# Word2Vec/Glove/FastText를 활용한 임베딩 구축과 한국어 특성에 맞는 패러미터 구축

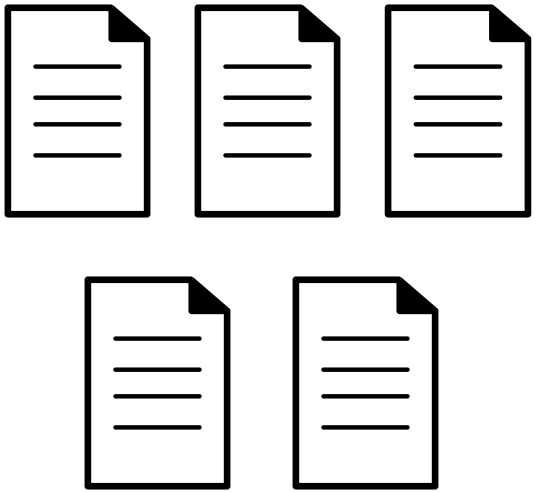
김재영
조민철

# I. Introduction

# Task description

한국어 뉴스 기사를 모아 놓은 기사 파일을 대상으로 word vector를 만드는 3가지 알고리즘 Word2Vec, GloVe, FastText를 사용하여 최적의 embedding size와 iteration을 찾고, word similarity와 word analogy를 통해 세 모델을 비교한다.

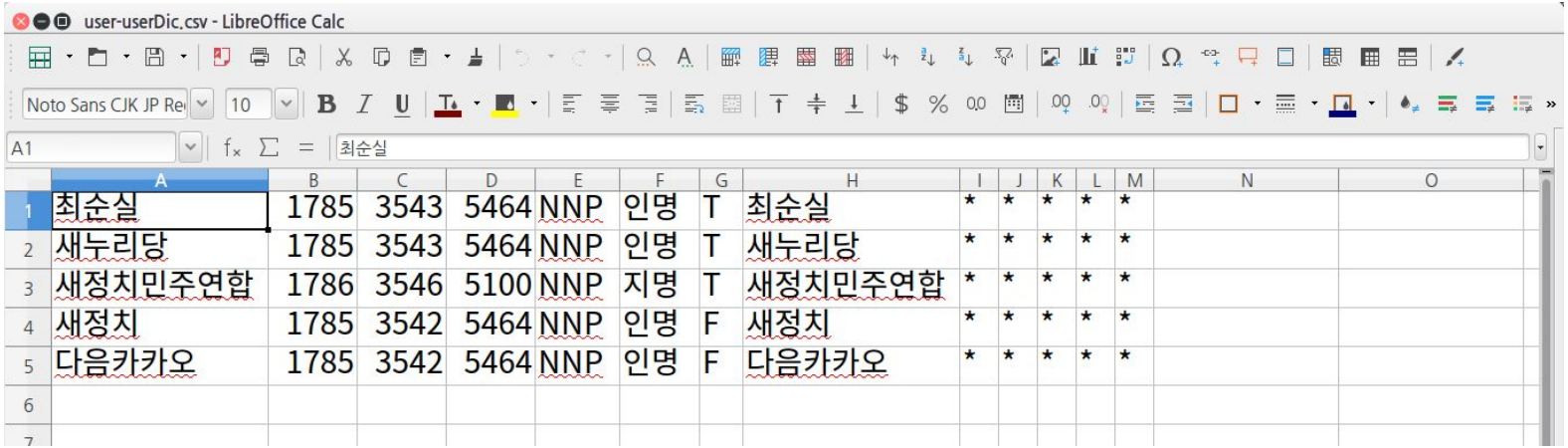
# Data description



- chosun_full.txt
- donga_full.txt
- hani_full.txt
- joongang_full.txt
- kh_full.txt

# Cleaning

- MeCab 사전에 고유명사 추가




| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 최순실 | 1785 | 3543 | 5464 | NNP | 인명 | T | 최순실 | * | * | * | * | * | | |
| 2 | 새누리당 | 1785 | 3543 | 5464 | NNP | 인명 | T | 새누리당 | * | * | * | * | * | | |
| 3 | 새정치민주연합 | 1786 | 3546 | 5100 | NNP | 지명 | T | 새정치민주연합 | * | * | * | * | * | | |
| 4 | 새정치 | 1785 | 3542 | 5464 | NNP | 인명 | F | 새정치 | * | * | * | * | * | | |
| 5 | 다음카카오 | 1785 | 3542 | 5464 | NNP | 인명 | F | 다음카카오 | * | * | * | * | * | | |
| 6 | | | | | | | | | | | | | | | |

