토픽모델링과 주성분 분석을 활용한 온라인 쇼핑 검색 질의 유형 분류

강현아*, 임희석**
고려대학교 컴퓨터정보통신대학원
khariss@korea.ac.kr, limhseok@korea.ac.kr

A Study on the Types of Online Shopping Queries using Topic Modeling and Principal Components Analysis

Hyeonah Kang*, Heuiseok Lim**
Dept. of Computer & Information Technology, Korea University

요 약

검색 질의 연구 분야의 대부분 선행 연구는 검색 질의 주제 분류에 집중되어 있으며 질의 자체에 대한 연구자의 정성적인 판단으로 분석되었다. 이는 검색 이후 클릭 된 문서를 고려하지 않고 진행되었다는 점과 분석 주제 및 활용 데이터가 제한적이라는 것에 한계가 있다. 이에 본 연구는 국내 대형 온라인쇼핑몰의 1년간의 검색로그를 활용하여 검색 질의와 검색 이후 조회한 문서명 정보를 기반으로 토픽모델링을 수행하여 검색 질의 주제를 정의하였다. 또한 검색 행동특성에 따른 주제별 성격을 정의하기 위하여 주성분 분석을 통해 주요 변수 추출 후 각 주제별 검색 행동특성을 분석하였다. 본 연구 결과는 효과적인 검색 서비스 구축 및 검색 시스템 개발에 기여 할 것으로 기대된다. 향후 연구로는 텍스트 분류기 모델링 실험을 통해 자동 분류 시스템을 구현할 수 있을 것이다.

1. 서론

국내 온라인커머스 시장은 매년 두 자릿수의 높은 성장률을 보이며 성장하고 있고 통계청 자료에 따르면 2019년도는 매출액 기준 유통시장의 점유율 28.2%를 차지하여 오프라인 대형마트의 아성을 무너뜨리고 있다. 온라인커머스 분야의 급성장에 따라 이와 관련된 다양한 연구들이 진행되었다. 권혁인 외 3명의 연구[1]에 따르면 e-커머스의 산업 생태계의 활성화 요인을 가중치 내림차순으로 '검색서비스 개발(0.0970)' > '추천서비스 개발(0.0805)' > '소비자 니즈 분석(0.0534)' > '고객 소비 패턴 분석(0.0505)' > '타 플랫폼 연계 서비스 개발(0.0450)'로 선정하였다. 해당 연구에서 언급된 요인 중 검색 시스템, 모델링 또는 추천서비스 연구 등은 이미 학계, 산업에서 연구가 활발한 분야인 반면에 '소비자 니즈 분석'의 관한 연구는 선행 연구가 많지 않은 실정이다. 특히 온라인커머스에서 소비자 니즈의 구체적인 발현은 검색 질의라고 할 수 있는데 국/내외 대부분의 연구는 적은 양의 데이터를 대상으로 연구자의 정성적인 판단에 근거하여 검색 질의 유형을 분류하는 제한적인 방법으로 연구되어왔다.

본 연구는 온라인커머스에서 고객의 니즈를 가장 집약적으로 나타내는 검색 질의 유형 분석을 위해 검색 질의뿐만 아니라 검색 이후 클릭로그를 활용하여 정량적인 측면과 정성적인 측면에서 다각적으로 분석하고자 한다. 이를 위해 19년 월평균 세션수 2.6억 규모의 국내 쇼핑사이트에서 1년간 발생한 빅데이터 검색 로그를 활용하여 검색 질의와 검색 후 조회 문서명의 비정형 데이터를 수집하고 이를 문서명과 문서 내용의 관계로 간주하여 텍스트에서 자동으로 주제를 추출해주는 토픽모델링을 수행한다. 이를 통해 검색 질의 자체뿐만 아니라 검색 이후 조회된 문서까지 고려하여 소비자의 의도를 명확하게 담아 질의 유형을 정의할 수 있으며 기계 모델을 사용하여 자동으로 분류하기 때문에 대량의 데이터를 활용할 수 있다는 데 연구의 의의가 있다. 또한 주제별 검색 행동특성을 분석하기 위해 검색 행동특성 변수 대상 주성분 분석을 수행하여 제 1, 2 주성분을 기준으로 총 4개의 행동특성별 유형을 정의하였다. 제2장에서는 관련 연구 및 이론에 대해서 다루고 제3장에서는 연구 프레임 및 방법론을 기술한다. 제4장에서는 연구 결과에 대해 분석하고 제5장에서는 결

론 및 향후 연구 제에 대해 기술한다.

