



AI Speaker Market Analysis

Unstructured Data Analysis
2018.06.14



이주현
최현석
정하은

Table of Contents

1. Project Overview
2. Data
3. Text Analysis
4. Conclusion

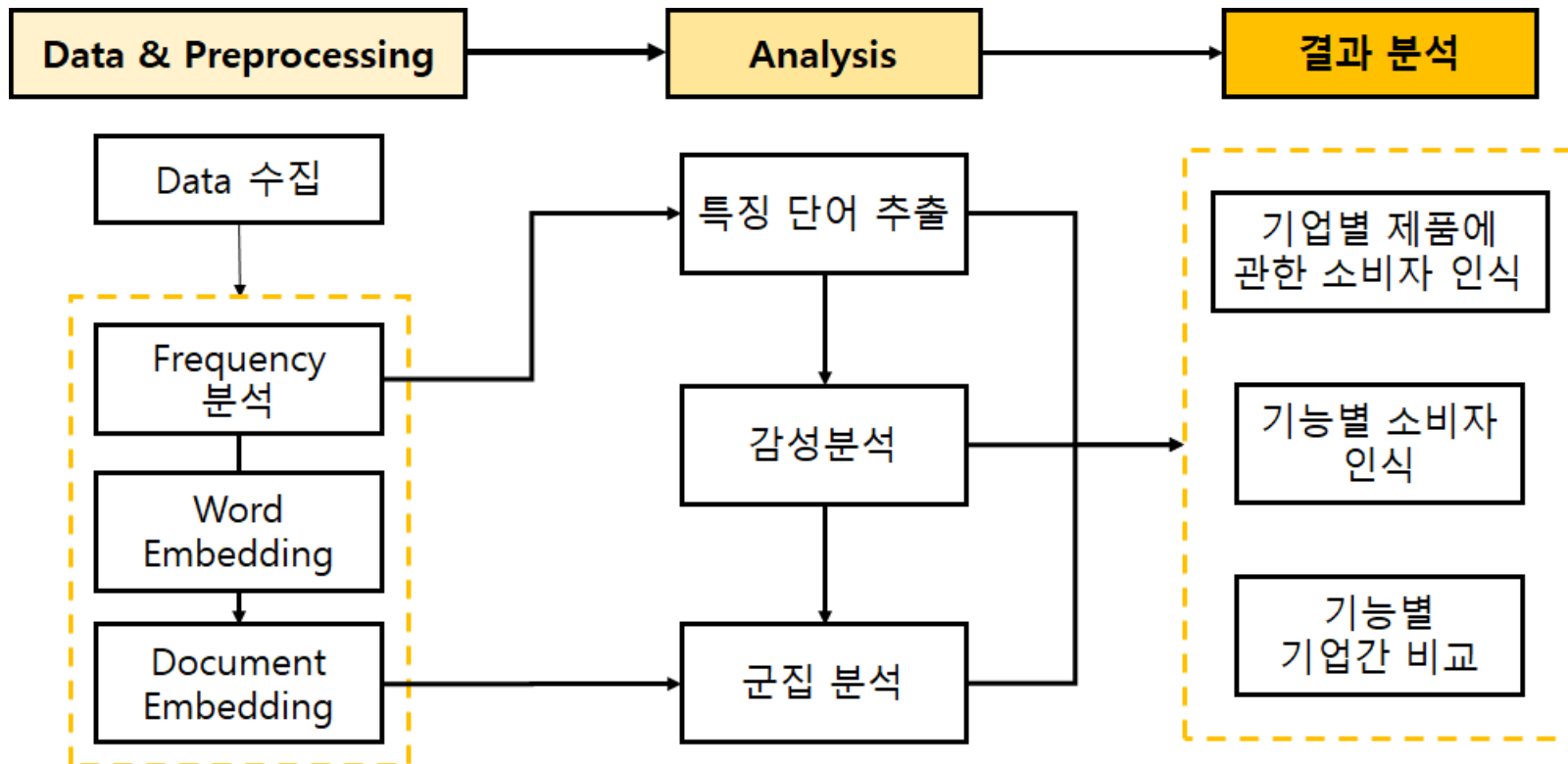
Project Goal

소비자리뷰를 활용한 시장분석

- 소비자 반응 이해
- 마케팅
- R&D
- 경쟁사 비교

* AI 스피커 시장 점유율이 가장 높은 Google Home과 Amazon Echo에 대한 리뷰 분석

Process/Algorithm



Data

Data

Product Line-up

Google 제품 (리뷰 20K, 블로그 4)

- Google Home
- Google Mini



Amazon 제품 (리뷰 20K, 블로그 8)

- Amazon Echo
- Amazon Echo Dot



Data Distribution: Google & Amazon Reviews

Google (Bestbuy)	
rating	n
1	3968 (1.03%)
2	92578 (24.18%)
3	57027 (14.89%)
4	81813 (21.37%)
5	147491 (38.52%)
Total	382,877

Amazon (Bestbuy)	
rating	n
1	2112 (0.7%)
2	2347 (0.8%)
3	11358 (4.07%)
4	66684 (23.89%)
5	196606 (70.44%)
Total	279,107



