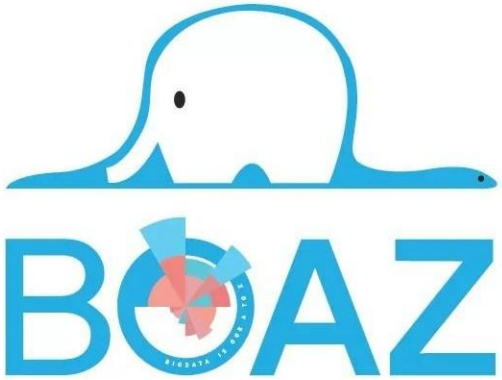
텍스트 전처리

BOAZ D조
이예림, 홍예지, 김강민, 김용규

목 차



1. 자료특성

1-1. 데이터의 종류

1) Unstructured : 비정형데이터

- 미리 정의된 데이터 모델(구조)을 가지고 있지 않은 데이터
- 빅데이터의 대부분을 차지

Ex) 텍스트, 이미지, 동영상 등

2) Semi-Structured : 반정형데이터

- 완벽한 구조는 아니지만, 어느 정도 모양이 갖춰진 데이터
- 태그를 사용하는 웹페이지에서 많이 보이는 형태

Ex) HTML, XML

3) Structured: 정형데이터

- 정해진 구조에 따라 데이터가 배열되어 있음

Ex) 데이터베이스

1-2. 데이터 탐색 방법

1) Search : 검색

- 기존 데이터를 필요에 맞게 가공하는 것

Ex) SQL Query, 웹 검색

2) Discovery : 발견

- 데이터가 갖고 있는 숨겨진 의미를 탐색하는 것

Ex) 데이터마이닝, 텍스트마이닝

1-2-1. 데이터종류 x 데이터 탐색 방법

	Structured	Semi-Structured	Unstructured
Search	데이터베이스 (Database)	정보 검색 (Information Retrieval)	
Discovery	데이터 마이닝 (Data Mining)	웹 마이닝 (Web Mining)	텍스트 마이닝 (Text Mining)

1-3. 비정형데이터 마이닝

"빅데이터에서의 데이터 마이닝은 비정형 데이터 마이닝에 초점"

- 통계 기반의 데이터 분석 도구를 사용하여 데이터 사이의 숨겨진 관계와 패턴, 경향 등을 추출

=> 비정형 데이터를 일단 정렬 과정을 통해 정형 데이터로 만든 뒤,

데이터마이닝 작업인 분류, 군집화, 회귀분석, 요약, 이상감지 등 에 적용하여 의미 있는 정보를 발굴

1-4. 텍스트 마이닝의 정의

- 텍스트 마이닝이란?

: 대규모의 텍스트에서 **고품질의 정보를 도출**하는 과정

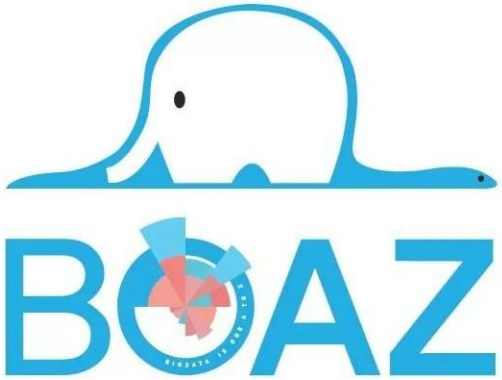
(ex. 문장의 의미 추출, 감성 분석, 초록 추출 등)

- 기법(Techniques)