# Cleaning

- 형태소 분석 결과 수정
  - "새" / "정치" => "새정치"

```
for news_name in ['chosun_full', 'donga_full', 'hani_full', 'joongang_full', 'kh_full'] :
    with open('./Data/news/'+ news_name + '.txt', 'r') as f:
        lines = f.read()
        sentences = lines.split('.')

        for i in range(len(sentences)):
            dic = mecab.morphs(sentences[i])
            dicSize = len(dic)

            for idx in range(dicSize-1):
                if(idx==dicSize):
                        break

                else:
                    if(dic[idx]=='새' and dic[(idx+1)]=='정치') :
                        dic.remove('새')
                        dic.remove('정치')
                        dic.insert(idx,"새정치")
                        dicSize = dicSize-1

                text.append(dic)
    print(news_name + "finished")
```

# II. Method

# Word2Vec

```
import time

start_time = time.time()


modelCbow= Word2Vec(text, min_count=15, size = 100, sg=0, iter= 5) #sg=0 : cbow
modelSkip = Word2Vec(text, min_count=15, size = 100, sg=1, iter= 5) #sg=1 : skip gram
print("word2vec success")

vectorsCbow = modelCbow.wv
vectorsSkip = modelSkip.wv
print("model.wv assign success")

vectorsCbow.save('modelCbow.bin')
vectorsSkip.save('modelSkip.bin')
print("model saved success")


print("Elapsed time: %s seconds" %(time.time() - start_time))
```

# FastText

articleMorphs.txt

```
mecab = Mecab()
articlesMorphs = open("articlesMorphs.txt","w")

for news_name in ['chosun_full', 'donga_full', 'hani_full', 'joongang_full', 'kh_full'] :
    with open('./Data/news/'+ news_name + '.txt', 'r') as f:
        lines = f.read()
        sentences = lines.split('.')

        for i in range(len(sentences)):
            dic = mecab.morphs(sentences[i])
            dicSize = len(dic)

            for idx in range(dicSize-1):
                if(idx==dicSize):
                        break

                else:
                    if(dic[idx]=='새' and dic[(idx+1)]=='정치') :
                        dic.remove('새')
                        dic.remove('정치')
                        dic.insert(idx,"새정치")
                        dicSize = dicSize-1

            for word in dic:
                articlesMorphs.write("%s " % word)



articlesMorphs.close()
```

# FastText

```
skipgram = fasttext.skipgram("../articlesMorphs.txt","fastSkip",dim=100,epoch=5)
print("fastSkip model saved!!!")
cbow = fasttext.cbow("../articlesMorphs.txt","fastCbow",dim=100,epoch=5)
print("fastCbow model saved!!!")


model1 = KeyedVectors.load_word2vec_format('fastSkip.vec')
print("fastSkip model loaded")
model2 = KeyedVectors.load_word2vec_format('fastCbow.vec')
print("fastCbow model loaded")
```

# III. Results

# Word Similarity

| 모델 | Word2Vec | | | | FastText | | | |
|---|---|---|---|---|---|---|---|---|
| | CBOW | | SKIP GRAM | | CBOW | | SKIP GRAM | |
| Embedding Size | 100 | 300 | 100 | 300 | 100 | 300 | 100 | 300 |
| Iteration | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 박근혜 | 이명박 | 이명박 | 당선인 | 당선인 | 이명박 | 이명박 | 당선인 | 대통령 |
| 오바마 | 푸틴 | 푸틴 | 버락 | 버락 | 클린턴 | ・오바마 | 버락 | 버락 |
| 김정은 | 김정일 | 김정일 | 장성택 | 김정일 | 김정일 | 김정일 | 최룡해 | 김정일 |
| 아베 | 노다 | 노다 | 신조 | 신조 | ・아베 | ・아베 | 신조 | 신조 |
| 청와대 | 인수위 | 인수위 | 김기춘 | 비서관 | 와대 | 와대 | 비서관 | 비서관 |
| 백악관 | 국무부 | 국무부 | 좌관 | 버락 | ・백악관 | ・백악관 | 오바마 | 오바마 |
| 새누리당 | 민주당 | 민주당 | 새정치민주연합 | 새정치민주연합 | 새정치민주연합 | 새정치민주연합 | 새정치민주연합 | 한나라당 |
| 최순실 | 강난희 | 최옥자 | 노재헌 | 강난희 | 최순일 | 최순옥 | 최태민 | 윤회 |

