Journal of the Korean Society of Civil Engineers
Vol. 38, No. 4: 595-599/ August, 2018
DOI: https://doi.org/10.12652/Ksce.2018.38.4.0595

ISSN 1015-6348 (Print)
ISSN 2287-934X (Online)
www.kscejournal.or.kr

Construction Management **시공관리**

Latent Dirichlet Allocation 기법을 활용한 해외건설시장 뉴스기사의 토픽 모델링(Topic Modeling)

문성현* · 정세환** · 지석호***
Moon, Seonghyeon*, Chung, Sehwan**, Chi, Seokho***

Topic Modeling of News Article about International Construction Market Using Latent Dirichlet Allocation

ABSTRACT

Sufficient understanding of oversea construction market status is crucial to get profitability in the international construction project. Plenty of researchers have been considering the news article as a fine data source for figuring out the market condition, since the data includes market information such as political, economic, and social issue. Since the text data exists in unstructured format with huge size, various text-mining techniques were studied to reduce the unnecessary manpower, time, and cost to summarize the data. However, there are some limitations to extract the needed information from the news article because of the existence of various topics in the data. This research is aimed to overcome the problems and contribute to summarization of market status by performing topic modeling with Latent Dirichlet Allocation. With assuming that 10 topics existed in the corpus, the topics included projects for user convenience (topic-2), private supports to solve poverty problems in Africa (topic-4), and so on. By grouping the topics in the news articles, the results could improve extracting useful information and summarizing the market status.

Key words : International construction market, News article, Text-mining, Topic modeling, Latent dirichlet allocation

초 록

해외건설 프로젝트를 기획하고 수행하는 과정에서 현지 시장의 상황을 신속하고 정확하게 파악하는 것은 수익성 창출에 매우 큰 영향을 미친다. 뉴스기사 데이터는 정치, 경제, 사회 등 다양한 관한 정보를 담고 있기 때문에 시장의 상황을 파악하는 데 사용할 수 있는 좋은 데이터이다. 텍스트의 형태로 존재하는 대량의 뉴스기사 데이터로부터 정보를 추출하고 내용을 요약하는 과정에서 인력, 비용, 시간의 소모를 줄이기 위해 텍스트마이닝 기술이 필요하다. 본 연구에서는 뉴스기사에 다양한 주제가 공존한다는 특성으로 인해 발생하는 정보 추출의 한계를 극복하기 위해 잠재 디리클레 할당(Latent Dirichlet Allocation) 방법론을 사용하여 토픽 모델링을 수행했다. 문서 집단에 존재하는 주제의 개수가 10개라고 가정했을 때, 이용자들의 편의 증진을 위한 프로젝트(2번 주제)와 아프리카 지역의 빈곤 문제를 해결하기 위한 민간 차원의 지원(4번 주제) 등의 주제 집단이 존재하는 것을 확인했다. 이와 같이 문서 집단의 주제를 구분함으로써 더욱 의미있는 정보를 추출하고, 요약 결과의 활용성을 높일 수 있다.

검색어 : 해외건설시장, 뉴스기사, 텍스트 마이닝, 토픽 모델링, 잠재 디리클레 할당

* 정회원 · 서울대학교 건설환경공학부 석박통합과정 (Seoul National University · blank54@snu.ac.kr)
** 정회원 · 서울대학교 건설환경공학부 석사과정 (Seoul National University · hwani751@snu.ac.kr)
*** 종신회원 · 교신저자 · 서울대학교 건설환경공학부 교수, 서울대학교 건설환경종합연구소 겸임교수
(**Corresponding Author** · Seoul National University, The Institute of Construction and Environmental Engineering (ICEE) · shchi@snu.ac.kr)

Received April 7, 2018/ **revised** May 8, 2018/ **accepted** May 14, 2018

1. 서 론

2010년 해외건설 수주액이 약 716억 불에 달하여 최고치를 경신한 이래 해외건설 수주액과 수주 건수는 전반적으로 감소하는 추세이다(ICAK, 2018). 하지만 2017년 총 수주액이 290억 불에 달하는 등 여전히 건설산업의 큰 비중을 차지하고 있고, 많은 건설기업들이 해외 시장에 진출하기 위해 노력하고 있다(Lee et al., 2015). 해외건설사업은 국내건설사업에 비해 더 많은 위험 요소를 가지는 것으로 알려져 있는데, 다양한 주체의 참여로 인한 문화 및 제도의 차이와 현지 상황에 대한 이해도 부족 등이 그 원인으로 지적되고 있다(Kim et al., 2009; Taroun, 2014).

