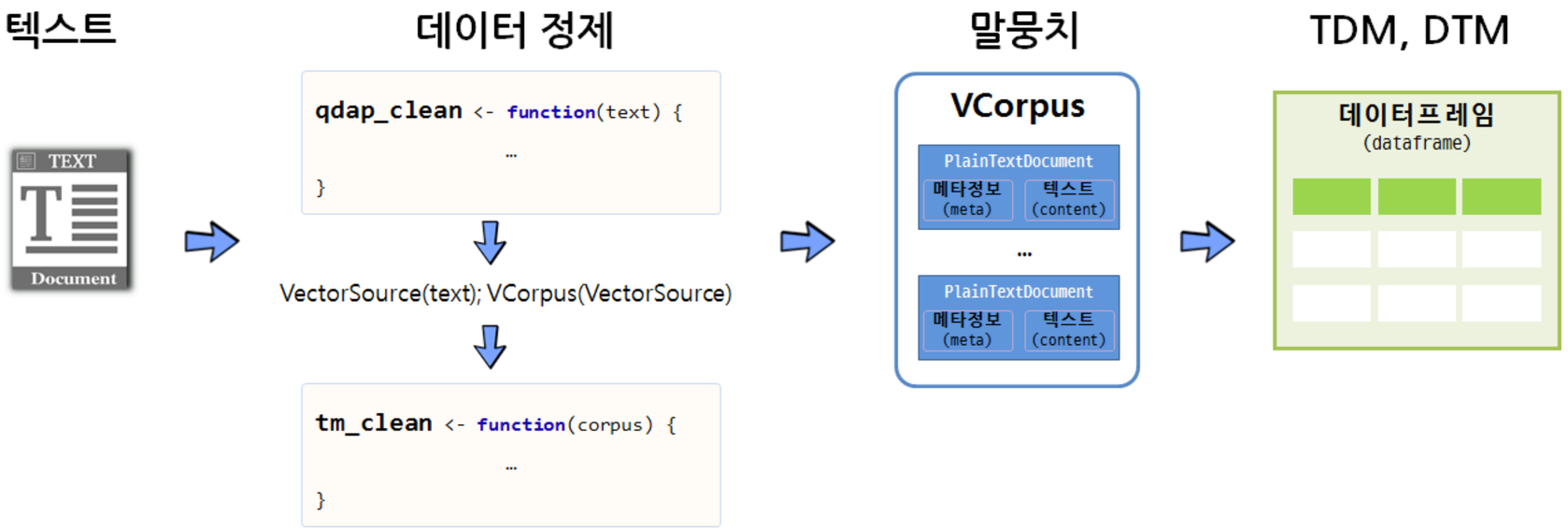
텍스트 전처리

231007_BDA 7기 파문응

텍스트 전처리란?


텍스트
데이터 정제
말뭉치
TDM, DTM
TEXT
Document
qdap_clean <- function(text) {
…
}
VectorSource(text); VCorpus(VectorSource)
tm_clean <- function(corpus) {
…
}
VCorpus
PlainTextDocument
메타정보
(meta)
텍스트
(content)
…
PlainTextDocument
메타정보
(meta)
텍스트
(content)
데이터프레임
(dataframe)

토큰화

Text 문서

restrained enthusiasm catch from one bystander to another. They swing and bow to right and left, in slow time to the piercing treble of the Congo women. Some are responsive; others are competitive. Hear that bare foot slap the ground! one sudden stroke only, as it were the foot of a stag. The musicians warm up at the sound. A smiting of breasts with open hands begins very softly and becomes vigorous. The women's voices rise to a tremulous intensity. Among the chorus of Franco-Congo singing-girls is one of extra good voice, who thrusts in, now and again, an improvisation. This girl here, so tall and straight, is a Yakoff. You see it in her almost Hindu features, and hear it in the plaintive melody of her voice. Now the chorus is more piercing than ever. The women clap their hands in time, or standing with arms akimbo receive with faint courtesies and head-liftings the low bows of the men, who deliver them swinging this way and that.

