Text analysis/Mining

텍스트 분석

류영표 강사
youngpyoryu@dongguk.edu

강사 소개

류영표
Youngpyo Ryu

동국대학교 수학과/응용수학 석사수료

現 Upstage AI X 네이버 부스트 캠프 1기,2기 멘토

現 Innovation on Quantum & CT(IQCT) 이사

前 한국파스퇴르연구소 Image Mining 인턴(Deep learning)

前 ㈜셈웨어(수학컨텐츠, 데이터 분석 개발 및 연구인턴)

강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- (재)윌튼블록체인 6일 과정 (파이썬기초, 크롤링,머신러닝)
- ㈜ 모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 공공데이터 청년 인턴 / SW공개개발자대회 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의
- 메가 IT 아카데미(파이썬, 빅데이터 강사)
- 이젠 종로 아카데미(파이썬, ADSP 강사)
- 최적화된 도구(R/파이썬)을 활용한 애널리스트 양성과정(국비과정) 강사
- 한화, 하나금융사 교육
- 인공지능 신뢰성 확보를 위한 실무 전문가 자문위원

주요 프로젝트 및 기타사항

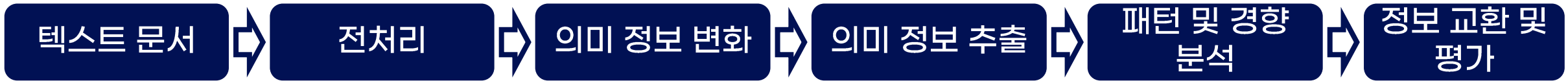
- 제1회 인공지능(AI)기반 데이터사이언티스트 전문가 양성과정 최우수상 수상(Q&A 챗봇)
- 인공지능(AI)기반 데이터사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

INDEX

1. Text Analysis

Text Analysis

- 대규모 텍스트 형태의 비정형 데이터로부터 유용한 정보를 추출 하는 것.
- 방대한 텍스트 뭉치(Corpus)에서 의미 있는 정보를 추출해 내고, 다른 정보와의 연계성을 파악하여, 텍스트가 가진 카테고리를 찾아내는 등 단순한 정보 검색 그 이상의 결과를 얻어 낼 수 있다.



Text Mining

- 의미 있는 패턴을 찾아내거나 통찰력을 얻어내는 방법
- 구조화 과정이 필요
- 비구조화 텍스트에서 구조화된 데이터로 변환 프로세스
- 텍스트 데이터 → 문서/문단/문장/단어
- 텍스트 데이터는 '구조화 시키는 과정 ' 이 필수적

Text Mining

	구조화된 데이터	비구조화된 데이터
내부	고객 (이름, 휴대폰, 이메일 등) 구매 (시간, 매장, 상품, 상품개수 등) 결제 (결제방식, 이용카드, 제휴할인 등) 온라인 (아이디, 로그인) 쿠폰 (발행, 사용, 보유) 멤버십 (등급, 포인트, 적립, 사용) 이벤트 (조회, 참여, 응모, 당첨) 평점 (상품, 평점)	검색 로그 (조회, 클릭, 장바구니, 관심상품 등) 위치 텍스트 (후기, VOC 등) 이미지 (후기 등) 신호 (카메라 센서, 와이파이, 사물인터넷)
외부	공공 (인구, 가구, 상권, 건설, 교통, 수출입 등 통계청 데이터) 날씨 상권, 부동산, 교통 등 공공 데이터	소셜 미디어 (블로그, 카페, 유튜브, 인스타그램) 포털 (검색어, 연관 검색어) 온라인몰 (상품정보, 가격 등)