여러 연구들의 결과에 따르면 프로젝트를 수행하는 현지 시장의 상황을 신속하고 정확하게 파악하는 것이 현장의 수요에 대응하고 리스크를 저감하는데 매우 중요한 요소이다(Javernick-Will and Scott, 2010). 해외건설시장의 상황에 관한 정보는 제도적 요소, 기술적 요소, 사회적 요소, 경제적 요소의 4가지로 구분될 수 있으며, 이러한 상황 정보는 주로 수치 데이터(GDP, 물가, 환율 등)나 문서 데이터(뉴스기사, 보고서, SNS)로부터 파악할 수 있다(Javernick-Will and Scott, 2010). 수치 데이터의 경우 간단한 통계 분석만으로도 시장의 변화를 살펴볼 수 있지만, 시장의 상황이 해당 수치로 변환되기까지 일정한 시간을 필요로 하기 때문에 좋은 데이터가 아니다. 문서 데이터 중 전문가들이 작성한 보고서 또한 작성되기까지 오랜 시간을 필요로 하며, SNS의 경우 데이터의 신뢰성에 관한 질문을 피해갈 수 없다. 뉴스기사는 생성 당시의 정치, 경제, 사회 등의 상황을 포괄적으로 담고 있기 때문에 역사학자들이 과거의 상황을 분석하거나 사회학자들이 현재의 현상을 설명할 때 사용되는 중요한 데이터이다(Yang et al., 2011). 하지만 정보를 추출하기 위해 읽어야 하는 텍스트 데이터가 너무 많기 때문에 인력, 비용, 시간의 소모가 상당하며(Ferreira et al., 2014), 해외건설시장에 관한 뉴스기사에서 다루고 있는 내용 중 중요한 내용을 요약해서 제공할 수 있다면 해당 시장의 상황을 파악하는 데 유용하게 사용될 수 있다(Goldszmidt et al., 2011).

텍스트 데이터를 요약한다는 것은 데이터의 크기를 줄이면서 전체 문서의 내용을 잘 반영하고 있는 새로운 문서를 생성한다는 것을 의미한다(Ferreira et al., 2014; Pal and Saha, 2014). 요약 방식은 요약된 결과를 제공하는 방식에 따라, 원래 문서에 등장하는 문장으로 재구성하는 추출식(Extractive)과 완전히 새로운 문장을 생성해내는 추상식(Abstractive)의 두 가지로 구분된다(Pal and Saha, 2014). 두 가지 방식 모두 원래 문서에서 중요하다고 여겨지는 키워드를 보존하는 것이 요약의 핵심적인 논리인데(Gambhir and Gupta, 2017), 뉴스기사는 다양한 주제를 동시에 다루기 때문에 주제 별로 구분하지 않고 텍스트 분석을 수행하면 특정한 주제에 관해 설명하고 있는 뉴스기사의 정보가 퇴색된다. 그 예시로, 본 연구를 수행하기 위해 수집한 11,491 건의 월드뱅크(Worldbank) 뉴스기사에 전처리를 가한 후 단어들의 출현 빈도에 따라 중요한 키워드를 추출한 결과, development, countries, project, government, economic 등 시장의 상황에 관한 내용보다는 다소 포괄적인 의미의 단어들이 상위권에 위치해 있음을 확인했다. 이러한 정보로는 시장의 상황을 충분히 파악할 수 없으며, 뉴스기사를 주제에 따라 구분한 뒤 각 주제별로 어떤 내용이 다루어지고 있는지를 파악해야 한다.