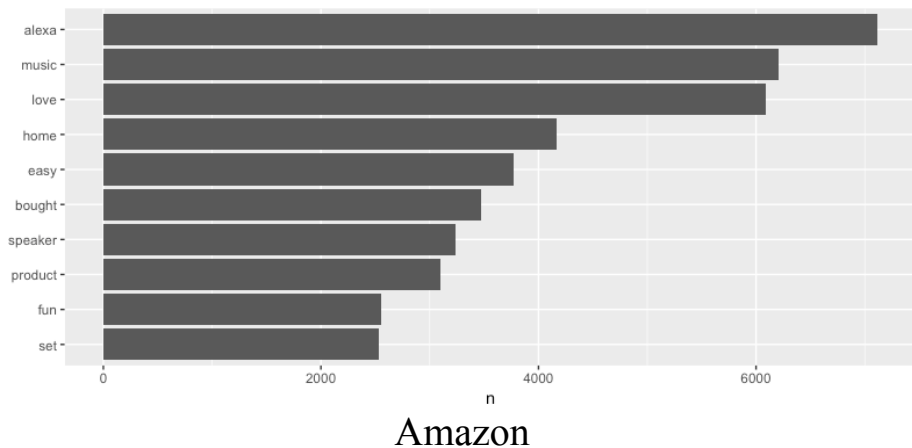
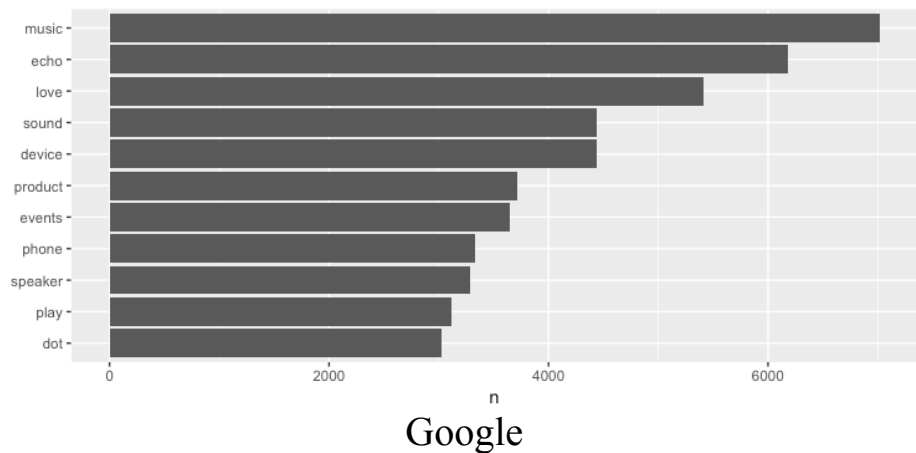
Bestbuy Sentiment		
Sentiment	Google	Amazon
Positive	39947 (74.3%)	38783 (83.86%)
Negative	13817 (25.7%)	7462 (16.14%)
Total	53764	46245

Bing Lexicon

별점별 토큰 분포
전반적으로 긍정적인 리뷰가 다수임.

