**Analysis of Fire Accident Factors on Construction Sites**

**Using Web Crawling and Deep Learning**

**딥러닝으로 바꿨다 ㅋㅋ**

Jaehong Kim1, Sangpil Youm2, A 3, A4

1 PhD Student, School of Civil & Environmental Engineering, Oklahoma State University, 207 Engineering South, Stillwater, OK 74078, PH (405) 780-3866; Email: [jaehong.kim@okstate.edu](mailto:jaehong.kim@okstate.edu)

2 여긴 니 정보 적어놔라 ㅋㅋ

3

4

5

6

\*Corresponding author:

**ABSTRACT**

Recently,

* 이번 주말에 대략적인 앱스트랙 잡아 놓으께

**KEYWORDS:** Construction sites, Safety, Fire accidents, Web crawling, Deep learning

1. **Introduction**

In recent years

* 사람마다 다른데, 나는 인트로를 젤 마지막에 적는다 ㅋㅋ 왜냐하면 분석하고 작성하다보면 뒷내용이 자꾸 바뀌는데 미리 적어놓으면 나중에 다시 수정해야함.

1. **Background (Literature Review)**

* 일단 아래 3개 테마로 해서 각 10개씩 논문 찾아서 정리하면 될 듯

***Construction Safety /*** Construction site accidents

* 슬슬 관련 논문들 찾아놓고 한두 문장씩 쓰면 좋을듯
* 레퍼런스들은 목록주면 End note로 정리하면 될듯

***Data collection - Web crawling***

The

* 논문 10개 정도 + 내용 한페이지

***Word Embedding (Word2Vec)***

To

* 논문 10개 정도 + 내용 한페이지

1. **Methodology**

* 일단 대략적인 가이드라인을 잡아 놨음. 혹시 더 추가해야 되는 방법론이나 분석방법이 있으면 자유롭게 추가 가능
* 데이터 이용한 논문들 읽어보니, 방법론 하나하나 설명할때 수학적 설명 + 다른논문 레퍼런스를 꼭 넣더라.

***Data collection procedures and sample***

Based on

***Web crawling***

* ***Selenium***
* ***Beautiful Soup***
* ***Data facts***

|  |  |
| --- | --- |
| Issued duration | 2000.01.01 – 2020.04.22 |
| News Source | The New York Times |
| Number of Article | 861 |
| Number of words | 453,283 |

Table 1 : news data information

* 여기서 기간을 2019.12.31일로 해도 괜찮나? 뭔가 딱 20년이 나을듯
* 그리고 조금 필요없는 정보라도 data information을 추가하는게 좋을듯. 우리가 말하고 있는 기사의 계절별 뭐 요일별 등등 추가정보(같이 생각해보자)
* 그리고 이게 원래부터 861개가 아니자나. 처음에 추출된게 몇개였는데 블로그 형식 빼고 이렇게 해서 최종적으로 861개의 기사를 사용했다. 이런식의 설명이랑 표가 필요할듯
* ***Data preprocessing***

The

* 니가 밑에 쭉 나열해놓은 방법론들 중에 묶을수 있는건 묶어서 설명하는 것이 좋음. 이론적인 설명이 꼭 들어가야하고 레퍼런스 필요

***Word Embedding - Word2Vec***

***- Why Word2Vec***

***- Theoretical grounds in Word2Vec***

***-- neural network***

***-- skip-gram vs cbow***

***- Setting of Word2Vec***

***- cosine similarity***

The

***UMAP***

* ***Why UMAP***
* ***Dimensional Reduction***

The

1. **Results**

* 결론 부분에는 그래프, 표, 이미지 다수 포함. 소제목들은 일단 그래프 표 등 정리한 후에 결정

1. ***Basic statistical facts***Figure 1 : frequency of keywords

* 아직 어디에 제출할지 모르니, 모든 그래프는 흑백으로 작성하는게 좋을듯
* X축 y축 설명 넣어야하고, 각 그래프 위에 몇개인지 숫자를 적어주면 좋을듯
* 모든 그래프, 표에는 citation이 입력되니 이미지에 제목을 넣을 필요는 없음.

1. ***Word Embedding with word2vec***
   1. ***Retrieve similar words by cosine similarity***