일반적으로 각 뉴스기관에서는 자체적으로 구축한 카테고리에 따라 뉴스를 구분하여 제공하지만, 이러한 탑다운(Top-down) 방식은 두 가지 이유로 인해 현실에서의 활용 가능성이 낮다. 첫 째로, 처음부터 잘 설계된 카테고리가 아니라 새로운 뉴스기사가 작성되면 필요에 따라 새로운 주제가 형성되는 방식이기 때문에, 각 주제의 범위가 너무 포괄적이거나 너무 협소해서 분석이 여의치 않다. 월드뱅크의 경우 431개의 주제에 따라 기사를 분류하고 있는데, ‘Agriculture’와 같이 900건 이상의 기사가 포함되는 주제와 ‘Tigers’와 같이 10건 미만의 기사가 포함되는 주제가 공존한다. ‘International Law’의 경우 국가 간의 채무관계에 관한 2건의 기사만을 포함하고 있으며, 이는 관련 기사들이 주로 ‘International Economics’ 또는 ‘Business-regulation’ 등의 카테고리에 포함되어 있기 때문이다. 이런 상황에서 카테고리별로 텍스트 분석을 수행하면 각 카테고리에 포함된 텍스트 데이터의 양에 따라 결과가 흐려지게 된다. 둘 째로, 현지 시장에 관해 편향되지 않은 정보를 획득하기 위해서는 다양한 기관의 뉴스기사를 총체적으로 수집하여 분석해야 하는데, 뉴스기사의 제공 기관에 따라 카테고리가 완전히 다르기 때문에 연구 결과의 확장성이 떨어진다.

해외건설시장의 상황 정보를 신속하고 정확하게 파악하기 위해 뉴스기사를 분석하여 중요한 정보를 요약 제공하는 과정이 필요하다. 뉴스기사 데이터에는 여러 주제가 섞여있기 때문에 이를 구분하는 선행 작업이 요구되며, 뉴스기관에서 자체적으로 제공하는 분류 체계만을 사용할 경우 시장상황에 관한 정보 추출 및 연구 결과물의 확장에 한계가 있다. 따라서 본 연구는 사전에 구축된 주제 카테고리가 아닌, 뉴스기사의 내용에 기반해서 주제를 분류하는 것을 목표로 하며, 추후 해외건설시장의 상황 정보를 추출하는 연구의 기반을 마련하고자 한다.

2. 연구 방법

2.1 웹 크롤링(Web Crawling)

웹 크롤링(Web Crawling)은 웹사이트로부터 특정한 형태의 데이터를 자동으로 다운로드하는 기법이다(Manning et al., 2008). 웹사이트는 기본적으로 하이퍼텍스트 마크업 언어(Hypertext

Markup Language, HTML)로 구축되어 있는데, 이 프로그래밍 언어는 웹사이트의 글자 크기, 글꼴, 색깔, 그래픽, 하이퍼링크 등 다양한 기능을 태그(tag)의 형태로 일일이 정의하고 있다. 그렇기 때문에 어떤 데이터를 수집할 것인지가 결정된다면 해당 데이터의 태그를 파악하여 같은 태그를 가진 항목을 일괄적으로 수집하는 것이 가능하다.

뉴스기사를 수집하는 웹 크롤링은 크게 두 가지의 절차로 수행된다. 먼저, 뉴스기사의 목록을 제공하는 사이트에 접근하여, 하이퍼링크 태그의 형태로 존재하는 각 뉴스기사의 고유 주소(Uniform Resource Locator, URL)를 추출한다. 이후에는 각 뉴스기사에 접근하여 텍스트 태그의 형태로 존재하는 기사의 제목, 카테고리, 본문 등을 수집한다(Manning et al., 2008).

2.2 텍스트 전처리

텍스트 데이터를 분석하기 위해서는 토큰화(Tokenization)와 불용어 처리(Stopword Removal) 등의 전처리를 수행해야 한다(Manning et al., 2008). 인간이 사용하는 '자연어'를 컴퓨터가 이해할 수 있는 언어로 변환해준다고 하여, 이러한 과정을 자연어 처리(Natural Language Processing; NLP)라고 부른다.

토큰화는 텍스트 데이터를 분석 가능한 작은 단위로 분할하는 작업이다. 문장부호 등 텍스트 데이터의 의미를 분석하는 데 영향을 미치지 않는 요소들이 제거되며, 모든 문장은 띄어쓰기에 따라 단어 수준으로 분할된다. 본 연구에서는 파이썬의 자연어 처리 패키지 중 하나인 Natural Language Toolkit (NLTK)의 Tweet Tokenizer를 사용했다(Bird et al., 2009).