Text Analysis

Word Frequency

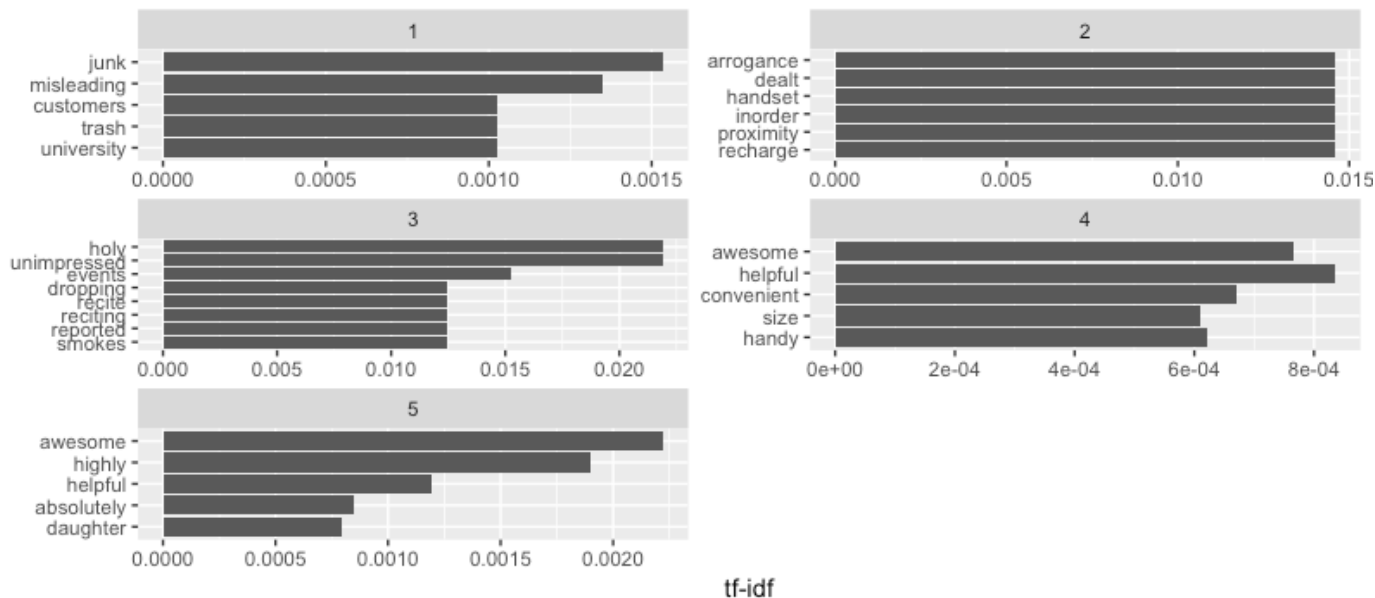


음악 관련 키워드 (music, sound, play, speaker) 빈도수가 상대적으로 높음

음악 듣기에 많이 사용한다라고 추측 할 수 있음

TF-IDF: Google

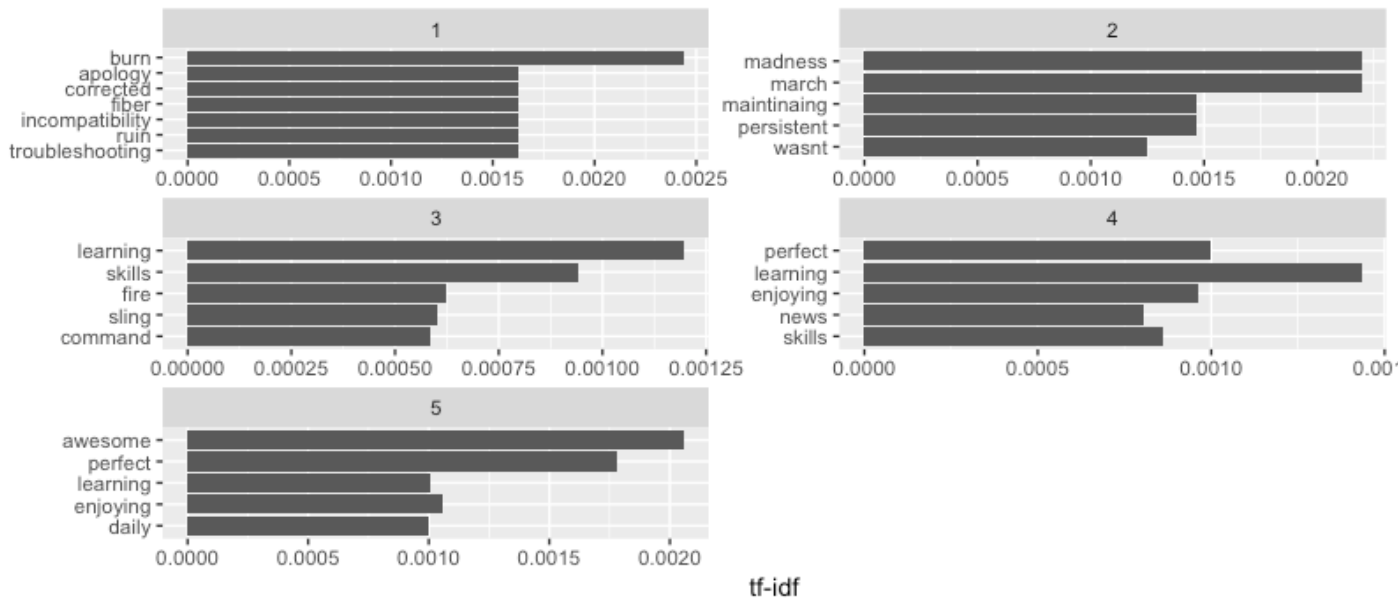
별점 별 TF-IDF



부정 적 단어 junk, trash, unimpressed 등 별점 1~3 고루 분포되어 있음.
긍정 적 단어 awesome, helpful, convenient 등 편의성을 강조한 단어들.

TF-IDF: Amazon

별점 별 TF-IDF



강한 부정 단어 burn, incompatibility, troubleshooting, madness (별점 1~2)
편의성에 이어 news, daily, enjoying 등 사용성을 강조하는 단어.

Google: N-gram

Bigram Sample

bigram	rating	n
Christmas gift	5	167
calendar events	3	1417
Day events, set time	3	708
chromecast audios	2	708

- 아마존 제품과 구글 제품 비교 빈도수가 높음 (별점별 Echo dot 또는 Amazon echo 언급)
- 음악 (Youtube, Spotify, Chromecast 등) 감상을 위한 사용
- Calendar 및 스케줄 관리에도 용이함
- 소비자들의 스피커 사용 또는 구매 동기들을 파악 할 수 있음
- 기업의 마케팅 방향과 일치하는 제품 사용 (기업 마케팅 시 캘린더, 음악 강조)

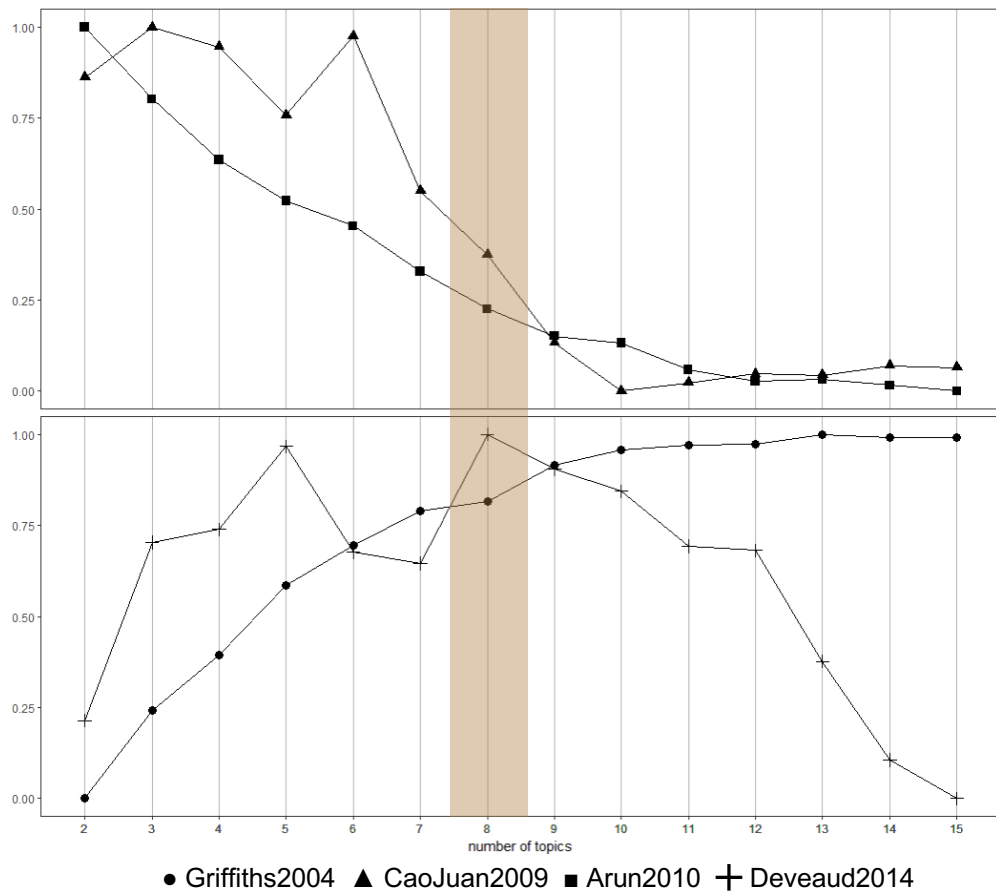
Amazon: N-gram

Smart Home Sample	bigram	rating	n
	Smart home	5	776
	Smart home	4	229
	Smart home	3	38