출처 : https://brunch.co.kr/@parkkyunga/51

Text Mining VS Text Analysis

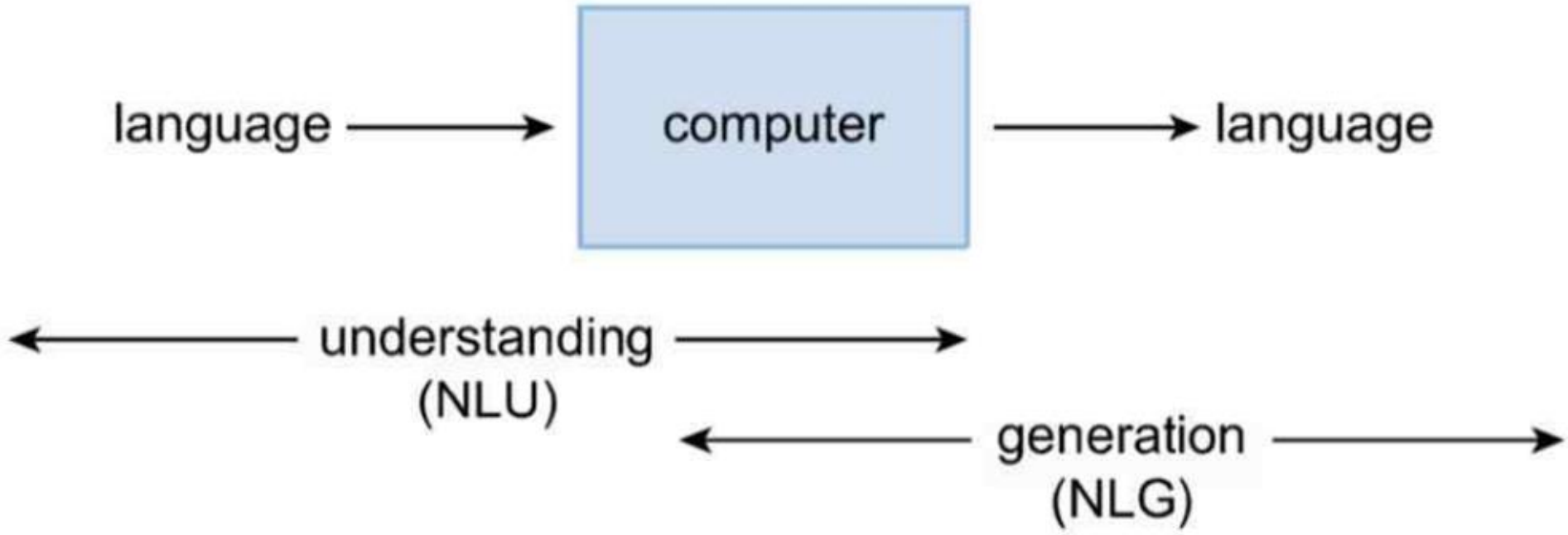
- 공통점: 머신 러닝이나 통계학, 언어학을 활용하여 비정형 데이터 내의 텍스트 분석을 통해 구조화된 데이터 형식으로 변환

- Text Mining: 데이터 내의 패턴과 추세를 찾아냄.

- Text Analysis: 데이터 에 대한 통찰력을 제공,
  시각화 기술을 활용한 이해도 향상.

자연어 처리(NLP: Natural Language Processing)

- 사람들 사이에서 이루어지는 대화는 일반적으로 자연어가 사용되지만, 컴퓨터는 그렇지 않기 때문에 정제작업이 필요
- NLP 분야는 컴퓨터가 문장어를 이해하거나 생성할 수 있도록 하는 학문