불용어 처리는 텍스트 데이터의 분석 수행 시 결과물을 흐릴 수 있는 불필요한 토큰들을 제거하는 작업이다. 전치사(in, with, on 등), 관사(a, an, the 등), 지시대명사(he, they 등)와 같은 요소는 다른 토큰들에 비해 압도적으로 많이 등장하지만 의미를 가지지 않기 때문에 키워드 추출 시 잘못된 결과를 도출할 가능성이 있어, 전처리 과정에서 제거한다. 본 연구에서는 NLTK의 stopwords 패키지에 기반하여 불용어를 제거했으며(Bird et al., 2009), 특히 모든 뉴스기사에서 'You have clicked on a link to a page … on worldbank.org.'와 같은 문구가 등장했기 때문에, 이 또한 일괄적으로 제거했다.

2.3 토픽 모델링(Topic Modeling)

인터넷이 발달하면서 분석 가능한 텍스트 데이터의 양이 급증했고, 필요한 정보를 효율적으로 획득하기 위해 텍스트를 자동으로 요약하는 기술에 관한 연구가 다수 이루어졌다(Newman et al., 2006). 토픽 모델링(Topic Modeling)이란 텍스트를 요약하는 기법 중 하나로, 문서 집단에 잠재되어 있는 토픽, 즉 주제들을 도출해내기 위해 텍스트 데이터에서 단어들이 등장하는 패턴의 확률을 모형화한다(Hong and Davison, 2010). 초기에는 트위터(Twitter) 등 SNS 데이터를 분석하는 데 사용되었지만, 점차 뉴스기사 분류, 유사 논문 추천 등의 분야로 확장되었다(Hong and Davison, 2010; Newman et al., 2006; Yang et al., 2011).

토픽 모델링을 구현하는 알고리즘에는 Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM), Probabilistic Latent Semantic Indexing (pLSI) 등이 있는데, 연구 결과에 따르면 일반적으로 LDA가 가장 좋은 성능을 보이는 것으로 밝혀졌다(Hong and Davison, 2010).

2.4 잠재 디리클레 할당(Latent Dirichlet Allocation)

잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)은 2003년에 제안된 토픽 모델링 방식으로, 최근까지 가장 보편적으로 사용되는 기법 중 하나이다(Blei et al., 2003). LDA는 토픽의 단어 비중과 문서의 토픽 비중이라는 두 가지 변수의 결합 확률분포에 따라 문서의 토픽을 찾는 과정이라고 할 수 있다. 토픽의 단어 비중이란 각 토픽에서 등장하는 단어들의 비중을 뜻하며, 문서의 토픽 비중이란 문서에 등장하는 단어들이 어떤 토픽에서 등장했을 것인지를 뜻한다. 두 변수 모두 양의 실수를 요소로 가지며, 모든 요소를 더한 값이 1이 되는 디리클레(Dirichlet) 분포를 따른다.

LDA에서 가정하는 문서 생성과정이 합리적이라면 토픽의 단어 비중과 문서의 토픽 비중을 결합한 확률이 가장 클 것이고, 그 문서는 이 확률을 가장 크게 만드는 토픽에 할당된다(Newman et al., 2006). 이 내용을 수식으로 표현하면 아래의 Eq. (1)과 같다. P(t|d)는 문서 d의 토픽 t에 대한 비중을 뜻하며(Σtp(t|d)=1), p(w|t)는 토픽 t에 등장하는 단어 w의 비중을 뜻한다(Σwp(w|t)=1). 두 확률을 결합한 p(w|d)는 문서 d에서 어떤 단어 w가 등장할 것인지에 대한 비중을 뜻하며, 여기에서 높은 비중을 가지는 단어들의 토픽이 해당 문서의 토픽으로 할당된다(Newman et al., 2006).

$$p(w|d) = \sum_{t=1}^{T} p(w|t)p(t|d) \quad (1)$$

3. 연구 결과

3.1 웹 크롤링 결과

웹크롤링 기법을 사용하여 월드뱅크 뉴스 웹사이트(www.worldbank.org/news)로부터 2010년 3월 31일부터 2017년 6월 1일 사이에 발생한 11,491건의 뉴스기사를 수집했다. 월드뱅크의 뉴스기사는 주로 개발도상국에 대한 지원과 인프라 구축에 관한

내용을 다루고 있기 때문에 본 연구의 실험 데이터로 설정했다. 토큰화 및 불용어 제거를 수행한 이후, 가장 많이 등장한 50개의 단어에 대해 등장 빈도에 따라 폰트 크기를 시각화하면 Fig. 1과 같다.

