NLP – 텍스트 전처리

Index

Preprocessing

전처리

텍스트 전처리 (Text Preprocessing)

- 자연어 처리를 위해 용도에 맞도록 사전에 표준화 하는 작업

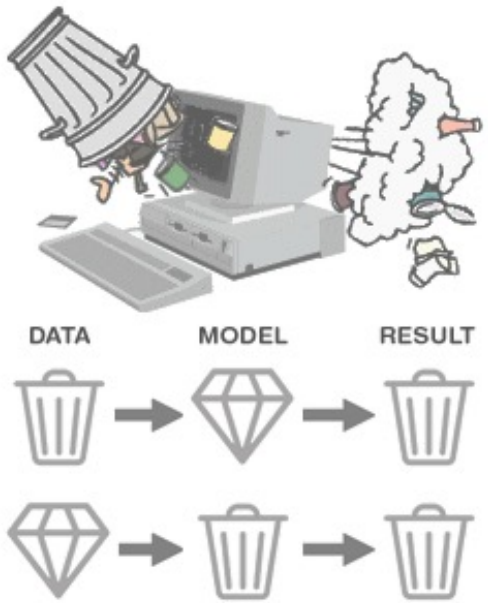
필요성

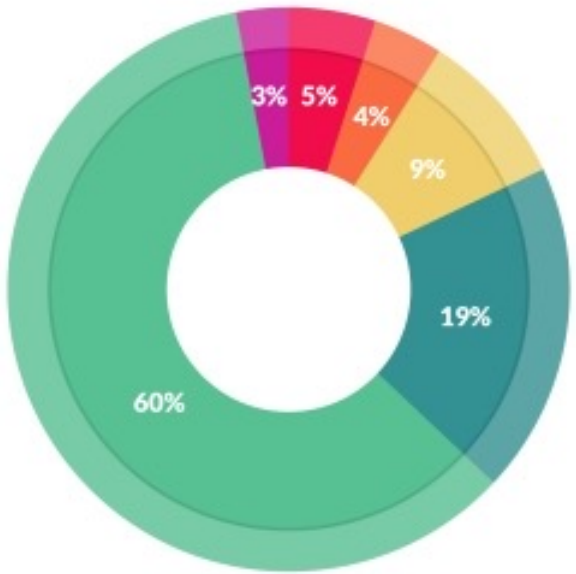
- 텍스트 내 정보를 유지하고, 분석의 효율성을 높임

- 분석하기 전 텍스트를 분석에 적합한 형태로 변환하는 작업
- 전처리 단계는 텍스트를 토큰화하고 자연어 처리에 필요 없는 조사, 특수문자, 불용어의 제거과정을 포함
- 전처리는 분석결과의 모델 성능에 직접 영향을 미치기 때문에 전처리 단계는 매우 중요

전처리의 중요성

- garbage in garbage out
- 데이터 과학자들은 그들의 79%의 시간을 데이터 전처리에 사용





What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*

79%

- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Tokenization

토큰화

토큰화

- 구두점으로 문서를 문장으로 분리하는 **문장 토큰화**
- 단어 단위로 분리하는 **단어 토큰화**

문장 토큰화 (Sentence Tokenization)

- 문장(Sentence)를 기준으로 토큰화
- 온점(.), 느낌표(!), 물음표(?) 등으로 분류하면 해결 될 것으로 생각됨
- 하지만 단순하게 분리할 경우 정확한 분리가 어려움

Barack Obama likes fried chicken. He don't like spicy chicken.

Barack Obama likes fried chicken.

He don't like spicy chicken.

단어 토큰화 (Word Tokenization)

- 단어(word)를 기준으로 토큰화
- 영문의 경우 공백을 기준으로 분리하면 유의미한 토큰화가 가능
- 반면 한글의 경우 품사를 고려한 토큰화(=형태소분석)가 필요

영문 토큰화

Barack Obama likes fried chicken very much.

Barack	Obama	likes	fried	chicken	very	much	.

한글 토큰화

버락 오바마는 후라이드 치킨을 너무 좋아한다.

버락	오바마	는	후라이드	치킨	을	너무	좋아한다	.

단어 토큰화 고려사항

- 특수문자 여부
  (구두점 및 특수문자를 단순하게 제외해서는 안됨)

특수문자	원문	토큰화 예제1	토큰화 예제2
'	Don't	Do/n't	Don/'/t
-	State-of-the-art	State/of/the/art	State-of-the-art