- 데이터 마이닝, 머신 러닝, 정보검색, 통계학, NLP 등등
- 기존의 정형 데이터 마이닝 기법 활용

- 기능

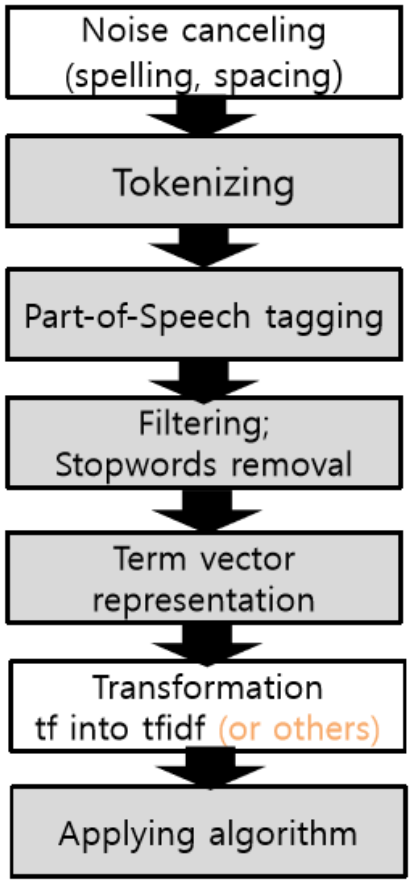
- 문서 요약, 문서 분류/군집, 특성 추출



2. 문장파싱(분해)

(spelling, spacing, tokenizing)

2-1. Text Mining process



2-2. Spelling

Spelling (철차, 맞춤법을 확인하는 단계)

- 같은 개념이 다른 단어로 표현
- 미등록단어 (Out of vocabulary) 문제

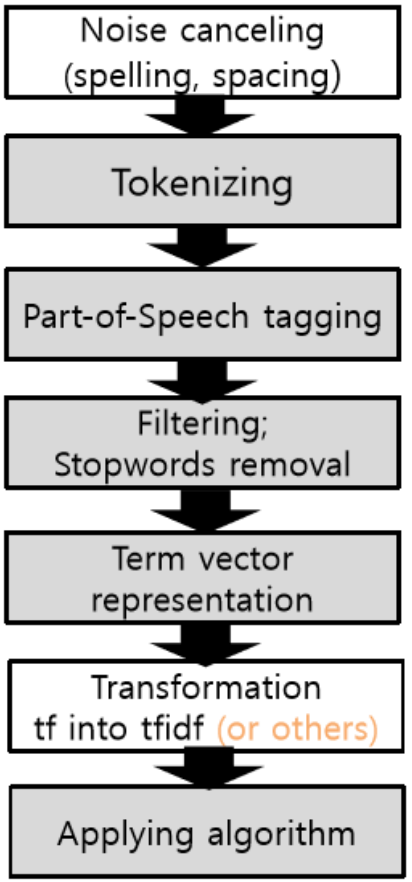
- 등록되지 않거나 희귀한 단어를,
  사전에 존재하는 올바른 단어들로 연결

-Edit distance 등으로 수정할 수 있음

(Edit distance : 단어의 유사도를 측정하는 방법)

-Word embedding 을 이용할 수도 있음

(Word embedding : 하나의 단어를 벡터 공간상의 하나의 점으로 맵핑하는 기법)



2-2. Spelling

- 수기로 입력된 데이터의 경우 자주 발생

- Edit distance등의 sting distance measure을 이용한 교정

데이터
제조외
제조, 도매, 부동산
건설업
조립금속제조, 기타화학제조
서빗스 도소매
편의점, 담배
소매업, 서비스업, 부동산업
식음료
제조 및 도소, 부동산업
제조업

