour, holes, shot, birdie, bogey, three, 67, 68, second, open
9	her, she, open,set, slam, so, williams, match, sharapova, serena, halep,
10	world meters, gold race, champion, seconds, olympic, her, women, medal, downhill, the, olympics, bronze, his
11	game, the, scored, points, minute, win, cleveland, goal, season, we, games, their, lead, came baseman

<Proposed method : Topic words>

5. Result

- **Topic words(tokens without stopwords)**

topic id	topic words
1	.. .., I, ", ", s, round, two, one, first, open, (,), it, back
2	.. .., ", ", I, said, s, ', we, would, n't, team, think, it, want
3	.. .., athletes, said, doping, ", ", the, athletics, iaaf, hernandez, russia,), (, s
4	fifa, blatter, .. soccer, .. president, platini, committee, corruption, swiss, u.s., said, officials, s, zurich
5	.. nfl, .. league, patriots, bowl, football, brady, super, said, players, university, new, the, uarterback
6	points, .. game, scored, goal, .. yard, quarter, gauard, minute, touchdown, minutes, goals, half, rebounds
7	-,), (, .. :, season, league, .. series, •, runs, manager, chicago,
8	gold, champion, world, seconds, meters, .. medal, .. france, silver, olympic, women, s, race, bolt
9	formular, race, hamilton,mercedes, ferrari, prix, ... car, rosberg, driver, mclaren, bull races, team
10	mathc, djokovic, murray, 6-3, 6-4, slam, set, seed, federer, wimbledon, williams, tennis, grand 6-2, 7-6
11	games, city, host, olympics, olympic, bid, ioc, rugby, stadium cup, tokyo, winter, summer, .. 2024

<LDA-vis : Topic words>

topic id	topic words
1	world, cup, team, race, match, players, club, hockey, states, finals, france, sky, statement, tour, minute
2	olympics, olympic, games, rio, international, host, ioc, party, costs, committee, beijing, event tokyo government, 2020
3	nfl, said, bowl, university, lawsuit, football, league, contract, washington, companies, murder, baseball, seattle, attorney, rodriguez
4	baseball, league, pitcher, manager, cubs, hits, inning royals, era, runs, run, baseman, season, blue, innings
5	world, olympic, gold, champion, race, emters, silver, slalom, seconds, beijing, pharoah, american, freestyle, downhill, gatlin
6	set, open, mathc, slam, wimbledon, seed, final, serve, title, federer, round, court, murray, finals, nadal
7	fifa, soccer, blatter, president, corruption, zurich, officials, cup, swiss, america, request, bribes, concacaf, warner, 2022
8	formula, prix, race, ferrari, mercedes, car, driver, season, team, future, renault, briton, massa, laps, baldwin
9	points, nba, games, quarterback, win, season,nhl, goal, second, conference, stanley, quarter, touchdown, first, point
10	doping, athletics, iaaf, coe, wada, moscow, agency, athletes, russian, banned committee, anti, olympic, farah, report
11	par, under, holes, golf, tour, open, round, masters, major, shot, hole, championship, mcilroy, whistling, spieth

<Proposed method : Topic words>

5. Result

- **Topic words(only noun tokens)**

topic id	topic words
1	game, points, season, coaach, team, bowl, nba, yards, super, quarterback, seattle, bleague, new, quarterback, patriots
2	olympic, athletes, ioc, games, sports, olympics, officials, russia, federation, president, sports, bid, committee, u.s., international
3	golf, pga, round, open, tour, woods, spieth, masters, course, holes, day, mcilroy, hole, champion, birdies
4	match, set, open, djokovic, slam, murray, seed, federer, wimbledon, williams, tennis, nadal, number, court, serve
5	race, world, formula, hamilton, mercedes, iaaf, champion, ferrari, champion, prix, rosberg, meters, seconds, one, races
6	nfl, players, league, sports, new, football, bull, school, police, court, university, state, teams, judge, case
7	fifa, blatter, soccer, president, platini, committee, football, zurich, sepp, qatar, uefa, ethics, election, corruption, body
8	cup, women, world, united, canada, gold, states, japan, team, australia, rugby, vancouver, rio, minute, goal
9	hernandez, baseball, runs, league, series, manager, ining, cubs, inings, hits, lloyd, home, mets, baseman, royals
10	team, tour, france, race, stage, froome, ryder, rider ,spain, sky, captian, presidents, riders, bt, cycling
11	fight, mayweather, pacquiao, backhand 7-6, fedexcup, chambers, baseline, boxing, vegas, floyd chelsea, nicklaus, rod, las