# Word Analogy

| 모델 | Word2Vec | | | | FastText | | | |
|---|---|---|---|---|---|---|---|---|
| | CBOW | | SKIP GRAM | | CBOW | | SKIP GRAM | |
| Embedding Size | 100 | 300 | 100 | 300 | 100 | 300 | 100 | 300 |
| Iteration | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 서울 + 일본 - 한국 | 오사카 | 도쿄 | 종로구 | 도쿄 | 도쿄 | 도쿄 | 오사카 | 오사카 |
| 서울 + 미국 - 한국 | 워싱턴 | 워싱턴 | 맨해튼 | 뉴욕 | 강남부 | 캘리포니아주 | 맨해튼 | 뉴욕 |
| 서울 + 북한 - 한국 | 선릉역 | 이촌 | 선부 | 평양 | 이촌역 | 이촌역 | 왕재산 | 평양 |
| 박근혜 + 일본 - 한국 | 아베 | 이명박 | 대통령 | 아베 | 여당 | 이명박 | 아베 | 아베 |
| 박근혜 + 미국 - 한국 | 오바마 | 오바마 | 오바마 | 버락 | 오바마 | 오바마 | 오바마 | 버락 |
| 박근혜 + 북한 - 한국 | 여당 | 여당 | 대통령 | 당선인 | 청와대 | 청와대 | 국방위원회 | 국방위원회 |
| 새누리당 + 미국 - 한국 | 민주당 | 민주당 | 공화 | 민주당 | 민주당 | 민주당 | 민주당 | 공화 |
| 새정치 + 미국 - 한국 | 민주당 | 새정치 | 공화 | 민주당 | 민주당 | 민주당 | 민주당 | 민주당 |
| 청와대 + 미국 - 한국 | 백악관 | 백악관 | 백악관 | 백악관 | 와대 | 백악관 | 백악관 | 백악관 |
| 민주주의 + 북한 - 한국 | 자유민주주의 | 법치주의 | 적화통일 | 법치주의 | 자유민주주의 | 자유민주주의 | 적화통일 | 신민주주의 |
| 김무성 + 새정치 - 새누리당 | 김한길 | 김한길 | 김한길 | 김한길 | 김한길 | 김한길 | 김한길 | 김한길 |
| 보수 + 새정치 - 새누리당 | DJP | 진보 | 좌파 | 진보 | 진보 | 진보 | 진보 | 진보 |

# IV. Conclusion

# Word Analogy

| 모델 | Word2Vec | | | | FastText | | | |
|---|---|---|---|---|---|---|---|---|
| | CBOW | | SKIP GRAM | | CBOW | | SKIP GRAM | |
| Embedding Size | 100 | 300 | 100 | 300 | 100 | 300 | 100 | 300 |
| Iteration | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 서울 + 일본 - 한국 | 4 | 6 | 5 | 8 | 5 | 6 | 7 | 9 |
| 서울 + 미국 - 한국 | | | | | | | | |
| 서울 + 북한 - 한국 | | | | | | | | |
| 박근혜 + 일본 - 한국 | | | | | | | | |
| 박근혜 + 미국 - 한국 | | | | | | | | |
| 박근혜 + 북한 - 한국 | | | | | | | | |
| 새누리당 + 미국 - 한국 | | | | | | | | |
| 새정치 + 미국 - 한국 | | | | | | | | |
| 청와대 + 미국 - 한국 | | | | | | | | |
| 민주주의 + 북한 - 한국 | | | | | | | | |
| 김무성 + 새정치 - 새누리당 | | | | | | | | |
| 보수 + 새정치 - 새누리당 | | | | | | | | |

# Conclusion

Embedding size와 Iteration에 따라 모델 학습 시간이 매우 오래 걸려
fasttext의 경우 (embedding=300, iteration=10)에서 학습하는 데만
7시간이 넘게 걸려서, iteration=5로 설정해서 모델을 비교했다.

즉, 다양한 hyperparemter를 비교하기가 힘들었다.

하지만 몇가지 모델과 hyperparameter를 비교해본 결과
최적의 조합은 **"fasttext (300/10)"**였다.

# END