2. 관련 연구

웹 검색 분야에서 빅데이터 수준의 트랜잭션(transaction) 로그를 활용하여 검색어를 분석한 연구로 Silverstein et al.(1999)이 1998년 8월 2일부터 6주간의 알타비스타 이용자들이 남긴 약 3억개 수준의 이용자 세션과 약 10억 개의 질의를 분석하였다.[2] 해당 연구는 지금까지 트랜잭션 로그를 활용한 연구 중 가장 방대한 데이터를 기반으로 진행된 연구였고 세션 정의 등과 같은 로그 분석 방법론을 제시하였다는데 의의가 있다. Spink et al.(2001)은 1997년 9월 16일 익사이트 이용자들이 남긴 약 100만개의 질의 대상으로 2,414개를 무작위로 추출 후 이를 11개의 범주로 분류하는 체계를 도출하였다.[3] 박소연, 이준호, 김지승(2005)은 2003년 7월부터 2004년 6월까지 1년간 네이버에서 발생한 18,200개의 질의 로그와 검색 이후 검색결과에서 이용자가 조회한 문서 등의 로그를 바탕으로 질의 형태 및 주제를 분류하였다.[4] 해당 연구는 국내 최초로 대량의 검색로그를 기반으로 검색 질의 분류를 수행하였다는데 의의가 있으나 여전히 질의 분류 시 연구자의 정성적인 판단에 의존하는 방법으로 분류한다는 점에서 한계가 있다.

3. 연구 방법

3.1 연구 개요

연구 프레임워크는 (그림 1)과 같다.

(그림 1) 연구 프레임워크

데이터는 19년 월평균 세션수 2.6억 규모의 국내 온라인쇼핑몰의 1년간의 검색로그를 활용하였다. 데이터 수집 단계에서 검색엔진DB에서 검색 질의와 검색결과의 클릭로그를 추출하여 분석마트를 구축한다. 데이터 수집 환경은 HDFS이며 HiveQL을 사용하였다. 검색 질의는 대표성 및 인기도에 초점을 맞추어 기간 내 누적 검색횟수 기준 상위 200,000개의 검색 질의를 수집하였다. 이는 연간 발생 검색 질의 중 0.24% 수준이며 검색횟수 비중으로는 61.6%를 차지하는 short-head query로 구성되어있다. 검색 질의를 기준으로 검색횟수(QC), 검색결과 문서 조회수(CC), 검색결과 대비 문서 클릭수(CTR) 등 검색 행동특성을 나타내는 변수 70개를 선정하고 수치를 집계하였다. 마지막으로 검색 이후 조회 문서명을 높은 조회 순으로 최대 50개까지 수집하였다.

데이터 전처리 과정에서는 텍스트인 검색결과 조회 문서명을 토픽모델링 입력값으로 활용하기 위해 일련의 자연어처리 과정을 거치는데 데이터 클렌징, 토큰화, 품사태깅, 불용어제거, TF-IDF matrix 생성을 통해 단어 임베딩(Word Embedding)을 수행하여 텍스트를 벡터로 변환한다. 검색 질의 유형 분류 단계는 토픽모델링 기법 중 LDA(Latent Dirichlet allocation)를 수행하여 검색 질의의 주제 유형을 정의하고 이후 주제별 검색 행동특성 기반 주성분 분석을 수행하여 검색 행동특성별 유형을 분석한다.

3.2 분석 방법

자연어처리의 한 분야인 LDA(Latent Dirichlet allocation)는 비지도 학습을 수행하는 확률 그래프 모델로서 주어진 문서에 대하여 Dirichlet 분포를 이용하여 각 문서에 어떤 주제들이 존재하는지를 서술하는 대한 확률적 모델링 기법이다.[5]

PCA(Principal Component Analysis)는 차원축소(Dimensionality Reduction Method) 기법으로 직교변환을 이용하여 고차원 공간의 표본들을 선형 연관성이 없는 저차원 공간(주성분)의 표본으로 변환한다. 한 개의 축으로 사상시켰을 때 그 분산이 가장 커지는 축을 첫 번째 주성분(PC1)으로 설정하고 PC1과 직교하는 모든 방향 중 분산을 최대화하는 방향을 두 번째 주성분(PC2)으로 정의한다.[6]