<LDA-vis : Topic words>

topic id	topic words
1	formula, prix, bull, mclaren, ferrari, race, car, driver, mercedes, circuit, team, rosberg, india,f1, hamilton
2	athletics, iaaf, doping, athletes, wada, moscow, anti, stubbs, blood, coe, russia, allegation, vladimir, kenya, sergei
3	djokovic, nadal, murray, federer, set, wawrinka, seed, open, andy, match, year, aces, finals, atp, court
4	olympic, games, committee, rio, sports, olympics, beijing, ioc, government, bach, bid, tokyo, janeiro, international, pyeongchang
5	holes, spieth, round, golf ,masters, open, champion, pga, ko ,woods, shots, hole, birdies, tour, number
6	bisiness, match, league, team, cup, nhl, marco, leon, arrest, las, police, eagles, los, country, alberto
7	fifa, president, blatter, officials, soccer, corruption, committee, platini, authorities, football, conmebol, sports, joshua, blazer, qatar
8	serena, williams, seed, sharapova, open, wimbledon, halep, finals, set, match, court, tennis, slam, wta, muguruza
9	nba, points, rebounds, guard, quarter, shots, james, cleveland, warriors, basketball, golden, wizards, istanbul, game, clippers
10	nfl, quarterback, bowl, yards, seahawks, pass, super, patriots, passes, game, colts, head, brady, broncos, receiver
11	world, championships, old, olympic, meters, beijing, seconds, slalom, gatlin, downill, olympics, gymnastics, jump, beaver, vonn

<Proposed method : Topic words>

5. Result

- **Document classification**
  - 토픽을 구성하는 단어들이 문서 내용을 잘 나타내는지를 확인하기 위해 해당 단어들을 변수로 설정하여 문서 분류 작업 수행
  - 2015 reuters 'politics', 'technology' 및 'health' 뉴스를 사용, 각각의 분야를 class label로 사용
  - 사용한 모델 : Logistic regression(LR), Decision tree(DT), Support vector machine(SVM) with 5-fold cross validation
  - **Methods**
    - LDA(tf-idf) : LDA 토픽을 구성하는 단어들을 변수로 사용하여 tf-idf를 input X로 사용
    - LDA(topic distr.) : 각각의 문서가 토픽의 분포로 표현되어 토픽 분포를 input X로 사용
    - Proposed method : 제안한 방법으로 토픽 단어들을 선택, 해당 단어들을 변수로 사용하여 tf-idf를 input X로 사용

method / model	LDA(tf-idf)	LDA(topic distr.)	Proposed method
Logistic regression	0.77 (+/- 0.10)	0.91 (+/- 0.09)	0.90 (+/- 0.10)
Decision Tree	0.70 (+/- 0.09)	0.87 (+/- 0.10)	0.85 (+/- 0.10)
Support Vector machine	0.76 (+/- 0.11)	0.91 (+/- 0.09)	0.89 (+/- 0.12)

<Document classification results>

✓ 제안한 방법의 단어들이 기존 LDA 토픽 구성 단어보다 높은 문서 분류 정확도 성능을 가짐
→ 제안한 방법으로 선택된 단어들이 뉴스 기사 내용을 더 잘 표현

✓ Tf-idf 보다 문서 분류 정확도 성능이 좋다고 알려진 LDA topic distribution과 비슷한 성능을 가짐

5. Result

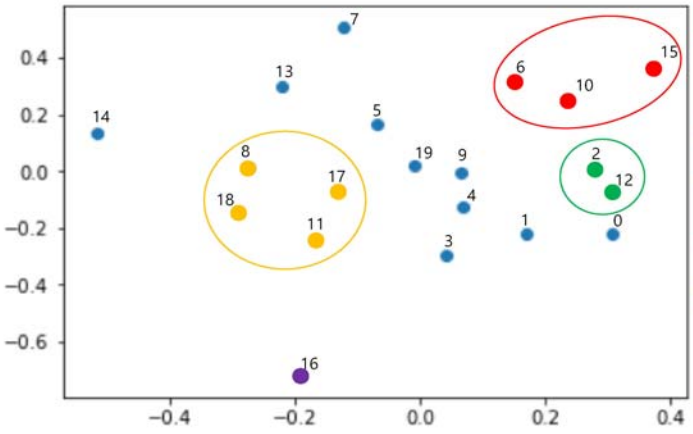
- **Multi-words representation**
  - 제안하는 방법의 input X는 tf-idf 형태이기 때문에 bigram, trigram의 적용이 가능
  - 연어(collocation)은 단일 단어보다 풍부한 정보를 전달