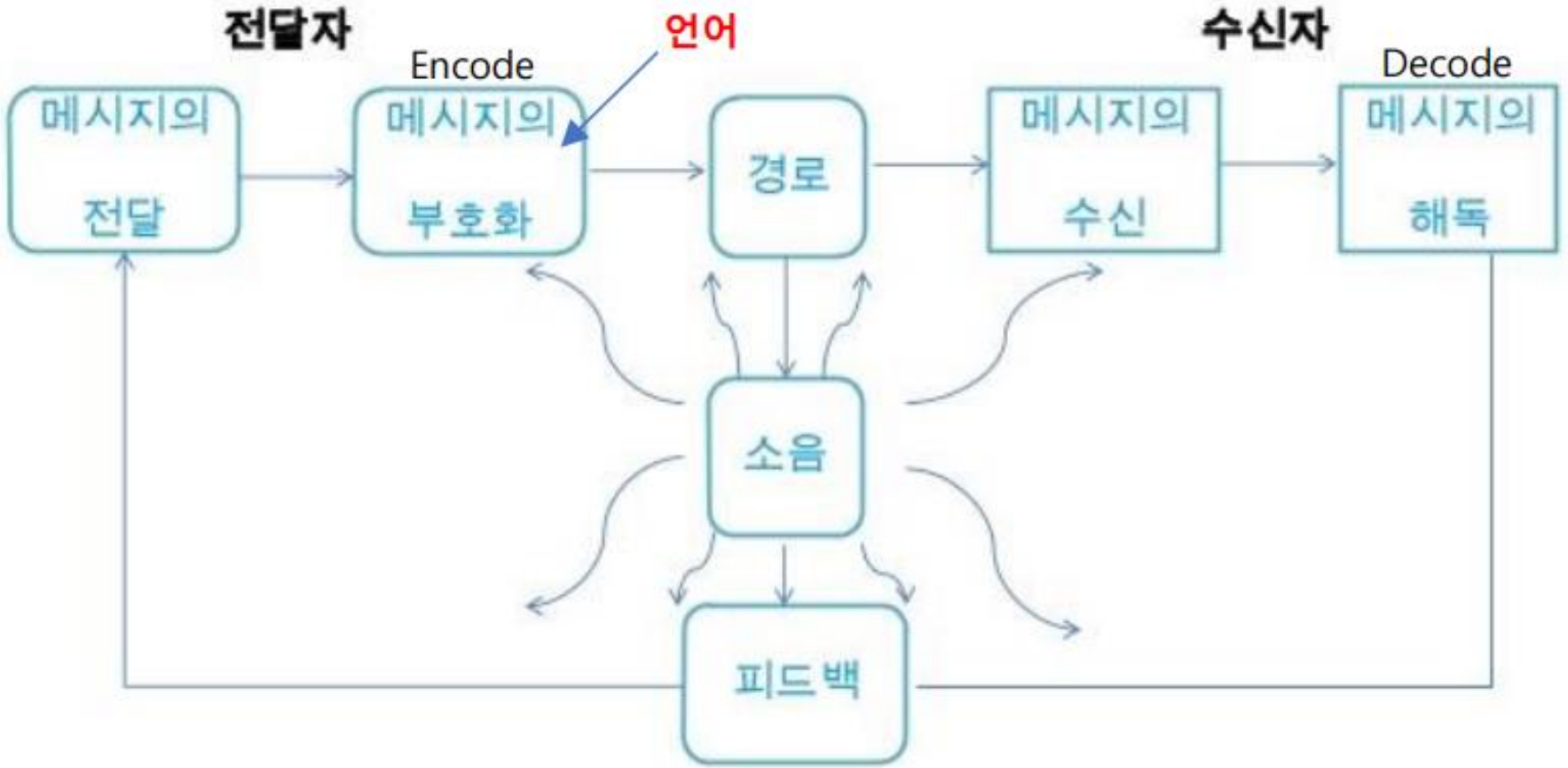


출처 : https://blog.kakaocdn.net/dn/P2jwU/btqQsucF3f4/e4JN9e8als0kPkmnzkB6Ek/img.png

자연어 처리 1)언어의 정의

언어[1] 言語

명사 생각, 느낌 따위를 나타내거나 전달하는 데에 쓰는 음성, 문자 따위의 수단. 또는 그 음성이나 문자 따위의 사회 관습적인 체계.



자연어 처리 2)자연 언어란?

자연 언어 自然言語

언어 일반 사회에서 자연히 발생하여 쓰이는 언어.

