회의분석 서비스를 위한 노이즈 처리와 클러스터링

조수현

2020. 08

Outline

1. 논문 개요
2. 단어 임베딩 (Word Embedding)
   - tf-idf 를 이용한 회의 기반의 중요 키워드 추출
3. 잡담 (Noise) 처리
   - word2vec 을 이용한 블록단위 노이즈 처리
4. 클러스터링(Clustering)
   - word2vec 블록단위 유사도를 이용한 클러스터링
5. 결론

01
논문 개요

(1) 회의 분석 서비스
(2) 사용 데이터

회의 분석 서비스

사용 데이터

사용 데이터	데이터 수(문서)	특징	비고
뉴스 데이터	약 5000	분석 용이, 잘 정제된 글	크롤링하여 사용
국어국립원 전사 데이터	약 1500	화자 구분	많은 종류의 전사 파일
STT 및 전사 데이터	약 50	STT의 오류를 포함	회의 음성을 직접 녹음

```
<u who="P2"><s n="00001">예,</s>
<s n="00002">그럼 마지막으로,</s>
<s n="00003">이익환,</s>
<s n="00004">원장님의,</s>
<s n="00005">말씀을 듣겠는데요?</s>
<s n="00006">이익환 원장님은.</s>
<s n="00007">이번 행사의,</s>
```

국어국립원 전사 데이터 예시

```
그 제가 약간 이거 안넣어도 될거같은데
제생각에는 컬랙션을 나누어서
차라리 디비를 크게 만들면 퍼포먼스에는
지장이 없을것 같아요
메모리 걱정도 안해도 되고
그냥 디비를 크게만들면
거기 지원되는 메모리가 작으니까 ..
```

녹음 파일 전사 예시 – 개인별 녹음

02
단어 임베딩(Word Embedding)

(1) TF-IDF
(2) TF-IDSF

TF-IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

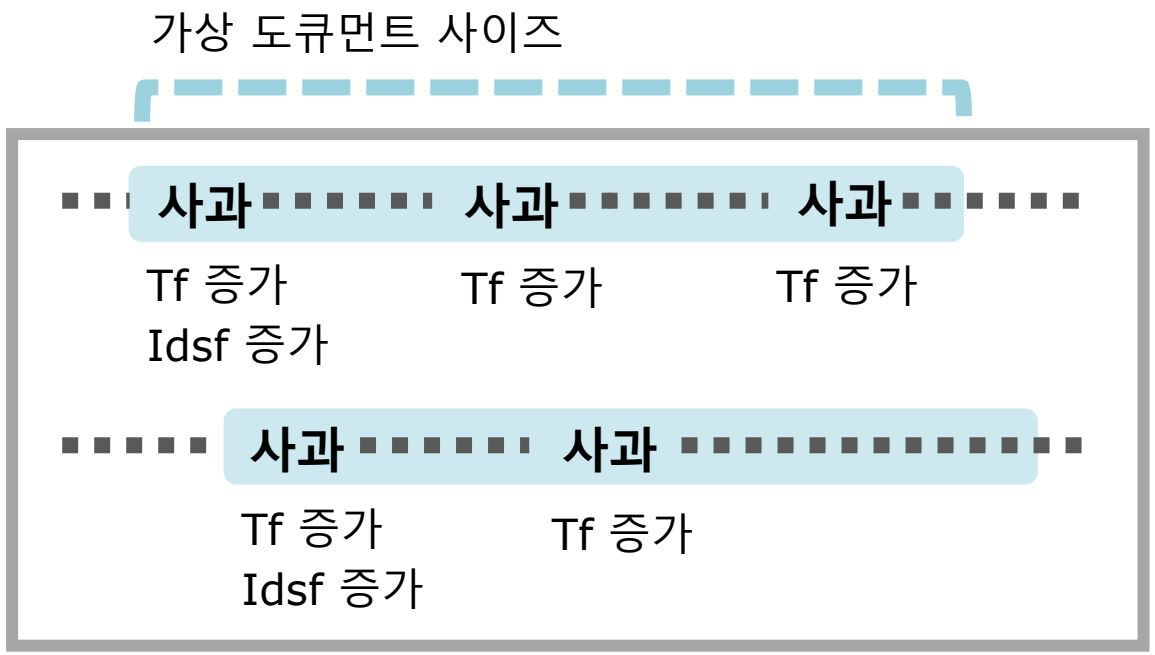
Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