서론에서 언급한 바와 같이, 'private', 'sector', 'well', 'million' 등 다양한 주제에서 포괄적으로 사용되는 단어들이 큰 비중을 차지하는 것을 확인할 수 있다. 이러한 결과만으로는 시장의 상황을 구체적으로 파악하기 어려우며, 건설 프로젝트의 의사결정을 지원하는 데 한계가 있다.

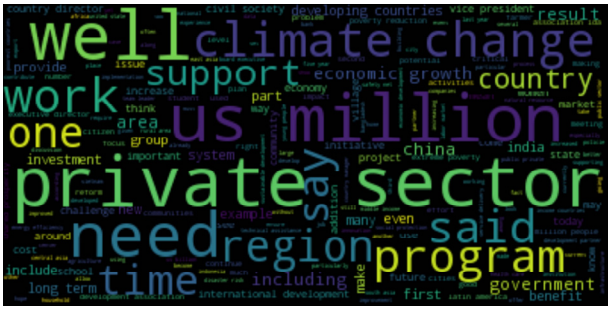

Fig. 1. Woldcloud of the Worldbank News Article

3.2 토픽 모델링 결과

문서 집단에 존재하는 주제의 개수는 10개라고 가정하고, 깁스 샘플링(Gibbs Sampling) 기법을 사용해서 각 토픽의 키워드를 추출했다. 깁스 샘플링이란, 모든 키워드에 임의의 토픽을 지정한 뒤에 토픽-단어 확률과 문서-토픽 확률을 반복적으로 계산하여 결과값을 조금씩 수정하는 방식이다. 반복의 횟수는 1,000으로 설정했다.

각 토픽에서 큰 비중을 차지하는 키워드들은 Table 1과 같이 도출되었다. 2번 토픽은 'project', 'million', 'improve', 'sector', 'service' 등의 키워드를 포함하고 있는 것으로 보아, '이용자들의 편의 증진을 위한 프로젝트'와 관련된 주제라고 추측할 수 있다. 4번 토픽은 'Africa', 'development', 'support', 'poverty', 'private' 등의 키워드를 포함하고 있는 것으로 보아, '아프리카 지역의 빈곤 문제를 해결하기 위한 민간 차원의 지원'과 관련된 주제라고 추측할 수 있다. 7번 토픽에는 'de', 'la', 'que', 'en', 'des' 등의 키워드가 주로 등장했는데, 스페인어나 라틴어 등 영어가 아닌 언어에서 사용되는 단어들이 모여 있는 것을 확인했다.

각 문서의 토픽 비중은 Table 2와 같이 도출되었다. 셀의 숫자는 각 토픽이 해당 문서에서 차지하는 비중을 의미한다. 한 문서 안에서 토픽의 비중을 비교하는 것은(열간 비교) 유의하지만, 한 토픽

Table 1. Topic-Word Count Scores

Topic	Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	…
1	Development	Public	Support	Government	Information	…
2	Project	Million	Improve	Sector	Service	…
3	Climate	Energy	Change	City	Carbon	…
4	Africa	Development	Support	Poverty	Private	…
5	People	Think	Know	Need	Time	…
6	Water	Project	Farmer	Community	Land	…
7	De	La	Que	En	des	…
8	Business	Report	Economy	Financial	Regulatory	…
9	Health	Education	Children	School	Social	…
10	Growth	Economic	Region	Investment	policy	…

Table 2. Document-Topic Count Scores

Doc	Topic									
	1	2	3	4	5	6	7	8	9	10
1	13	98	1	184	0	38	0	0	6	0
2	0	0	1	0	17	16	0	0	298	87
3	241	0	0	170	31	1	0	0	0	0
4	152	130	0	44	0	2	0	14	0	69
5	33	0	11	60	259	0	1	12	0	0
…	…	…	…	…	…	…	…	…	…	…

안에서 문서의 점수를 비교하는 것은(행간 비교) 의미가 없다. 3번 문서는 우크라이나의 전력 문제를 해결하기 위해 정부와 민간 업체가 협력하여 추가 발전소를 건설하는 내용의 뉴스기사인데, 1번 토픽(정부, 개발, 지원 등)과 4번 토픽(아프리카, 빈곤, 민간)의 점수가 높은 것을 확인할 수 있다.

4. 결 론

본 연구는 해외건설시장의 상황 정보를 파악하기 위해 뉴스기사 데이터를 분석하는 연구의 초기 단계로서, 뉴스기사 문서 집단에 존재하는 주제를 찾아내고 각 기사를 주제에 따라 분류했다. 이를 통해 현지 시장에서 어떤 주제가 이슈화되고 있는지 파악할 수 있으며, 이러한 정보는 건설 프로젝트의 기획 및 수행 시 유용하게 활용될 수 있다.