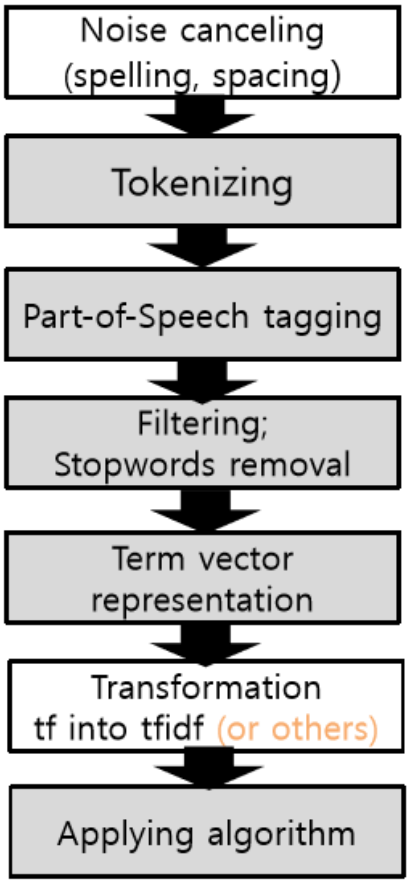
사전		
가구내 고용활동	보험	운수
가스	부동산	원료재생
개인	사업시설관리	음식점
건설	사업지원	임대
과학	사회복지	임업
광업	서비스	자가소비생산활동
교육	소매	전기
국제	수도사업	전문
금융	수리	정보
기술	숙박	제조
기타	스포츠	증기
농업	어업	출판
단체	여가	폐기물처리
도매	영상	하수처리
방송통신	예술	협회
보건	외국기관	환경복원

2-3. Spacing

Spacing

한국어의 어절은 띄어쓰기로 구분

- 띄어쓰기 오류는 자연어처리기의 정확도와 계산 시간에 영향을 줌
- 문제나 데이터 품질에 따라서는 띄어쓰기 오류를 교정해야 할 경우도 있음
- 영어의 경우 공백만으로 구분가능



2-4. Tokenizing

- 토크나이징은 어절에서 단어를 나누는 것

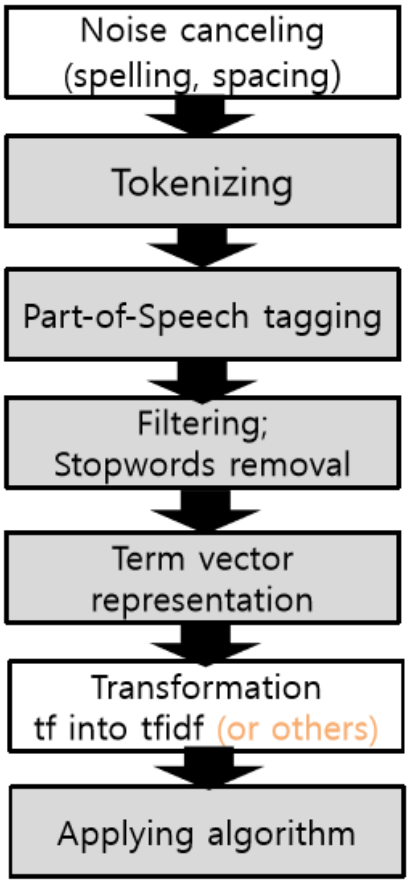
위 문장에서 [토크나이징,은,어절,에서,단어,를,나누는,것]

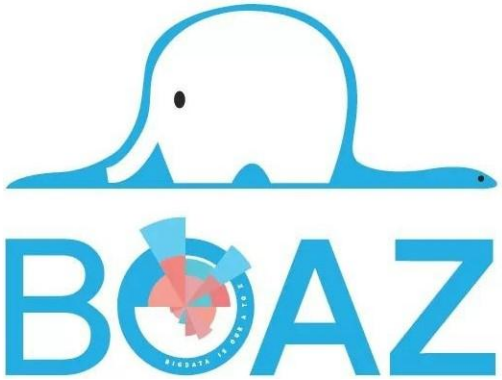
- 문제나 데이터 품질에 따라서는 띄어쓰기 오류를 교정 해야 할 경우도 있음

- 정확히는 "문장"을 "토큰"으로 나누는 것

- 토큰은 n-gram, 어절, 단어, 구 등으로 정의 가능

- 우리는 좁은 의미의 '토크나이저'로 단어를 나누는 것이라 이야기함





3. 불용어처리

3-1. 불용어 처리란

- **불용어란?**

: 주제색인어로서 의미가 없는 단어들을 의미

- 불용어 처리를 하는 이유

1. 워드 클라우드를 만들고자 할 때 필요 없거나 원하지 않는 단어 제거
2. 같은 명사인데 다른 조사나 보조사가 붙은 것들을 모두 제거하여 하나의 단어로 통일시킴