4. 연구 결과

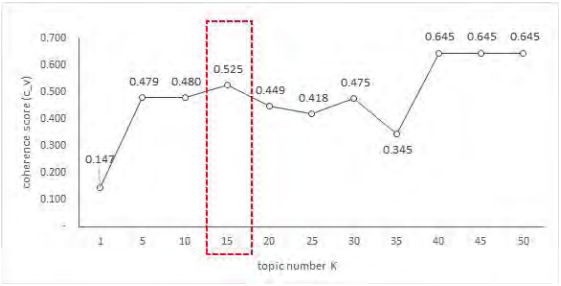
4.1 LDA 기반 검색 질의 주제 유형 정의

LDA 실험은 프로그래밍 언어로 Python을 활용하였고 통합 개발 환경은 Jupyter Notebook(Anaconda 3), 라이브러리는 gensim을 이용하였다. 다음 <표 1>은 검색 질의와 질의별 검색결과 이후 조회한 문서명의 Term-Document 행렬 일부를 나타내고 있다. 해당 데이터셋이 토픽모델링의 입력 데이터로 활용되어 검색 질의의 토픽을 도출하게 된다.

<표 1> 토픽모델링 실험 데이터셋

검색 질의	검색 결과 이후 조회 문서명
검색어1	나이키에어맥스 화이트 실버 [illegible] 4종 나이키 스포츠웨어 테크 플리스 … (이하생략)
검색어2	애플 에어팟 Apple Airpod MMEF2KH,블랙색상 제이스 엠씨스 JAYS 블루투스5.0 무선이어폰 … (이하생략)
검색어3	베베숲 센시티브 80매 캡 10팩 미마미아 프로즌2 [illegible] 물티슈 캡형 70매 도리도리 …(이하생략)
검색어4	ASUS 정품과우지 [illegible] ASUS 뉴젠북 UX433FN-A6053T 인텔 i5-8265U 치포스 대학생추천 … (이하생략)
검색어5	커네스트 미포07플러스 블루투스이어폰 블랙에디션 할인이벤트사은품 듀얼DAC탑재 … (이하생략)
검색어6	V2 새브 [illegible]400g 무용매파스타 캘리포니아 Pay's 아웃백디지털상품권10만원권 … (이하생략)

토픽모델링의 주요 파라미터는 토픽 개수로 K를 선정하는 것인데 이를 위해 토픽 수 최적화 실험을 수행하였다. 토픽 개수를 늘려가며 coherence score (응집성 지수, c_v)를 측정하였으며 해당 값이 높을수록 토픽이 의미론적으로 일관성이 높다고 할 수 있다.[7] 실험결과 (그림2)와 같이 K=15개에서 비교적 높은 coherence score 0.525가 측정되었고 40개 이상의 토픽수는 유의미한 주제 그룹화에 용이하지 않다고 판단하여 토픽 수를 15개로 결정하였다.



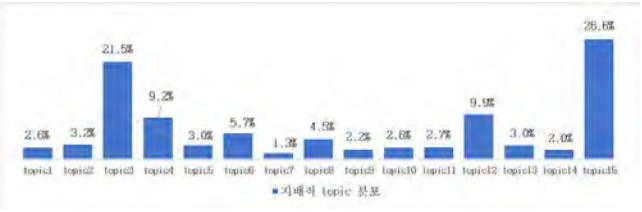
(그림 2) 토픽 수에 따른 coherence score

다음 <표 2>는 20만 개의 검색결과 조회문서 대상 토픽모델링으로 도출된 15개의 토픽별 가장 영향도가 높으며 유의한 해석이 가능한 키워드 top5이다.