향후 연구에는 토픽 모델링을 고도화하기 위해 품사 태깅 등 추가적인 전처리를 수행하고, 다른 기관의 뉴스 데이터를 추가하며, 문서 집단에 존재하는 주제의 개수도 최적화하는 과정이 필요하다. 또한, 본 기술의 실무 활용도를 높이기 위해 로봇 저널리즘 기술과 연계하여 현지 시장에서 주로 생성되는 뉴스기사의 주제와, 각 주제에서 중요하게 여겨지는 내용에 대한 설명을 자동적으로 생성해주는 연구를 수행할 계획이다.

감사의 글

본 연구는 국토교통부 국토교통과학기술진흥원 건설교통기술촉진연구사업의 연구비지원에 의해 수행되었습니다(17CTAP-C114956-02)

References

Bird, S., Loper, E. and Klein, E. (2009). “Natural language processing with python.” *O’Reilly Media Inc*.

Blei, D. M., Jordan, M. I. and Ng, A. Y. (2003). “Latent dirichlet allocation.” *The Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Ferreira, R., Freitas, F., Cabral, L. de S., Lins, R. D., Lima, R., Franca, G., Simske, S. J. and Favaro, L. (2014). “A context based text summarization system.” *2014 11th IAPR International Workshop on Document Analysis Systems, IEEE*, pp. 66-70.

Gambhir, M. and Gupta, V. (2017). “Recent automatic text summarization techniques: A survey.” *Artificial Intelligence Review*, Vol. 47, No. 1, pp. 1-66. DOI: 10.1007/s10462-016-9475-9.

Goldszmidt, R. G. B., Brito, L. A. L. and de Vasconcelos, F. C. (2011). “Country effect on firm performance: A multilevel approach.” *Journal of Business Research*, Vol. 64, No. 3, pp. 273-279. DOI: 10.1016/j.jbusres.2009.11.012.

Hong, L. and Davison, B. D. (2010). “Empirical study of topic modeling in Twitter.” *Proceedings of the First Workshop on Social Media Analytics*, ACM Press, New York, New York, USA, pp. 80-88.

International Contractors Association of Korea (ICAK). (2018). Available online: <http://www.icak.or.kr/sta/sta_0101.php> (31/03/2018).

Javernick-Will, A. N. and Scott, W. R. (2010). “Who needs to know what? institutional knowledge and international projects.” *Journal of Construction Engineering and Management*, Vol. 136, No. 5, pp. 546-557. DOI: 10.1061/(ASCE)CO.1943-7862.0000035.

Kim, D. Y., Han, S. H., Kim, H. and Park, H. (2009). “Structuring the prediction model of project performance for international construction projects: A comparative analysis.” *Expert Systems with Applications*, Vol. 36, pp. 1961-1971. DOI: 10.1016/j.eswa.2007.12.048.

Lee, K. W., Han, S. H., Park, H. and Jeong, H. D. (2015). “Empirical analysis of host-country effects in the international construction market: An industry-level approach.” *Journal of construction engineering and management*, Vol. 142, No. 3, DOI: 10.1061/(ASCE)CO.1943-7862.

Manning, C. D., Raghaven, P. and Schutze, H. (2008). *Introduction to Information Retrieval*, Cambridge University Press.

Newman, D., Chemudugunta, C., Smyth, P. and Steyvers, M. (2006). “Analyzing entities and topics in news articles using statistical topic models.” *International Conference on Intelligence and Security Informatics*, Springer, Berlin, Heidelberg, pp. 93-104.

Pal, A. R. and Saha, D. (2014). “An approach to automatic text summarization using wordnet.” *Advance Computing Conference (IACC), 2014 IEEE International*, pp. 1169-1173. DOI: 10.1109/IAdCC.2014.6779492.

Taroun, A. (2014). “Towards a better modelling and assessment of construction risk: Insights from a literature review.” *International Journal of Project Management*, Vol. 32, pp. 101-115. DOI: 10.1016/j.ijproman.2013.03.004.

Yang, T. I., Torget, A. J. and Mihalcea, R. (2011). “Topic modeling on historical newspapers.” *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics, pp. 96-104.