TF-IDSF



TF : 5
IDSF : 2

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

$$tf \times \log\left(\frac{N}{(df \mid v)}\right)$$

TF-IDSF

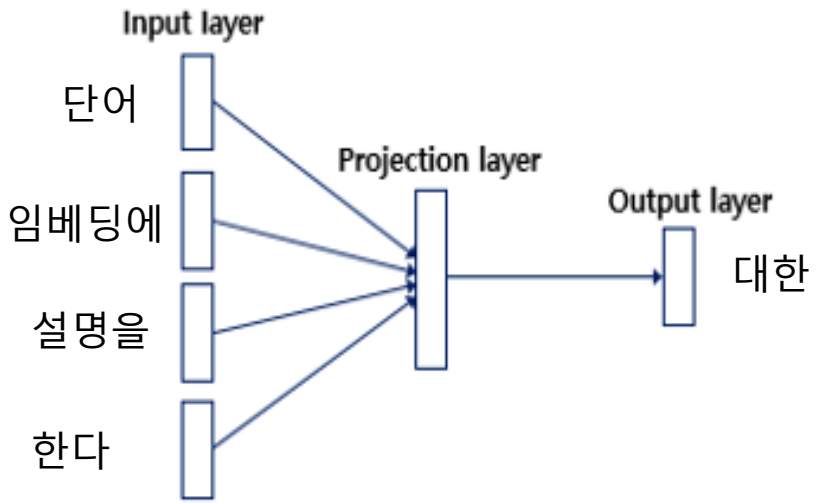
V = Virtual document size

Tern x within virtual document

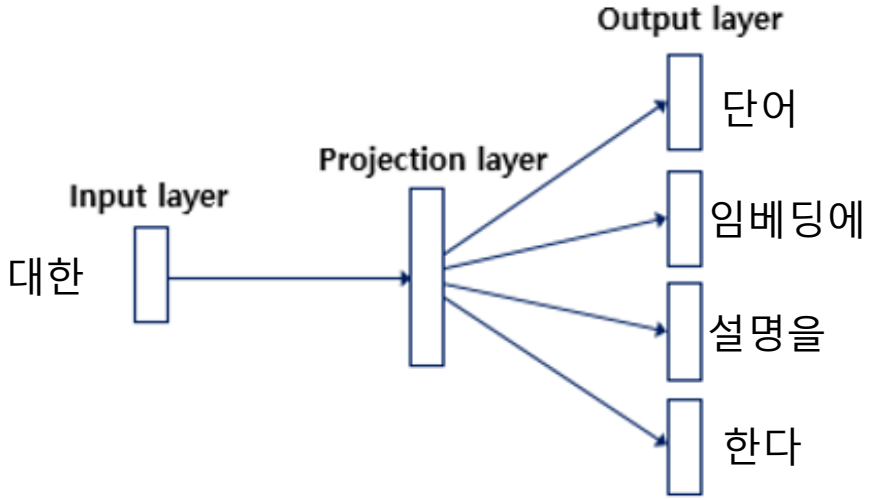
03
잡담(Noise) 처리

(1) Word2vec
(2) 토큰화(Tokenization)
(3) 토큰 유사도를 이용한 잡담 처리

word2vec



CBOW



Skip-gram

예시 문장 : 단어 임베딩에 대한 설명을 한다.

토큰화(Tokenization)

웹 개발은 이렇게 하면 되겠네 텍스트 분석의 경우에는 유사 논문이 잘 안보..