topic id	topic words
1	cup, officers, sports, match, draftkings, league, grand slam, team, speculation, basketball, woman, tony
2	league, major, inning, series, baseball, season, home run, game, major league
5	formula, prix, mercedes, grand prix, race, car, ferrari, formula one, hamilton, driver, circuit, red bull, season, drivers, bull
6	nfl, quarterback, hernandez, national football, football, yards, seahawks, bowl, super bowl, patriots, receiver, nfc, cowboys, steelers, passes colts
7	fifa, blatter, president, officials sports, soccer, president jeffrey, officials, committee, prosecutors, soccer world, football, uefa, corruption, caribbean, spokesman
9	nba, points, basketball, rebounds, guard, cleveland, all star, warriors, game ,knicks, coach, center, james, quarter, golden state, games, draft
11	iaaf, athletics, doping, athletes, wada, moscow, agency, jack stubbs, coe, anti doping, russia, code, moscow russia, drug, athletics federation

<Proposed method with bigram>

5. Result

- **Visualization**
  - 클러스터의 중심 좌표를 토픽의 좌표로 생각하여 Multidimensional scaling(MDS) 를 통해 2차원으로 축소하여 토픽 시각화
  - 토픽을 세분화 했을 때, 위치에 따른 토픽 의미 관계를 확인 할 수 있음



FIFA

Topic 06 : fifa, president, committee, corruption, investigation
Topic 10 : sccoer, fifa, corruption, extradition, indictment
Topic 15 : fifa, blatter, election, ethics, association, praag

Tennis

Topic 02 : set, open, Djokovic, nadal, federrer, seed, serve
Topic 12 : serena, set, open, Sharapova, finals, match

Major sports leagues

Topic 08 : quarterback, nfl, bowl, game, yard, touchdown
Topic 11 : points, nba, guard, shots, rebounds, basketball
Topic 17 : league, nhl, period, goal, balckhawks, hockey
Topic 18 : league, inning, run, pitch, hitter, shortstop

[Ex-nfl player Hernandez's murder incident]
Topic 16 : Hernandez, murder, Massachusetts, parole, lloyd

6. Conclusion

- 제안한 방법은 전처리에 따라 토픽 구성 단어가 크게 변화하지 않고 common words 및 불용어의 포함이 적음
- 모델 구조 변화나 추가적인 계산 비용 없이 토픽 구성 단어에 multi-word 적용이 쉬움
- 문서 분류 정확도를 통해 제안한 모델이 LDA보다 전반적인 문서의 내용을 잘 반영하는 단어들을 선택
- 문서 분류 정확도 성능이 좋다고 알려진 LDA topic distribution과 비슷한 성능을 가짐
- 클러스터의 중심 좌표를 차원 축소하여 토픽들의 위치 관계를 시각화 할 수 있고 거리에 따른 의미 관계를 확인할 수 있음

7. Reference

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

[2] Sievert, Carson, and Kenneth E. Shirley. "LDAvis: A method for visualizing and interpreting topics." *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. 2014.

[3] Bischof, Jonathan, and Edoardo M. Airoldi. "Summarizing topical content with word frequency and exclusivity." *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. 2012