- 아마존 소비자 역시 구글 제품과의 비교 빈도수가 높음 (별점별 home 또는 home mini 언급)
- 음악 감상 관련 bigram 이 많음
 - 별점 3, 4, 5 공통으로 play music, bluetooth speaker, sound quality 관련 bigram
- Smart home 용도로 스피커를 구매한 다수의 소비자

Topic Modeling

- 12개의 기술 블로그를 이용해 이슈를 파악
- 기술 블로그는 1,519개의 단어를 포함
- 8개의 Topic 이 최적임을 확인



Topic Modeling

Topic 1	Speaker	Topic 5	Assistant
Topic 2	function with other device	Topic 6	Design
Topic 3	Application / OS	Topic 7	Display
Topic 4	IoT	Topic 8	Audio

Topic Modeling



토픽 1
Speaker



토픽 2

Device

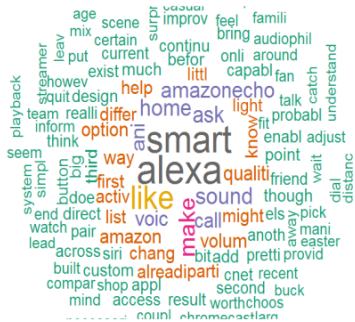


토픽 5
Assistant

Topic Modeling



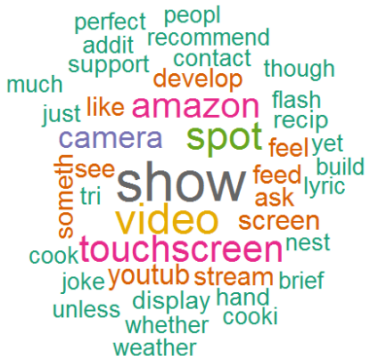
토픽 3



토픽 4



토픽 6



토픽 7



토포픽 8

Topic Modeling

Blog별 Topic에 배정될 확률 확인. 대체로 균등하게 분포한 것을 확인

	Blog1	Blog2	Blog3	Blog4	Blog5	Blog6	Blog7	Blog8	Blog9	Blog10	Blog11	Blog12
Topic1	0.00001	0.00002	0.00002	0.00013	0.00015	0.99989	0.99850	0.000015	0.00001	0.13000	0.145687	0.00002
Topic2	0.00001	0.00002	0.00002	0.00013	0.00015	0.00016	0.00002	0.000015	0.99914	0.13000	0.145687	0.00002
Topic3	0.00001	0.00002	0.00002	0.00013	0.00015	0.00016	0.00002	0.000015	0.00001	0.13000	0.99898	0.00002
Topic4	0.00001	0.00002	0.00002	0.9999	0.00015	0.00016	0.00002	0.99990	0.00001	0.13000	0.145687	0.00002
Topic5	0.00001	0.00002	0.00002	0.00013	0.00015	0.00016	0.00002	0.000015	0.00001	0.99991	0.145687	0.99987
Topic6	0.00001	0.00002	0.99989	0.00013	0.99990	0.00016	0.00002	0.000015	0.00001	0.13000	0.145687	0.00002
Topic7	0.00001	0.99893	0.00002	0.00013	0.00015	0.00016	0.00002	0.000015	0.00001	0.13000	0.145687	0.00002
Topic8	0.99993	0.00002	0.00002	0.00013	0.00015	0.00016	0.00002	0.000015	0.00001	0.13000	0.145687	0.00002

Clustering

Google과 Amazon 각각의 리뷰에 대해 Embedding된 공간에서의 **군집분석** 실시

- 최적의 군집수 도출
- 별점과 군집간 비교 분석: 별점이 높은 리뷰 중 부정적 내용은 기업에게 **Insight** 제공할 수 있음

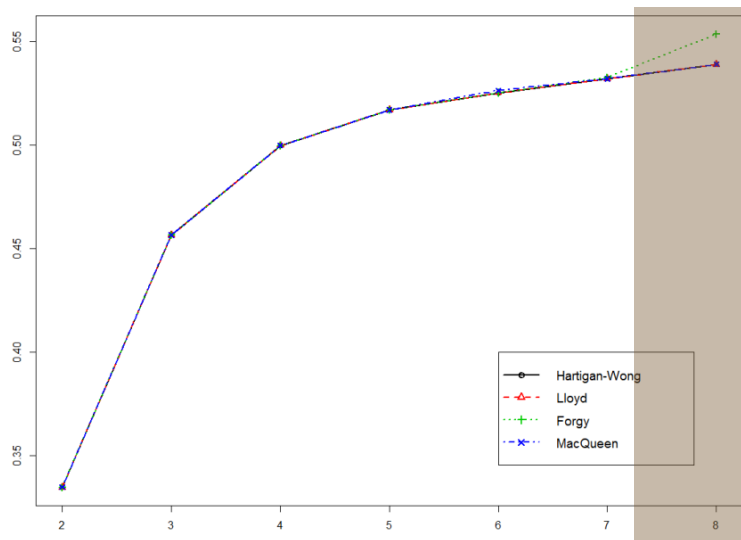
기업별 리뷰데이터		별점	
		Positive	Negative
군집분석 결과	Positive		
	Negative	Insight	

Clustering

- Google에 대한 리뷰는 21,623건이 존재하며 별점은 다소 불균형적인 분포를 이룸
- 21,623건의 리뷰에 대하여 전처리한 결과, 총 6,438건의 단어로 구성됨을 확인
- 최적 군집 수 탐색

별점	1	2	3	4	5	합
빈도	189	776	1,218	5,631	13,809	21,623
퍼센트(%)	0.88	3.59	5.63	26.04	63.86	100