<표 2> LDA 결과 topic1~15 주요 키워드 및 주제 정의

TOPIC	topic_1	topic_2	topic_3	topic_4	topic_5
주제 정의	주방관련 식품/가전	컴퓨터/교육/유아동	패션의류	가공/건강식품	생활가전/생필품
키워드1	0.001*"생수"	0.001*"노트북"	0.004*"나이키"	0.002*"포"	0.001*"이어폰"
키워드2	0.000*"두유"	0.001*"학년"	0.003*"빅사이즈"	0.002*"개"	0.001*"다이슨"
키워드3	0.000*"코베아"	0.001*"그램"	0.003*"티셔츠"	0.001*"정"	0.001*"무선청소기"
키워드4	0.000*"전기레인지"	0.001*"수학"	0.003*"자켓"	0.001*"캡슐"	0.001*"겸"
키워드5	0.000*"구"	0.001*"ASUS"	0.003*"원피스"	0.001*"오두기"	0.001*"생리대"
TOPIC	topic_6	topic_7	topic_8	topic_9	topic_10
주제 정의	라이프뷰티	취미	라이프플러스	스마트디지털	컴퓨터주변/e쿠폰
키워드1	0.002*"크림"	0.001*"레고"	0.001*"귀걸이"	0.005*"노트"	0.001*"GB"
키워드2	0.001*"로션"	0.001*"어벤져스"	0.001*"공기청정기"	0.004*"A"	0.001*"지포스"
키워드3	0.001*"헤라"	0.000*"아이언맨"	0.001*"목걸이"	0.003*"케이스"	0.001*"GTX"
키워드4	0.001*"에센스"	0.000*"마블"	0.001*"자동차"	0.003*"갤럭시"	0.001*"갤럭시탭A"
키워드5	0.001*"스킨"	0.000*"피규어"	0.001*"반지"	0.002*"아이폰"	0.001*"갤럭시탭S"
TOPIC	topic_11	topic_12	topic_13	topic_14	topic_15
주제 정의	생필품/식품	트랜드잡화	레포츠/아웃도어	여행/성인	가구/인테리어
키워드1	0.001*"백신"	0.002*"백팩"	0.001*"시마노"	0.001*"강원"	0.002*"단"
키워드2	0.001*"하기스"	0.002*"크로스백"	0.001*"행거"	0.000*"단제"	0.001*"원목"
키워드3	0.001*"기저귀"	0.002*"가방"	0.001*"다이와"	0.000*"리조트"	0.001*"인테리어"
키워드4	0.001*"카누"	0.001*"캐리어"	0.000*"등산화"	0.000*"성인용품"	0.001*"케이스"
키워드5	0.001*"제주"	0.001*"숄더백"	0.000*"장갑"	0.000*"테일러메이드"	0.001*"다용도"

topic별 주요 구성 키워드를 기반으로 해당 키워드의 카테고리 또는 주제 측면에서 상위 개념으로 카테고리화 하여 주제를 정의하였다. 토픽 주제 15개는 topic1부터 '주방관련 식품/가전', '컴퓨터/교육/유아동', '패션의류', '가공/건강식품', '생활가전/생필품', '라이프뷰티', '취미', '라이프플러스', '스마트디지털', '컴퓨터주변/e쿠폰', '생필품/식품', '트랜드잡화', '레포츠/아웃도어', '여행/성인', '가구/인테리어'이다. 해당 토픽 주제를 최종적으로 검색 질의의 주제로 정의하는 과정은 검색 질의별로 해당 질의를 구성하는 topic 중 가장 비중이 높은 topic을 지배적 topic으로 선정하였고 해당 topic의 주제를 검색 질의의 최종 주제 유형으로 정의하였다. 해당 기준으로 검색 질의별 지배적 topic 분포는 다음 (그림 3)과 같으며 topic15 26.6%, topic3 21.5%로 비중이 높았으며 그 다음 비중이 5%를 넘는 토픽으로 topic4, 6, 12, 그 외 토픽들은 5% 미만으로 분포하였다.



(그림 3) 검색 질의별 지배적 Topic 분포

4.2 PCA 기반 검색 질의 주제별 검색특성 정의

15개의 검색 질의 주제 유형에 대해 검색 행동특성을 분석하여 행동특성별로 유형을 구분하고자 한다. 이를 위해 검색 행동특성 관련 변수 12개에 대하여 주제 유형별 실적을 집계 후 주성분 분석(PCA)을 통해 제 1, 2 주성분을 도출하고 해당 기준으로 총 4개의 검색 행동특성별 유형을 정의하였다. PCA 실험을 위한 변수는 검색데이터 분석경력이 있는 연구자의 정성 평가로 <표 3>과 같이 선정하였다.