자연언어 : 한국어, 영어, 일본어

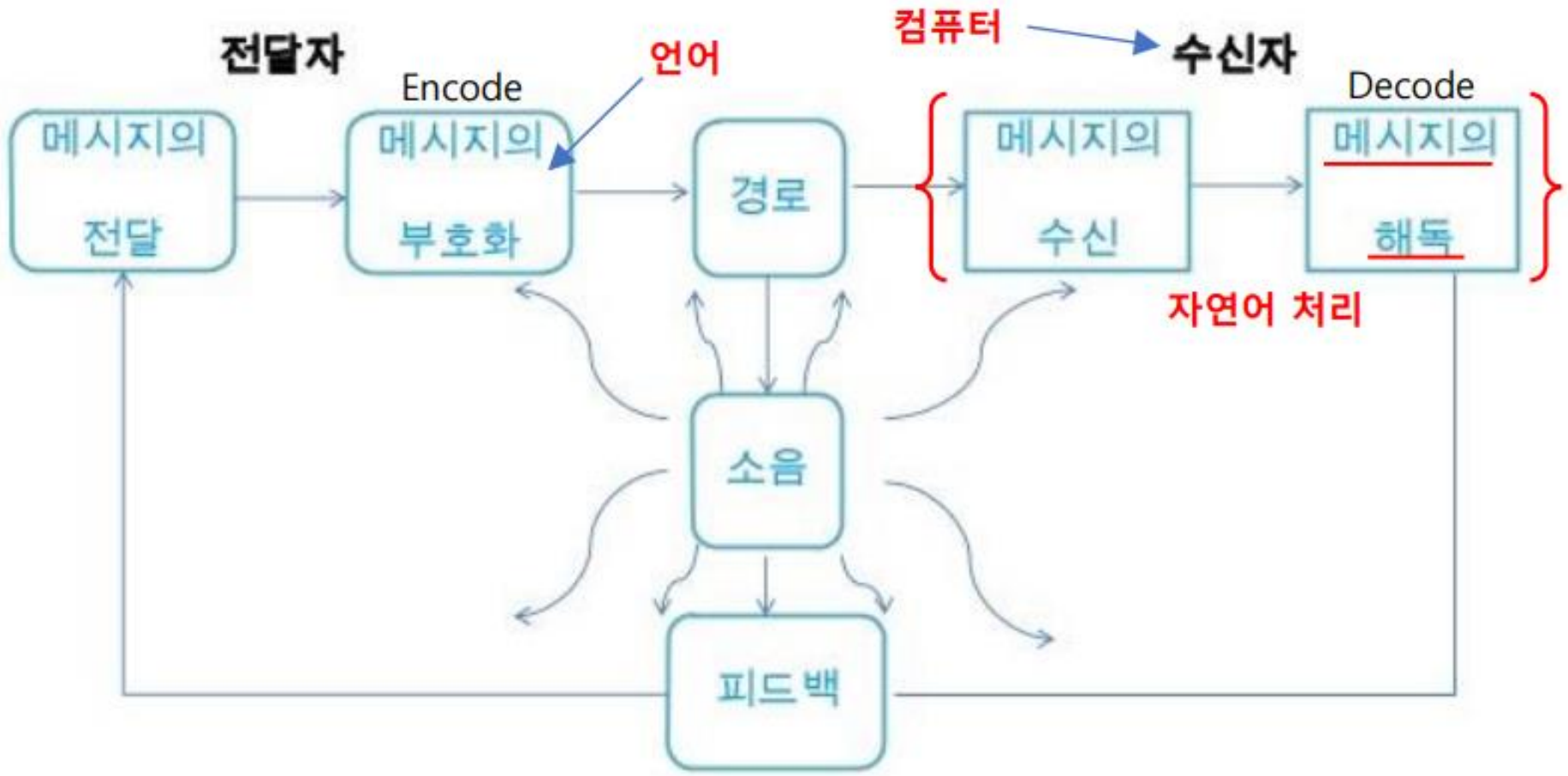
인공언어 : 프로그래밍 언어, 에스페란토어

* Chisung Song, 시나브로 배우는 자연어처리 바벨피쉬 송치성
https://www.slideshare.net/shuraba1/ss-56479835

자연어 처리 3)자연어 처리

자연어 처리 自然語處理

정보·통신 컴퓨터를 이용하여 인간 언어의 이해, 생성 및 분석을 다루는 인공 지능 기술.



자연어 처리 4)자연어 처리 접근법

- 자연어 처리란, '자연어를 컴퓨터가 해독하고 그 의미를 이해하는 기술'

Symbolic approach

- 규칙/지식 기반 접근법

```
@Override
boolean isAnswerableQuestion(String messagePattern) {
  boolean isAnswerable = false;

  if (messagePattern.matches("ChannelNm(NOW)?(PROGRAM)?WHAT")) {
    // 국회TV에서 지금 뭐해?
    isAnswerable = true;
  } else if (messagePattern.matches("ChannelNm(NOW)?WHATPROGRAM")) {
    // 국회TV에서 지금 무슨 방송해?
    isAnswerable = true;
  }  else if (messagePattern.matches("ChannelNm(NOW)?PROGRAMHOW")) {
    // 국회TV에서 지금 무슨 방송해?
    isAnswerable = true;
  }
  return isAnswerable;
}
```