See! Yonder brisk and *slavey* fellow has taken one short, nervy step into the ring, chanting with rising energy. Now he takes another, and stands and sings and looks here and there, rising upon his broad toes and sinking and rising again, with what wonderful lightness! How tall and lithe he is. Notice his brown shining through his rags. He too is a *zande*, and by the three long rays of tattooing on each side of his face, a Kianiba. The music has got into his feet. He moves off to the farther edge of the circle, still singing, takes the prompt hand of an unsmiling Congo girl, leads her into the ring, and, leaving the chant to the throng, stands her before him for the dance.

Will they dance to that measure? Wait! A sudden frenzy seizes the musicians. The measure quickens, the swaying, attitudinizing crowd starts into extra activity, the female voices grow sharp and staccato, and suddenly the dance is the furious Bamboula.

데이터 사전 가공 후
Feature Vectorization 수행

Feature Vectorization

Bag of Words

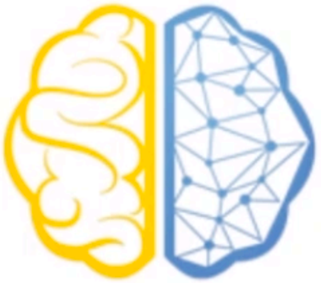
단어 #1	단어 #2	.....	단어 #n
3	4	0	1

또는

Word2Vec

Feature 기반의
데이터 셋 제공

ML 학습/예측/평가



다양한 토큰화 방법

문장 토큰화

단어 토큰화

정규 표현식을 이용한 토큰화 등

어간(stem) 추출 - 어형이 변형된 단어로부터

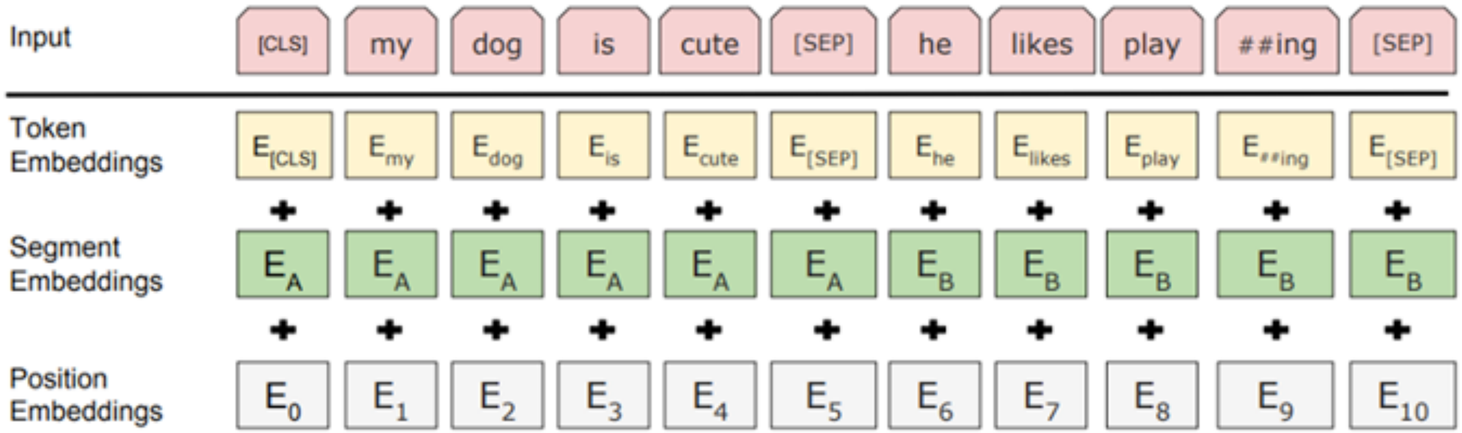
접사 등을 제거하고 그 단어의 어간을 분리해 내는 작업

표제어 추출 (Lemmatization) - 주어진 단어 기본형으로 변환 등

Believe, belief believe

	의미	기능	형태
단어	명사	체언	불변어
	대명사		
	수사		
	관형사	수식언	
	부사		
	조사	관계언	
	감탄사	독립언	
	동사	용언	가변어
	형용사		

BERT 간단한 예시


Input
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
Token Embeddings
E[CLS] Emy Edog Eis Ecute E[SEP] Ehe Elikes Eplay E##ing E[SEP]
Segment Embeddings
EA EA EA EA EA EA EB EB EB EB EB
Position Embeddings
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10