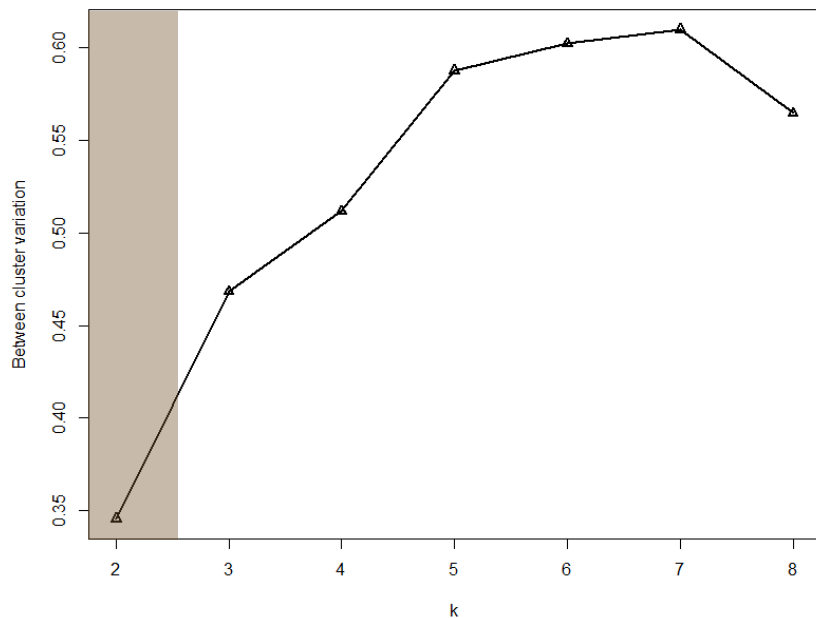
Clustering



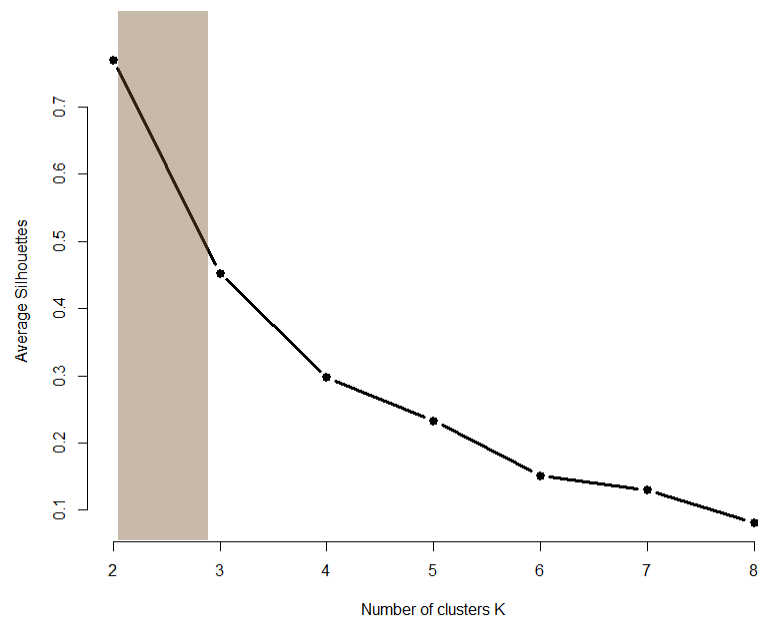
- K-means clustering Algorithm 중 최적의 Algorithm 탐색
- 탐색 결과, 설명력이 가장 높은 Forgy Algorithm 선택

Clustering

k=2부터 8까지 군집 간 변동의 차이 비교



k=2부터 8까지 군집의 평균 실루엣 비교

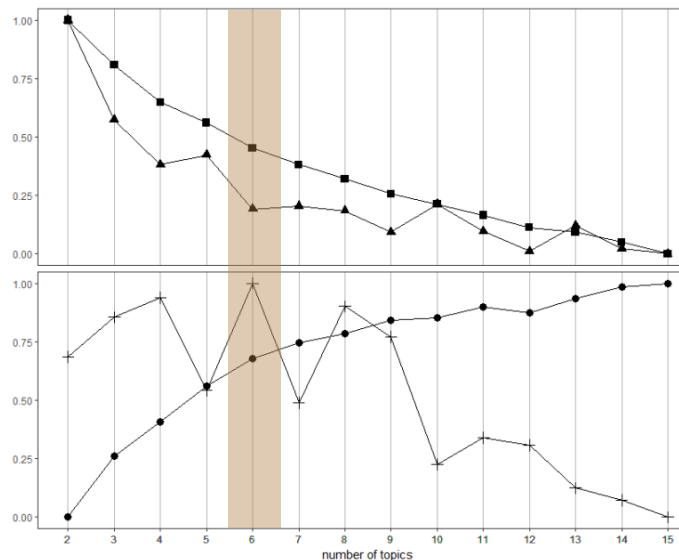


Clustering

- Google review 에 대한 최적의 K는 2로 결정, 따라서 별점도 3을 기준으로 나눔
- K-means clustering 결과, 351개의 Insight를 탐색할 수 있었음

구글		별점		
		Positive	Negative	Total
군집분석 결과	Positive	19,089	187	19,279
	Negative	351	1,996	2,347
	Total	19,440	2,183	21,623