<표 3> PCA 활용 검색 행동특성 관련 변수

구분	물리명	논리명
검색 활성도	qc_per_sess	세션당 검색회수
검색 결과 만족도	ctr	클릭율
검색 상품 만족도	cc_per_prd	상품당 클릭수
	buy_try_ratio	검색 후 구매시도율
	buynow_ratio	구매시도 중 바로구매 비중
	qc_cvr	검색 후 구매 전환율
문서(상품) 클릭위치	clk_pos	평균 클릭 위치
	fst_clk_pos	첫클릭 위치
문서(상품) 클릭 분산/집중도	m_clk_ratio	최다클릭 문서 클릭점유율
	avg_prd_clk_ratio	문서별 평균 클릭점유율
상품가격	amt_per_ord	주문당 평균 결제금액 *) 검색이후 결제로 이어진 주문
광고상품 집중도	ad_qc_ratio	검색광고 비중

PCA 수행은 통계분석 소프트웨어인 R의 prcomp, FactoMineR 패키지를 사용하였으며 변수 간 scale이 선형결합 시에 미치는 영향도를 제거하기 위하여 상관계수행렬(correlation matrix)를 사용하였다.

PCA 결과 <표 4>와 같이 주성분1, 2의 각 고유값(Standard deviation)이 각 1이 넘으며 주성분2까지의 누적 설명력(Cumulative Proportion)이 79.7%로 주성분 채택 기준인 80%에 근사하여 PC1, PC2를 검색 행동특성을 설명하는 주성분으로 채택하였다.

<표 4> Importance of PCA components

구분	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	2.736	1.442	0.947	0.782	0.708	0.457	0.382	0.217	0.140	0.111	0.081	0.030
Proportion of Variance	0.624	0.173	0.075	0.051	0.042	0.017	0.011	0.004	0.002	0.001	0.001	0.000
Cumulative Proportion	0.624	0.797	0.872	0.923	0.964	0.982	0.993	0.997	0.998	0.999	1.000	1.000

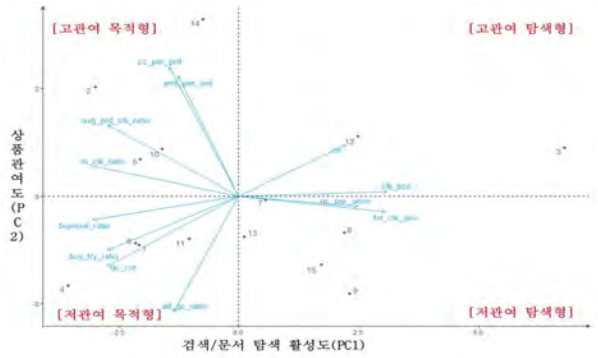
PC1, PC2의 특성을 파악하기 위해 PC1, PC2에 영향을 미치는 변수가 무엇인지 살펴보았다. PC1은 양의 상관관계로 평균 클릭 위치(0.354), 첫 클릭 위치

(0.351), 음의 상관관계로 구매시도 중 바로구매 비중(- 0.347), 최다클릭 문서 클릭점유율(-0.346)의 영향도가 커 검색 및 문서 탐색의 활성도를 나타내는 성분으로 특정 지었다. PC2는 상품당 클릭수(0.520), 주문당 평균 결제금액(0.483)의 영향도가 커 상품관여도를 나타내는 주성분으로 특정 지었다.

<표 5> PCA 제1, 2 주성분 특성 정의

주성분 구분	특성 정의	설명내용
PC1	검색/문서 탐색 활성도 (고활동/저활동)	클릭위치와 상위 관계로 연관이 깊고 문서집중도, 바로구매비중과 음의 관계로 연관이 깊음 즉 검색결과 하단에서 클릭할수록(탐색을 많이 할 수록) 문서집중도와 바로구매 비중은 떨어짐
PC2	상품 관여도 (적관여/고관여)	상품당 문서클릭수(검색결과클릭수)와 양의 관계, 주문건수당 거래액과 양의 관계로 연관이 깊음 즉 고객 결과에서에서 중심한 문서(상품)을 여러 번 클릭하는 경우 결제단가도 높은 경향

2차원의 x축은 PC1(검색/문서 탐색 활성도)으로, y축은 PC2(상품관여도)로 설정하여 각 축이 0이 되는 지점을 기준으로 사분면으로 영역을 구분 후 15개의 topic을 투사하였다.