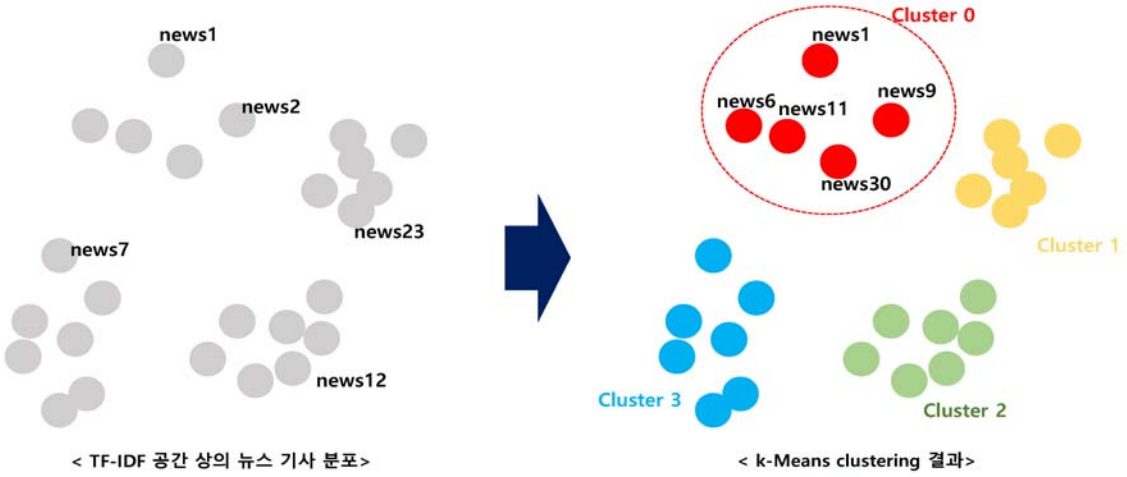
[4] Chuang, Jason, Christopher D. Manning, and Jeffrey Heer. "Termite: Visualization techniques for assessing textual topic models." *Proceedings of the international working conference on advanced visual interfaces*. ACM, 2012.

[5] Chuang, Jason, et al. "Topic model diagnostics: Assessing domain relevance via topical alignment." *Proceedings of the 30th International Conference on machine learning (ICML-13)*. 2013

+. Appendix – Proposed Method

+. Proposed Method

- **Step 1 : k-Means Clustering with TF-IDF**
  - TF-IDF로 표현된 모든 뉴스에 대하여 군집화(clustering) 수행
  - 같은 클러스터에 속한 뉴스기사들은 같은 topic을 가지고 있다고 가정



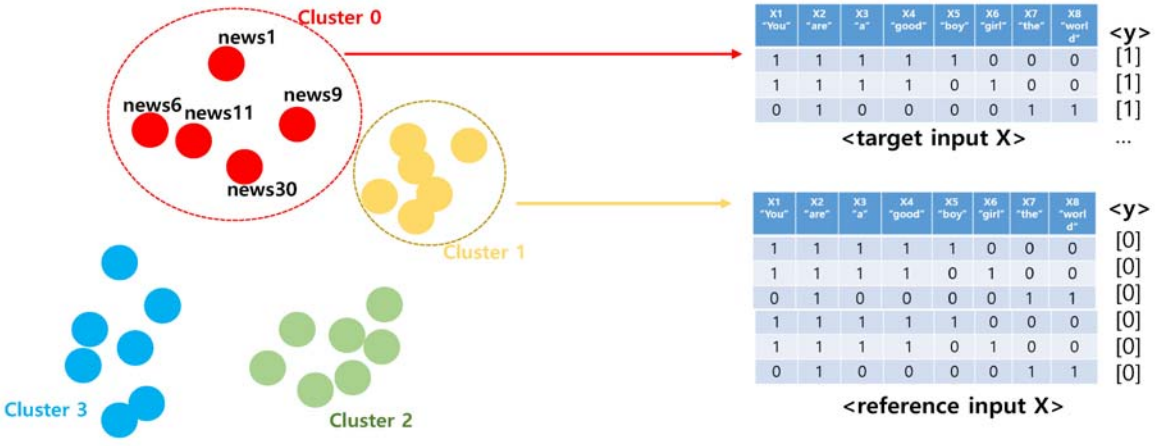
< TF-IDF 공간 상의 뉴스 기사 분포>

< k-Means clustering 결과>

SNU Data Mining Center

+. Proposed Method

- **Step 2 : Logistic regression with Lasso regularization**
  - Topic을 추출 하고 싶은 cluster news를 target documents로 지정하고 label를 1로 지정
  - 그 외 cluster news는 reference documents로 지정하고 label를 0으로 지정
  - Reference documents으로부터 target documents을 분류하는 문제로 설정



X1 "You"	X2 "are"	X3 "a"	X4 "good"	X5 "boy"	X6 "girl"	X7 "the"	X8 "world"	<y>
1	1	1	1	1	0	0	0	[1]
1	1	1	1	0	1	0	0	[1]
0	1	0	0	0	0	1	1	[1]

<target input X>

...