Topic Modeling for Insight

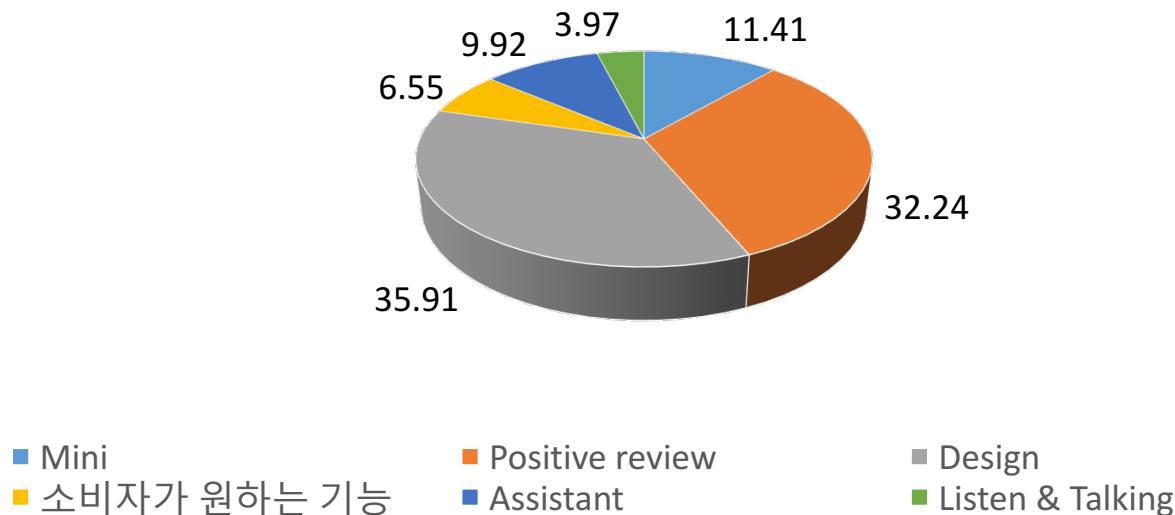


● Griffiths2004 ▲ CaoJuan2009 ■ Arun2010 + Deveaud2014

- Google review 중 351건의 Insight에 대한 Topic Modeling 실시
- 최적의 Topic 수는 6으로 결정

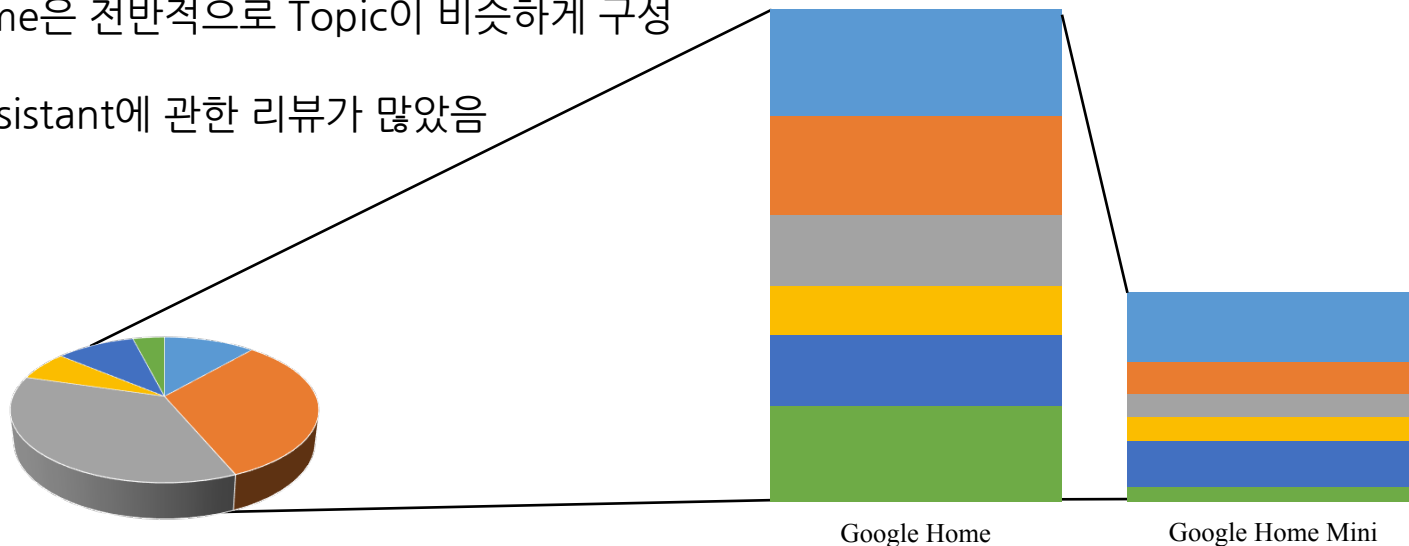
Topic Modeling for Insight

- Design에 관한 Topic이 가장 많은 비중을 차지하고 있음
- 6개의 Topic 중 Google Home Mini 제품에 대한 Topic이 따로 할당



Topic Modeling for Insight

- Google A.I. Speaker 제품 별 Topic의 비중
- Google Home은 전반적으로 Topic이 비슷하게 구성
- Mini에서 Assistant에 관한 리뷰가 많았음



■ Mini ■ Positive review ■ Design ■ 소비자가 원하는 기능 ■ Assistant ■ Listen & Talking

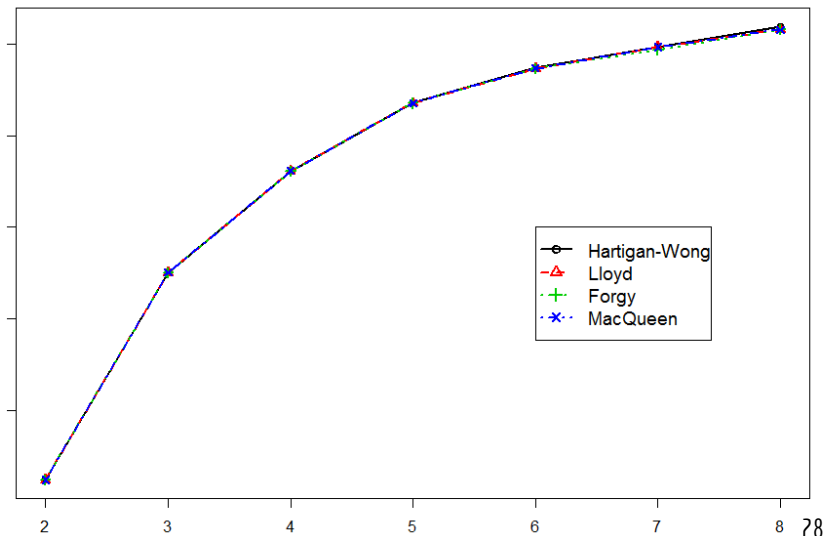
Clustering

- Amazon에 대한 리뷰는 21,129건이 존재하였으며 별점은 다소 균형적인 분포를 이룸
- 21,129건의 리뷰에 대하여 전처리한 결과, 총 5,739건의 단어로 구성됨을 확인
- 최적 군집 수 탐색