예시 - 영화리뷰

배우의 국적을 의심하게 만듬.여주를 잘 알려지지 않은 배우로 캐스팅한게 정말 탁월했습니다! 진짜 일본인이 연기하는 줄 착각할 정도로 연기가 정말 좋았
어서 신선 했어요. 배우들이 정말 일본인처럼 느껴질정도로 회화실력들도 엄청나시고 중간중간 재밌는요소도 있어서 지루하지 않게 관람하였습니다??탄탄
그냥 무자비하게 죽이기만 했을 거라 생각했지.. 일본 근대시대에 지진이 난 모습을 보여줘서 신박했다 그때도 지진이 났을텐데 왜 생각못했나 싶다진지하고
진짜큰일났을뻔잊지않겠습니다. 그리고 감사합니다.보세요. 유쾌하지만 슬픈영화..동주에 이어서 박열까지 역시나 이준익 감독님 작품은 믿보!!!!지루함... 두
일본애들은 박열말이면 쩔쩔매면서 다 들어주고, 억지감성에 연극톤 오버연기;; 말이되냐 이거 실화냐 ;;이것이 일본영화인가? 혹은 개인 자서전인가 싶을 정
옴 ㅠ너무오버해서....난별로였다우리가 당당할 수 있게 만들어준 바보 같았던 영웅들도대체 이놈의 평점 알바들은 초반은 아쉽지만이야 후반이 아쉽지만이야
. 하나부터 열까지 스토리 전개 전부 뜬금없고ㅋㄱㅋㄱ진심 너무합니다..보다가 어이가없어서 웃음이 나왔습니다..처음으로 평점 남깁니다 정말 감당할수없을
랑보러갔다가욕만먹었네요ㅡㅡ유명하지 않은 데엔 이유가 있는거다.. 딱히 뭔가를 한 건 없는 아나키스트인데 한줄 짜리 역사를 2시간 짜리 영화로 만들었으
이고 자문했다..이준익감독 리스펙〉배우와 감독의 자아도취 연기에 감탄했다. 내가 왜 박열이란 애국열사를 몰랐는지 알게해준 영화였다. 처음부터 끝까지 ㄱ
이버는 O점이 없네?이렇게 흡입력 떨어지는 영화는 본 적이 없음. 지루한 스토리 전개와 쓸데없이 배치된 촌스러운 유머. 오버스러운 캐릭터들에 가미된 과
연극보는 느낌! 잠~오랜만에본 한국영화 취지는 정말 좋으나 재미 없어요....요즘하도볼거없어서그나마기워라서봤는데그닥재미는없네요지루지루~~개지루...
영화라서 엄청 더 마음에 안든다. 내 평점에다 비공감을 눌러버려라.진심으로 이게 재밌다고? 일본인들이 예의바르고 조선인이 버릇없어보일정도로 쓰레기
점이왜케높은지이해안됨..지루해서 중간에졸고 뛰쳐나가고싶엇음....지루하고 배우들에 오버스러운 연기.여배우는 어눌하고 어색한 한국말연기할바엔 그냥 일
개인으로 미화한 영화첫 데이트였는데 개 똥밟았네 아 욕나온다 진짜 이딴 쓰레기가 재밌으면 집에서 클레멘타인이나 봐라 알바새끼들아기시간 잤어요 수면제
...태어나서 처음으로 중간에 보다가 나왔다.이제훈 팬이지만...전체적으로 흐름이 자주 끊기고 여주인공 연기는 나만 소름돋는건가?발가락이 꼼질거림..너무
돈아깝네소재,캐스팅,연기 다좋았는데 아쉬운건연출력이 ...,재미가 없음... 이런건 한줄로 정리할 수있는걸 막 미사여구가 붙고붙은 느낌임.개인적으로 후미코
때문에 영화보며 피가 끓어오르다가 식어버린채, 언제 재판 판결 나오는지만 기다리게 됨.후미코 일본인인줄알았네다 뭐야 이게?의미 있까?기점 진짜 많았는
해 워스트 무비였네요 저는모든게 어색한 영화. 으.. 내 돈...가벼움을 넘어서서 우습게 만들었다 우스워 죽겠음 쓰레기 같고 이거에 웃으면서 좋아하는 관
..어떻게평점ㅈ이이렇게높을수있는지 아이러니네애국심 유발도, 인물해석도 영 꽝일본간부들이 조선인학살을 천황암살범으로 교묘하게 비비듯, 영화감독은