***2.1.1 similarity between keywords***

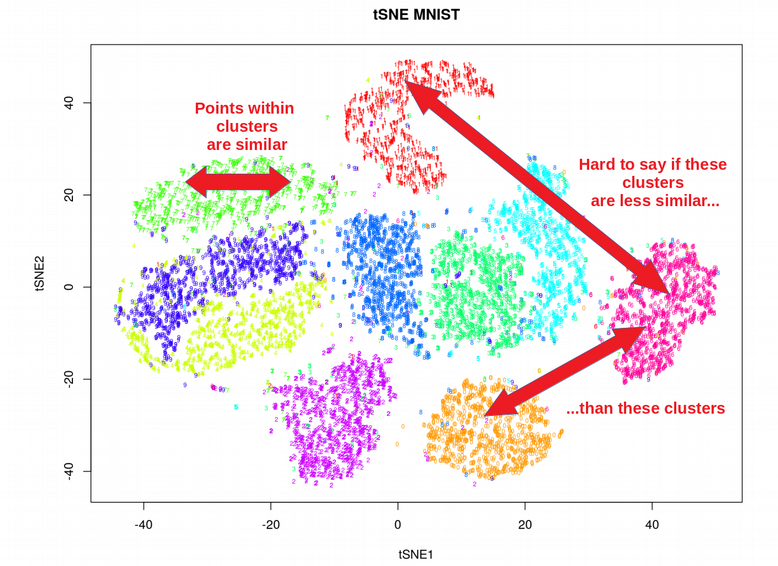
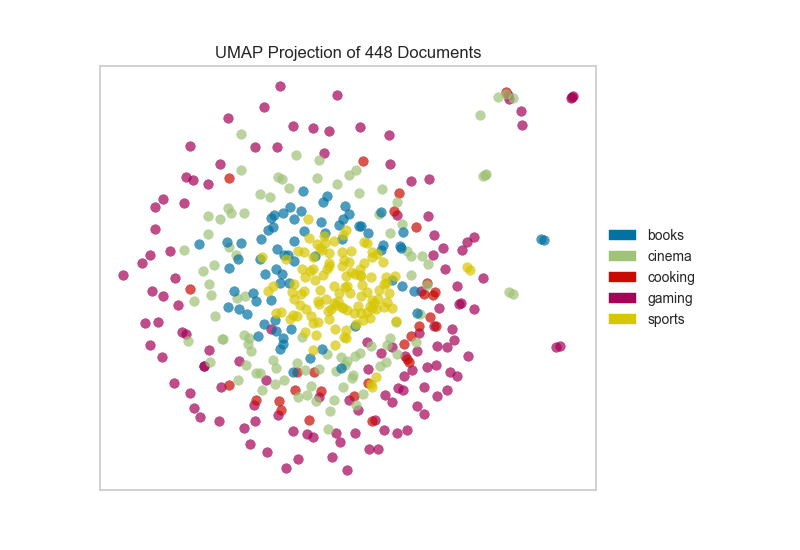
Table 2 : similarity between keywords

* 표는 이미지로 만들지 말고, 엑셀이나 작업한곳에서 그대로 복사해서 가져오는게 나중에 편집할때 좋음
* 이것도 그림으로 나타낼수 있을지에 대한 고민 필요할듯
* 모든 그래프, 표에는 citation이 입력되니 이미지에 제목을 넣을 필요는 없음.

***2.1.2 Visualizing with UMAP***



Figure 2 : word2vec result overview with umap

* 이게 올바른 예인지는 모르겠는데 이런식으로 뭔가 선명한게 논문에는 좋다. 이 그래프는 컬러가 들어가도 괜찮을듯.
* 키워드 말고 다른 단어들은 저런식으로 굵은 점으로 표현해도 좋을것 같다. 각 키워드에 종속되는 단어들 색깔을 달리해서 점으로 표현하고 키워드는 글자 크게 보여주면 될듯
* 어차피 이거 다음에 각각 키워드를 보여줄거라 여기서는 큰 그림이 어떻게 생긴지만 선명하게 보여주면 될듯
* 이 그래프가 논문의 하이라이트로 보일 수 있게 눈에 띄어야함 ㅋㅋㅋㅋㅋㅋㅋ

Subgraphs for keywords

A close up of text on a white background

Description automatically generated

Figure 3 : subgraphs for keywords

Figure 3 shows graph confining the scope of keywords.

* 일단 나중에 합치더라도, 처음엔 5개 키워드 이미지를 각각 크게 만드는게 좋을듯
* 최대한 각 5개 그래프가 각각 선명하게 보이는게 좋음
* 이 그래프에서는 근처의 단어들도 명확하게 보이는 것이 좋음
* 중간중간 색깔이 다른 단어들이 있는데, 이 그래프에서는 키워드 단어를 제외하고는 검정으로 표현하는 것이 나을듯.

***2.1.3 Similar words of keywords***

***- network analysis (Threshold : cosine similarity 0.5)***



Figure 4 : network of keyword's similar words

The size of nodes varies depend on degree.

* 각 키워드 옆에 몇 단어가 링크되어 있는지 표시
* 링크가 많을수록 진해진다고 했는데, 진하기 차이를 더 많이 주면 좋을듯
* 각 키워드별로 나눠서 보여줘야할듯 – 각 키워드별로 보여줄때는 각 링크된 단어도 식별이 가능하면 좋을듯
* 네트워크 분석을 UMAP 앞에 넣을지 Umap 뒤에 넣을지 결정. 나는 마지막에 유맵을 보여줄 생각이었는데, 분석쪽인 니가 전문가니까 의견을 도
* ***Top 20 list of similar words of keywords***



Table 3 : top 20 similar words

* 표는 이미지로 만들지 말고, 엑셀이나 작업한곳에서 그대로 복사해서 가져오는게 나중에 편집할때 좋음
* 이거는 내가 정리한걸로 버릴꺼 제외한 Top 20을 보여주는게 나을듯
* 유사 키워드나 우리가 정한 5개 키워드 색깔 표시 이런건 더 고민해보자. 엑셀에서 바로 붙여놓으면 수정하기 그나마 편함
* 모든 그래프, 표에는 citation이 입력되니 이미지에 제목을 넣을 필요는 없음.

1. **Discussion**
2. **Conclusion**

By investigating firefighters' perceptions of construction site fire accidents that require consideration of various characteristics and by exploring the structural relationships between the elements, it is possible to make useful data in developing more realistic and differentiated policy proposals.

**REFERENCES**