별점	1	2	3	4	5	합
빈도	4,677	406	746	3,554	11,749	21,129
퍼센트(%)	22.14	1.91	3.53	16.81	55.61	100

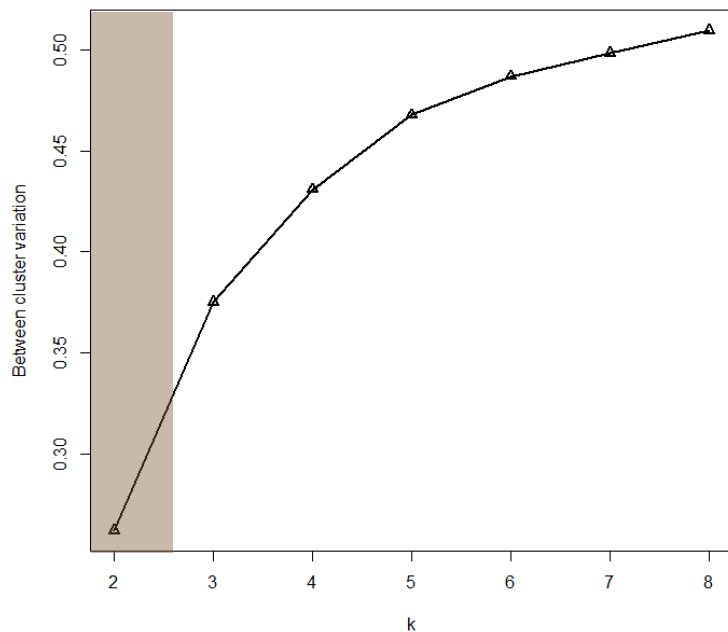
Clustering

- K-means clustering Algorithm 중 최적의 Algorithm 탐색
- 탐색 결과, Algorithm 별 설명력의 차이가 없어 Google과 같은 Forgy Algorithm 선택

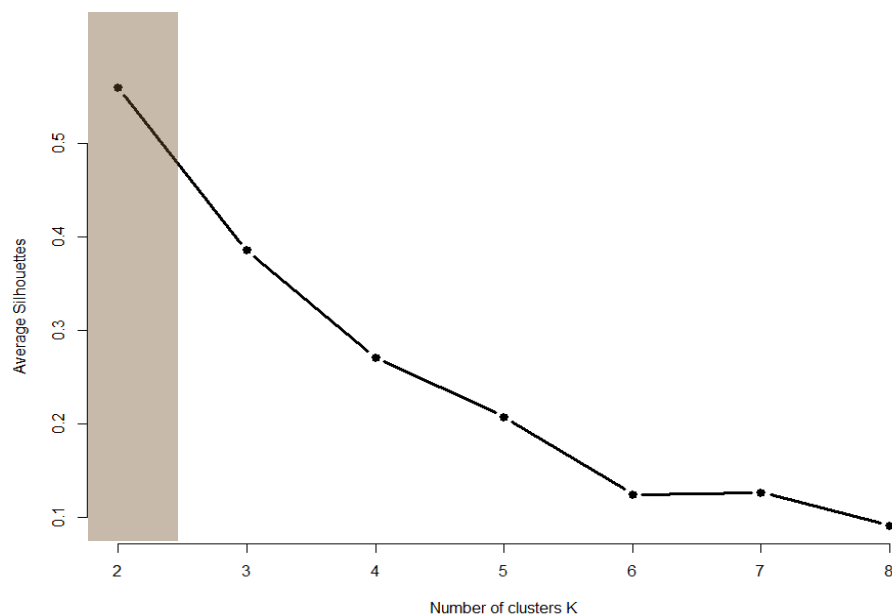


Clustering

k=2부터 8까지 군집 간 변동의 차이 비교



k=2부터 8까지 군집의 평균 실루엣 비교

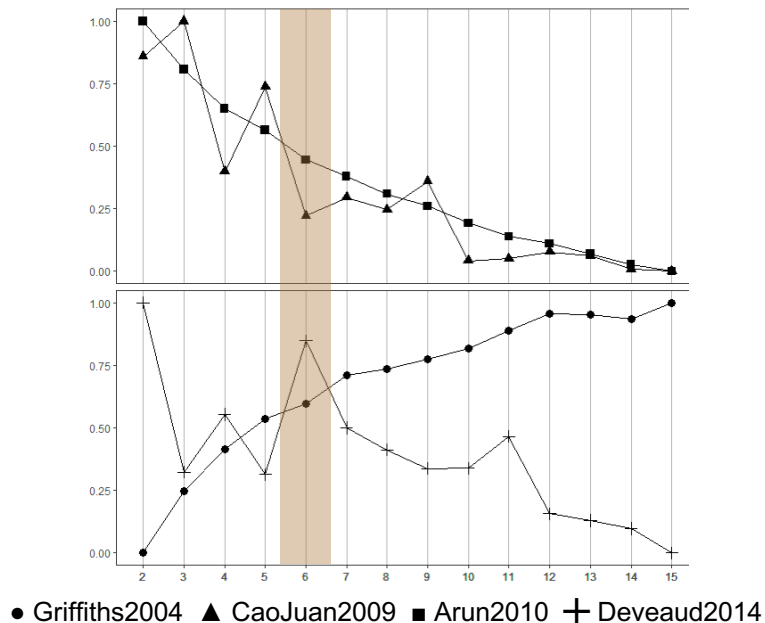


Clustering

- Amazon review 에 대한 최적의 K는 2로 결정, 따라서 별점도 3을 기준으로 나눔
- K-means clustering 결과, 1,347개의 Insight를 탐색할 수 있었음

기업별 리뷰데이터		별점		
		Positive	Negative	Total
군집분석 결과	Positive	13,956	526	14,482
	Negative	1,347	5,300	6,647
	Total	15,303	5,826	21,129