기능1 . 문자열은 “ “ 처리로 없는 것처럼 눈속임이 가능

```cs
data_unlist <- gsub("영화", "", data_unlist)
```

그러나 이 방법은 사실상 비효율적

'영화'는 지워지겠지만 '영화를' 이라는 단어는 '영화'가 ” "로 대체되어 '를' 이라는 단어가 됨
'이영화를' '그영화는' 이렇게 띄어쓰기를 하지 않아 한 단어로 분류되어버린 단어가 있었다면
'영화'가 ” "로 사라졌으므로 '이를', '그는' 이라는 전혀 딴판의 단어로 바뀌어 버림

여기서 정규표현식 필요!

기능1. ‘gsup’의 바꿀 단어 안에 정규식을 이용

```
1 data_unlist = gsub(“영화\\S*", “ ", data_unlist)
```

\\S : ‘영화’ 뒤에 붙은 공백,탭,행을 제외한 모든 문자
* : 뒤에 몇 글자가 오던 상관 없다는 뜻

즉,
“영화\\S*” : ‘영화’라는 단어 뒤에 뭐가 얼마나 붙었는지 간에 라고 해석 가능

기능2. 특수문자 삭제

1. 자음 삭제(ㅋㅋ,ㅎㅎ,ㅇㅇ 등)

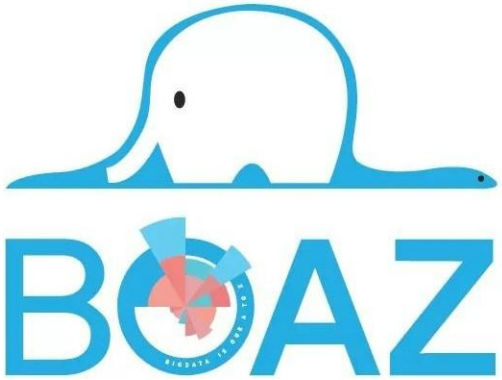
```
data_unlist <- gsub('[ㄱ-ㅎ]','',data_unlist)
```

2. ㅜㅜ나 ㅠㅠ 등 삭제

```
data_unlist<- gsub('(ㅜ|ㅠ)','',data_unlist)
```

3. 숫자 삭제

```
data_unlist <- gsub("\\d+","",data_unlist)
```



4. 품사 태깅

품사 태깅에 대해 보기 전에 형태소 분석이란?

4-1.형태소 분석이란?

1. 형태소?
: 의미의 기능을 부여하는 언어의 형태론적 수준에서의 최소 단위

2. 형태소 분석?
: 형태소를 비롯하여, 어근, 접두사/접미사, 품사 등
다양한 언어적 속성의 구조를 파악하는 것이다.

4-1. 형태소 분석

- 토큰들을 좀 더 일반적인 형태로 분석해 단어 수를 줄여 분석의 효율성을 높이는 작업

- Text Normaization이라고 부르기도 함

Word	stemming	lemmatization
Love	Lov	Love
Loves	Lov	Love
Loved	Lov	Love
Loving	Lov	Love
Innovation	Innovat	Innovation
Innovations	Innovat	Innovation
Innovate	Innovat	Innovate
Innovates	Innovat	Innovate
Innovative	Innovat	Innovative

Folding : 대문자에서 소문자로 바꿔 줌

Stemming : 단어를 축약형으로 바꿔 줌

Lemmatization : 품사정보가 보존된 형태의 기본형으로 변환하는 것

4-2. 품사 태깅이란?