검색 행동특성 유형	토픽
고관여 탐색형	[3]패션의류, [12]트랜드잡화
고관여 목적형	[2]컴퓨터/교육/유아동, [5]생활가전/생필품, [10]컴퓨터주변/e쿠폰, [14]여행/성인
저관여 탐색형	[7]취미, [8]라이프플러스, [9]스마트디지털, [13]레포츠/아웃도어, [15]가구/인테리어
저관여 목적형	[1]주방관련 식품/가전, [4]가공/건강식품, [6]라이프뷰티, [11]생필품/식품

(그림 4) 검색 질의 주제 유형별 검색 행동특성 유형

'고관여 탐색형'에 해당하는 토픽은 (topic3)패션의류, (topic12)트랜드잡화로 나타났다. 패션 관련된 검색 질의는 사용자가 찾고자 하는 상품에 대한 속성을 키워드에 적절히 반영하여 정보검색(information retrieval)이 이루어지는 것이 어렵기 때문에 보통 여러 번의 검색과 문서 클릭 등의 탐색 패턴을 보이는 것으로 사료된다. '고관여 목적형'에 해당하는 토픽은 (topic2)컴퓨터/교육/유아동, (topic5)생활가전/생필품, (topic10)컴퓨터/e쿠폰, (topic14)여행/성인으로 나타났다. '고관여 목적형' 유형은 '고관여 탐색형' 대비 검색 탐색 활성도 관련하여 더 적은 검색과 비교적 특정 문서(상품)에 집중된 문서클릭, 검색결과 상단의 상품을 클릭하는 등의 소극적 탐색활동 특징을 보인다. '저관여 탐색형'에 해당하는 토픽은 (topic7)취미, (topic8)라이프플러스, (topic9)스마트디지털, (topic13)레포츠/아웃도어, (topic15)가구/인테리어로 나타났다. 해당 유형에 속하는 토픽은 검색/문서 탐색 활성도는 높으나 상품관여도가 낮은 특성을 보인다. '저관여 목적형'에 해당하는 토픽은 (topic1)주방관련 식품/가전, (topic4)가공/건강식품, (topic6)라이프뷰티, (topic11)생필품/식품이다.

5. 결론 및 향후 연구 과제

본 연구는 선행 연구에서 정성적 방법과 제한적인 데이터로 진행되었던 검색 질의 분류 연구를 정량적인 기계학습 방법을 적용하고 빅데이터를 활용하여 의미 있는 연구 결과를 확인하였다는 점에서 의의가 있다. 또한 주제뿐만 아니라 검색 행동특성까지 고려한 검색 질의 유형을 정의하였다는 점에서 검색 질의 연구 분야의 다양성을 제고하였다. 본 연구 결과는 효과적인 검색 서비스 구축 및 검색 시스템 개발에 기여할 것으로 기대된다. 향후 연구 과제로는 검색 주제 유형을 자동으로 분류하는 기계학습 모델링을 통하여 여러 분류기를 학습시키고 성능을 평가하여 최적의 분류 시스템을 제안하고자 한다.

참고문헌

[1] 권혁인, 백보현, 안예진, 이진형, "e-커머스 서비스 혁신을 위한 발전전략 연구", 한국콘텐츠학회논문지: 20(1), 217-232.

[2] C Silverstein, H Marais, M Henzinger, M Moricz, "Analysis of a very large web search engine query log.", SIGIR Forum,1999, 33(1): 6-12.

[3] A Spink, D Wolfram, MBJ Jansen, T Saracevic, "Searching the web: The public and their queries", Journal of the American Society for Information Science and Technology, 52(3): 226-234, 2001.

[4] 박소연, 이준호, 김지승, "클릭 로그에 근거한 네이버 검색 질의의 형태 및 주제 분석", 한국문헌정보학회지 39(1), 2005, 265-278.

[5] S. Y. Bong and K. B. Hwang, "Applying Labeled LDA to Author Keywirds Recommendation", Proceedings of KIISE Spring Conference, Vol.37, No.1(C), pp.385-389, 2010.

[6] Hotelling H., "Analysis of a complex of statistical variables into principal components", Journal of Educational Psychology, vol. 24, no. 6, p.417, 1933

[7] Newman, D., Lau, J. H., Grieser, K., & Baldwin, T., "Automatic evaluation of topic coherence", NAACL HLT 2010 - Human Language Technologies, 2010, 100-108.