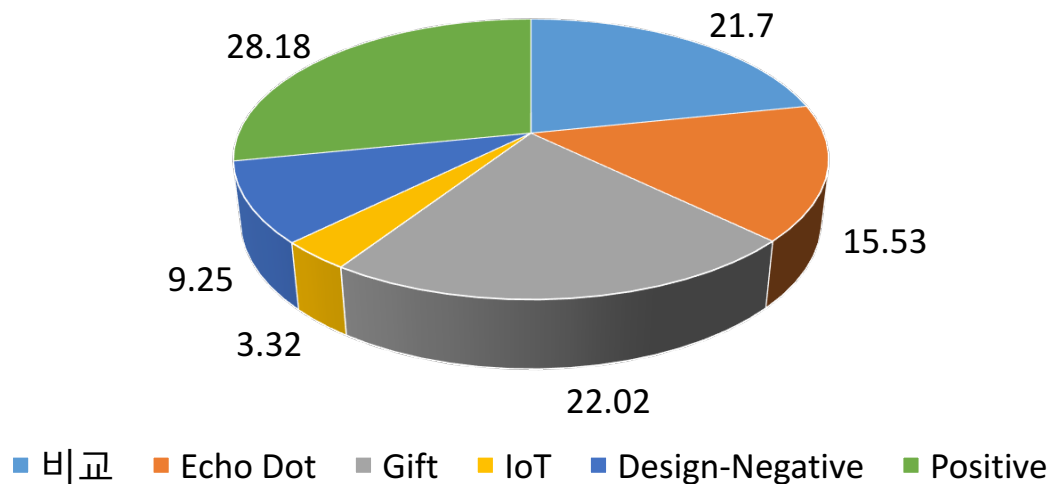
Topic Modeling for Insight



- Amazon review 중 1,347건의 Insight에 대한 Topic Modeling 실시
- 최적의 Topic 수는 6으로 결정

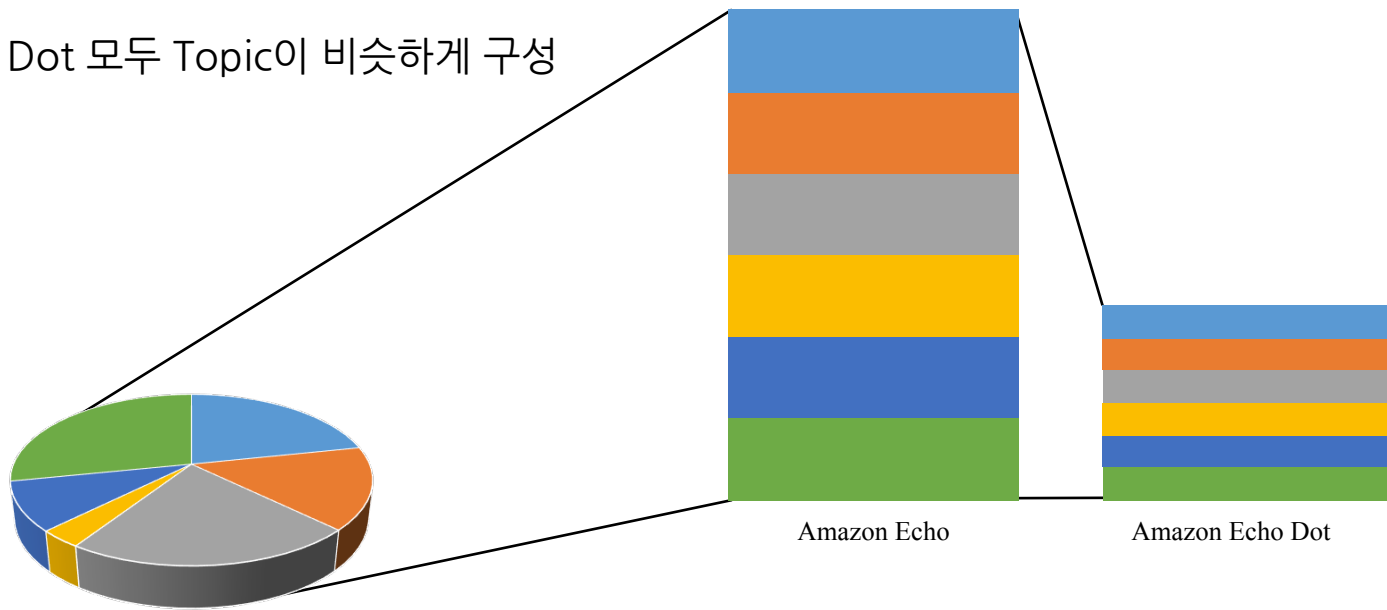
Topic Modeling for Insight

- 선물용으로 만족(Gift)한다는 Topic이 가장 많은 비중을 차지하고 있음
- 부정적 Design 토픽



Topic Modeling for Insight

- Amazon A.I. Speaker 제품 별 Topic의 비중
- Echo와 Echo Dot 모두 Topic이 비슷하게 구성



■ 비교 ■ Echo Dot ■ Gift ■ IoT ■ Design-Negative ■ Positive

Limitations & Improvements

- Amazon Echo와 Google Home에 관한 데이터 불균형
 - * 데이터 수를 맞추기 위해 다른 리뷰 사이트 크롤링
- 긍정적인 데이터와 부정적인 데이터의 수의 불균형
 - * 샘플링 등의 기법을 통해 불균형 완화
- 다의적 단어: Google, Design 등 추가 처리 필요
- 광고성 리뷰의 필터링 등을 통해 더욱 정확한 분석이 가능

Conclusion

- 초기 성장 단계이지만 아직은 긍정적인 반응이 있다고 판단
- 부정적인 리뷰의 다수는 소비자의 기대치에 맞지 않은 기능/편의성 (개선필요)
- 크리스마스 선물: “내가 사고 싶지는 않지만 선물 받으면 좋은” 제품일 수도 있음
- 단순 호기심에서 장기 고객으로 끌고 가는 것이 핵심
- 플랫폼/연동되는 서비스 및 스마트홈 등 일상생활에 녹아 드는 Lock-in 전략 실행
- 두 기업 비슷한 트렌드를 가지고 있음으로 아직 큰 차이점은 발견하지 못함

Thank You!