의미를 얻을려면 데이터가 필요한데

과연 어떤 데이터의 의미를 찾아야 하는가

1.토큰화(Tokenization)

띄어쓰기 를 통해 토큰화를 했는데 = 1차원적으로 생각

띄어쓰기만으로 토큰화 를 하는게 아니라

관용구나 어절

Piece of cake = 쉽다 라는 의미 인데 토큰화를 하면 안됨

San francisco = 지역 을 나타내는 단위

Dr.A = . 도 의미에 따라 사용할수 있어서 규칙에 맞게 해야한다.

고민을 많이 해야 한다

붙으면 안돼는 case

They’ev i’ev 같은 축약어 들도 붙으면 안된다.

먹다 먹고 먹자 = case 가 많음 먹 이라는 어간을 떨어 트릴수 있는 규칙이 필요함

토큰화의 규칙에 맞는 토큰 화 라이브러리를 사용해야 한다.

1.Spacy 라이브러리

토큰화를 할떄 주문제작을 할수 있다. = 개꿀이네

단점 = 한국어 지원이 안된다. ㅅㅂ

2.Stopwords(불용어)

2-1불용어 가 필요할때

문장의 의미가 훼손되고 성능이 저하 된다.

불용어가 중요도가 떨어지지만 관용구에서는 클수도 있다.

Nltk 라이브러리

Spacy 라이브러리로 해결할수 있다.

2-2 불용어가 필요하지 않을때

왜 필요하지 않냐

Zips Law

F = C(빈도수)/rank(빈도수 순위 높으면 낮음)

즉 C 는 뒤로 갈수록 작아지고 rank 는 뒤로 갈수록 커진다.

단어의 중요도가 올라가면서 정점을찍다가 내려간다.

3.정규화

3-1stemming(어간 추출)

Complete -> complet

Exception -> except

Alize -> AL

ANce -> 제거

ICal -> IC

단순한 규칙으로 작동으로 하는 알고리즘이다

일반화 수행을 못할수도있다.

Organization -> organ(장기)

Nltk. Porterstemmer 을 통한 어간 추출을 진행한다

3-2형태 추출 (lemmatization)

단어 안에서 형태를

과거 -> 현재

Is are were -> be(원형) 으로 바꿔줌

단수,복수 -> 원형

-> 어간과 형태 추출을 하면 문장의 의미가 달라진다.(단점)