- 단어 내 띄어쓰기가 있는 경우

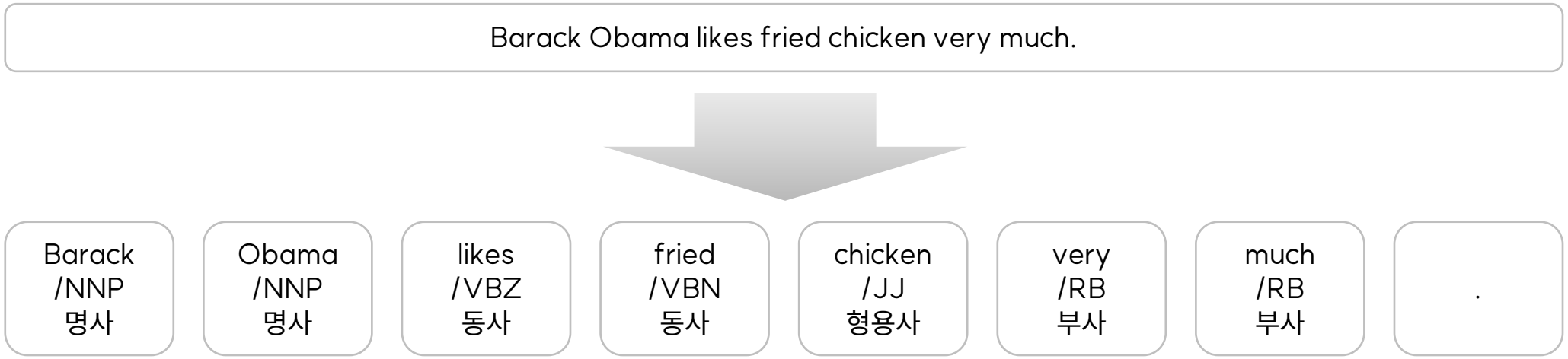
	원문	토큰화 예제1	토큰화 예제2
공백	New York	New/York	New York

Processing

토큰 처리

품사 태깅 (Pos Tagging)

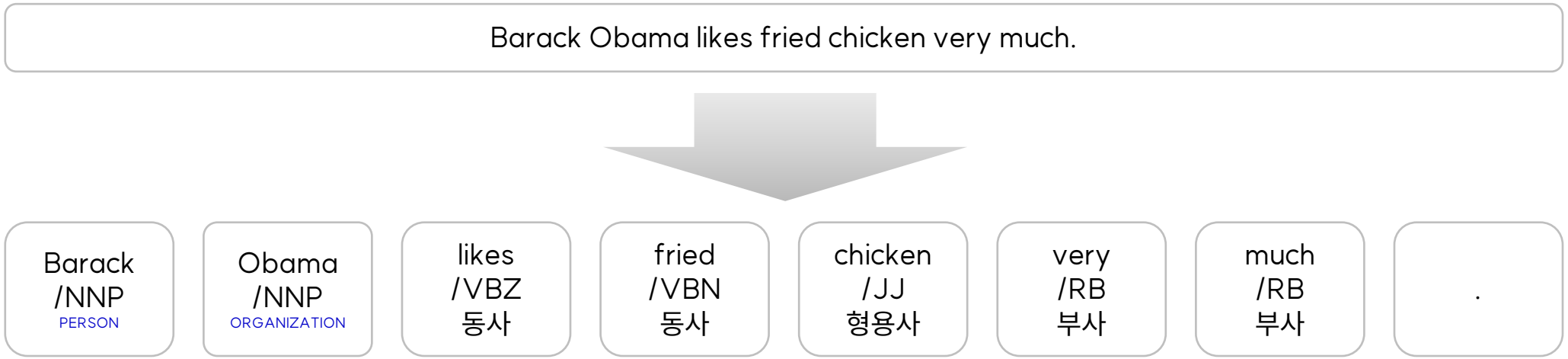
- 각 토큰에 품사 정보를 추가
- 분석시에 불필요한 품사를 제거하거나 (ex. 조사, 접속사 등) 필요한 품사를 필터링 하기 위해 사용



http://kkma.snu.ac.kr/documents/index.jsp?doc=postag

개체명 인식 (NER, Named Entity Recognition)

- 사람, 조직, 지역, 날짜, 숫자 등 개체 유형을 식별
- 텍스트가 무엇과 관련되어 있는지 구분하기위해 사용
- 검색 엔진 색인에 활용



※ 색인 : 검색을 빠르게 하기 위해 데이터를 일정한 순서로 나열한 목록.
즉, 특정 데이터를 빠르게 찾기 위해 일련의 순서를 유지한 상태로 저장하는 것을 인덱싱(indexing)이라고함

- chunking : 자연어 처리 기법중의 하나로 같은 의미를 한 덩어리로 인지하는 것

원형 복원

- 각 토큰의 원형 복원을 함으로써 토큰을 표준화, 불필요한 데이터 중복 방지
(= 단어의 수를 줄일 수 있어 연산의 효율성을 높임)

어간 추출 (Stemming)

- 품사를 무시하고 **규칙에 기반하여 어간을 추출**
- 규칙 : https://tartarus.org/martin/PorterStemmer/def.txt

원문	Stemming
running	run
beautiful	beauti
believes	believ
using	use
conversation	convers
organization	organ
studies	studi

표제어 추출 (Lemmatization)

- **품사정보를 유지하여** 표제어 추출 **(사전 기반)**

원문	Lemmatization
running	running
beautiful	beautiful
believes	belief
using	using
conversation	conversation
organization	organization
studies	study

Stopwords

불용어 처리

불용어 처리 (Stopwords)

- 불필요한 토큰을 제거하는 작업
- 분석 시 불필요한 품사를 제거하기도 함

- 문장을 구성할 때 자주 사용하지만 자주 사용하는 만큼 큰 의미를 가지지 않는 단어를 제거하는 과정 (the, a, an)