X1 "You"	X2 "are"	X3 "a"	X4 "good"	X5 "boy"	X6 "girl"	X7 "the"	X8 "world"	<y>
1	1	1	1	1	0	0	0	[0]
1	1	1	1	0	1	0	0	[0]
0	1	0	0	0	0	1	1	[0]
1	1	1	1	1	0	0	0	[0]
1	1	1	1	0	1	0	0	[0]
0	1	0	0	0	0	1	1	[0]

<reference input X>

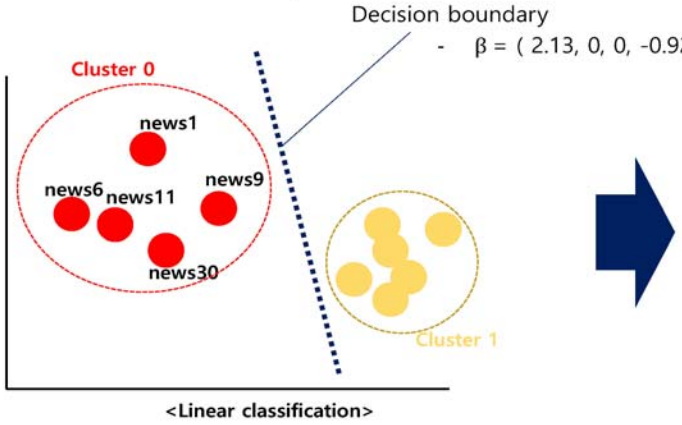
< k-Means clustering 결과>

SNU Data Mining Center

+. Proposed Method

- **Step 2 : Logistic regression with Lasso regularization**
  - Topic을 추출 하고 싶은 cluster news를 target documents로 지정하고 label를 1로 지정
  - 그 외 cluster news는 reference documents로 지정하고 label를 0으로 지정
  - Reference documents으로부터 target documents을 분류하는 문제로 설정

- Decision boundary를 구성하는 변수 중에 계수가 양수인 변수에 해당하는 단어 추출



<Linear classification>

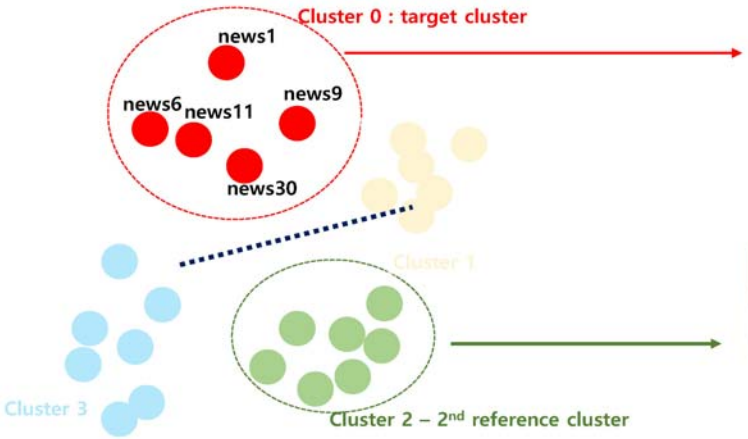
index	words	coefficient
0	open	0.4575
1	set	0.3112
2	seed	0.3101
3	djokovic	0.2051
4	slam	0.2003
5	number	0.1716
6	murray	0.1153
7	serena	0.0813
8	federer	0.0505
9	wimbledon	0.0012
10	finals	0.0009

< Features with β>0 >

SNU Data Mining Center

+. Proposed Method

- **Step 3 : One-vs.-one(OvO)**
  - 하나의 Target cluster에 대해 class balance를 고려하여 (k-1)개의 모델 학습
  - Decision boundary를 구성하는 단어들을 coefficient와 함께 누적



X1 "You"	X2 "are"	X3 "a"	X4 "good"	X5 "boy"	X6 "girl"	X7 "the"	X8 "world"	<y>
1	1	1	1	1	0	0	0	[1]
1	1	1	1	0	1	0	0	[1]
0	1	0	0	0	0	1	1	[1]

<target input X>

...

X1 "You"	X2 "are"	X3 "a"	X4 "good"	X5 "boy"	X6 "girl"	X7 "the"	X8 "world"	<y>
1	1	1	1	1	0	0	0	[0]
1	1	1	1	0	1	0	0	[0]
0	1	0	0	0	0	1	1	[0]
1	1	1	1	1	0	0	0	[0]
1	1	1	1	0	1	0	0	[0]
0	1	0	0	0	0	1	1	[0]

<reference input X>

SNU Data Mining Center

+. Proposed Method

- **Step 3 : One-vs.-one(OvO)**
  - 하나의 Target cluster에 대해 class balance를 고려하여 (k-1)개의 모델 학습
  - Decision boundary를 구성하는 단어들을 coefficient와 함께 누적