- 품사 태깅?

형태소 분석의 일종으로 형태소의 뜻과 문맥을 고려하여 그것에 표시하는 일

```
from konlpy.tag import Twitter

twitter = Twitter()
print(twitter.pos(u'아버지가 방에 들어가신다.'))
```

```
[('아버지', 'Noun'), ('가', 'Josa'), ('방', 'Noun'), ('에', 'Josa'), ('들어가신', 'Verb'), ('다', 'Eomi'), ('.', 'Punctuation')]
```

4-2. 왜 품사 태깅을 하는가?

1) 단어의 순서, 배치를 통해서 그 단어의 의미, 더 나아가 문장의 의미를 알 수 있다.

2) 품사 태깅이 되었다면 좀 더 쉽게 그 단어의 의미를 알아낼 수 있다.

5.단어빈도계산

5-1. 가장 많이 사용되는 단어 추출

1.문장에서 명사를 추출한다

: KoNLP의 extractNoun()를 이용 문장에서 명사를 추출

2.데이터프레임으로 변환 추출

: 명사추출->데이터프레임으로 변환 ->변수명 변환

-> 각 단어 빈도표생성(가장 많이 사용된 20개 단어)

3. 두 글자 이상으로 된 단어만 추출

: nchar()를 이용

예시-R


```
> extractNoun("대한민국의 영토는 한반도와 그 부속도서로 한다")
[1] "대한민국" "영토"     "한반도"   "부속도서" "한"
```

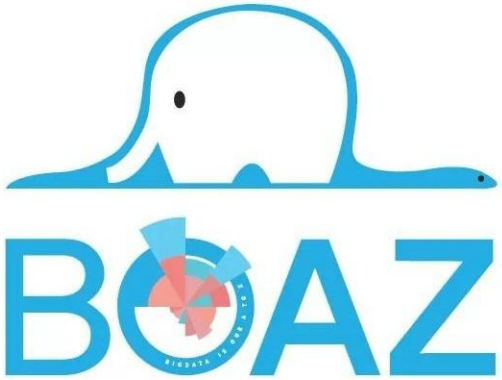
```
# 가사에서 명사추출
nouns <- extractNoun(txt)

# 추출한 명사 list를 문자열 벡터로 변환, 단어별 빈도표 생성
wordcount <- table(unlist(nouns))

# 데이터 프레임으로 변환
df_word <- as.data.frame(wordcount, stringsAsFactors = F)

# 변수명 수정
df_word <- rename(df_word,
                  word = Var1,
                  freq = Freq)
```

```
# 두 글자 이상 단어 추출
df_word <- filter(df_word, nchar(word) >= 2)

top_20 <- df_word %>%
  arrange(desc(freq)) %>%
  head(20)
```

```
> top_20
   word freq
1   you   89
2    my   86
3   YAH   80
4    on   76
5  하나   75
6  오늘   51
7   and   49
8  사랑   49
9  like   48
10 우리   48
11  the   43
12 시간   39
13 love   38
14   to   38
15   we   36
16   it   33
17   em   32
18  not   32
19 역사   31
20 flex   30
```

5-2.단어빈도 막대그래프-R


```
library(ggplot2)

order <- arrange(top20, freq)$word     # 빈도 순서 변수 생성
ggplot(data = top20, aes(x = word, y = freq)) +
  ylim(0, 2500) +
  geom_col() +
  coord_flip() +
  scale_x_discrete(limit = order) +     # 빈도 순서 변수 기준 막대 정렬
  geom_text(aes(label = freq), hjust = -0.3)     # 빈도 표시
```



6. 워드 클라우드

6-1. 워드 클라우드란?

1. 검색된 문서에서 **출연 빈도가 높은 단어**를 확인할 수 있도록 나타낸 것

2. 핵심 내용을 **빠르게 파악**할 수 있다.

3. 단어 **출현빈도**에 따라 글자의 크기가 달라진다.

감사합니다