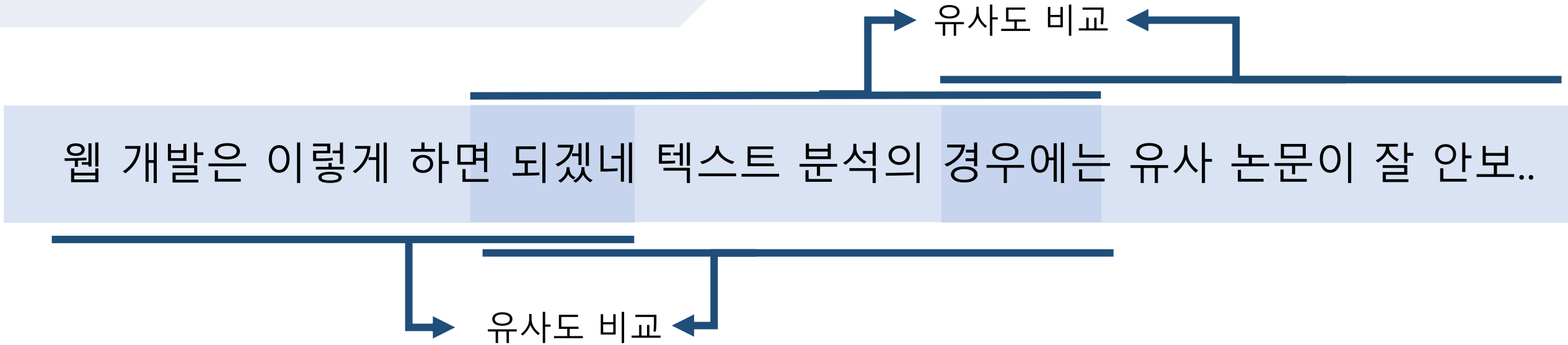
특정 개수로 자른 후 word2vec을 이용한 벡터화

웹 (0.123, …), 개발(-0.12…), …

각 벡터의 평균을 하나의 토큰으로 정의

$$\text{Token} = \frac{\sum_{k=0}^{n} word\_embedding}{word\ split\ size}$$

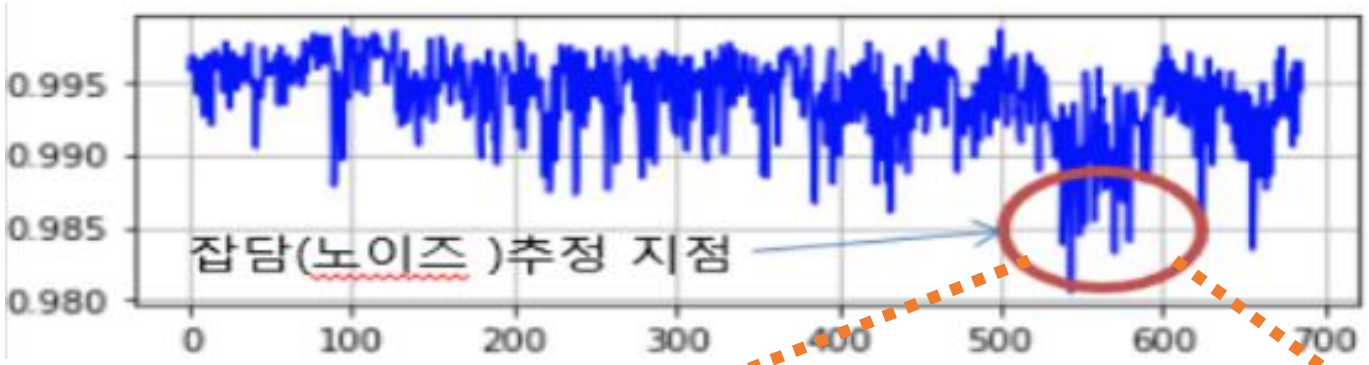
잡담 처리



Parameter

Token size	전체 단어 토큰의 수 * 0.1%	Ex) 100
step size	**Token size** 의 절반	Ex) 50

잡담 처리

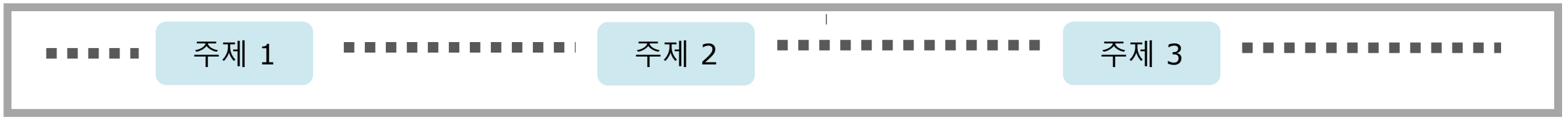


배가 고프네 회의 끝나고 곱창 아니면 피자 먹을까 여기 배달 가능한 곳인...

04
클러스터링(Clustering)

(1) 토큰 유사도를 이용한 클러스트링

클러스터링



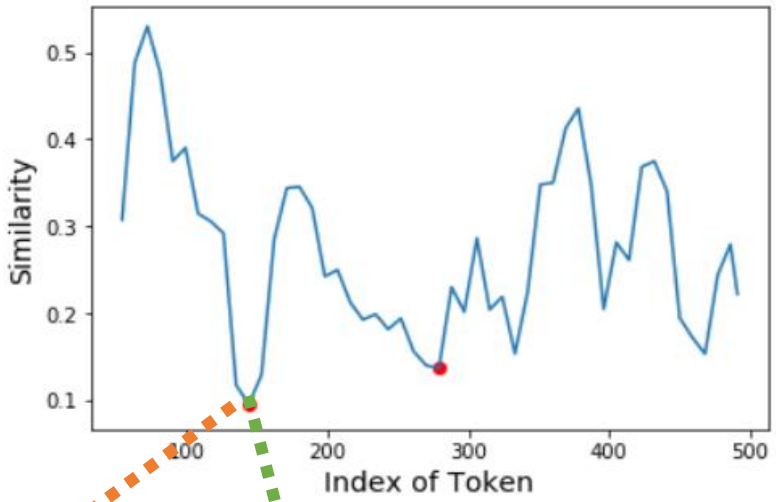
연속된 대화 데이터로 주제는 시간에 따라 변한다고 가정

Parameter

Token size	전체 단어 토큰의 수 * 5%	Ex) 100
step size	**Token size** 의 절반	Ex) 50

Token size 크게 조절

클러스터링



$\frac{1}{2}$

웹 개발은 이렇게 하면 되겠네 텍스트 분석의 경우에는 유사 논문이 잘 안보..

05
결론

(1) 기대 효과

기대 효과

1. 서비스 이용

- 회의 관리 서비스
- 회의록 요약 및 문서화
- 회의록 분석 서비스

2. 확장성

- 이전 회의와 연동하여 회의 주제 관리 시스템

-> 프로젝트 주제 분석

- 프로젝트 진행상황 도표화 및 개발 문서로도 이용 가능
- 집단(회사, 정부, 팀 단위 등)의 지식 관리 가능

감사합니다