100 원	100 (Number) + 원(G_ExchangeRateKRW : KRW=원)
100 달러	100(Number) + 달러(G_ExchangeRateKRW : USD=달러), S_UNIT_USD_money
100 m	100(Number) + m (S_UNIT_m_length)
100 미터	100(Number) + m (S_UNIT_m_length)

Statistical approach

- 확률/통계 기반 접근법
- TF-IDF를 이용한 키워드 추출
  - TF (Term frequency): 단어가 문서에 등장한 개수
    → TF가 높을수록 중요한 단어
  - DF (Document frequency): 해당 단어가 등장한 문서의 개수
    → DF가 높을수록 중요하지 않은 단어

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

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

* sujin oh, 실생활에서 접하는 빅데이터 알고리즘
https://www.slideshare.net/osujin121/ss-44186451?from_action=save

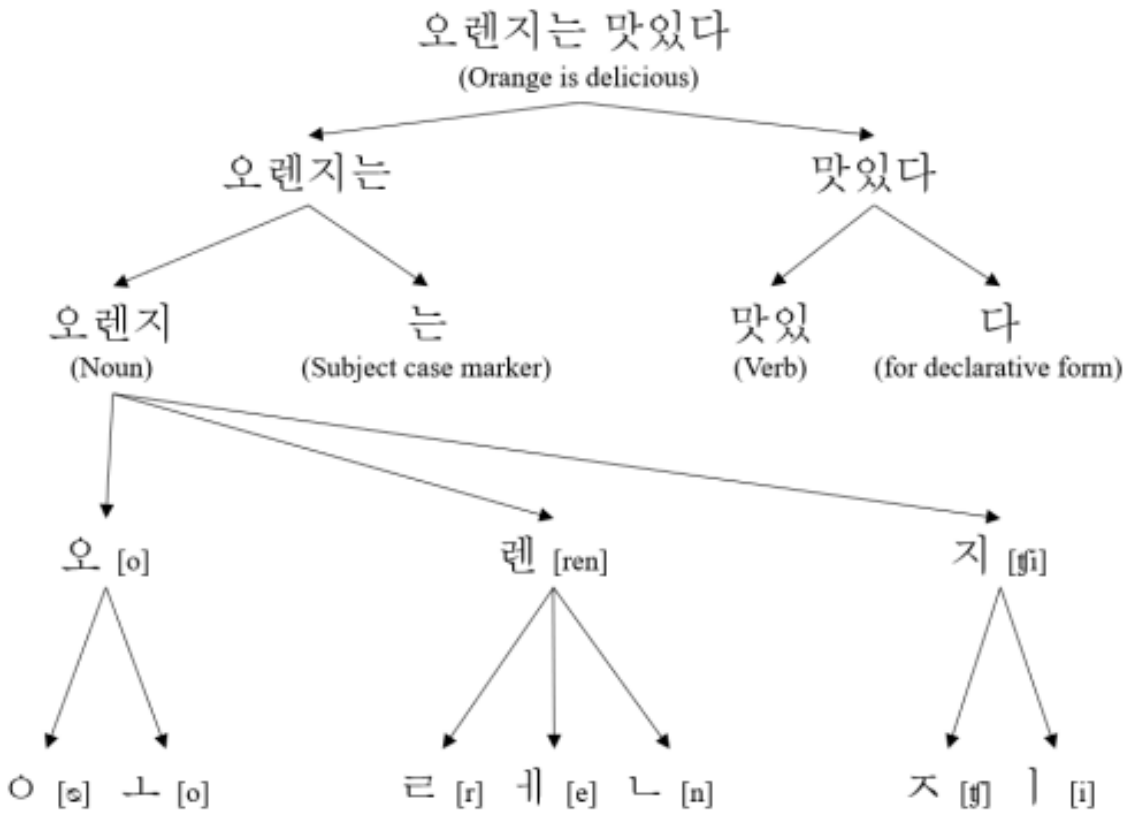
자연어 처리 5)전처리

- 전처리
  - 개행문자 제거
  - 특수문자 제거
  - 공백 제거
  - 중복 표현 제어 (ㅋㅋㅋㅋㅋ, ㅠㅠㅠㅠ, ...)
  - 이메일, 링크 제거
  - 제목 제거
  - 불용어 (의미가 없는 용어) 제거
  - 조사 제거
  - 띄어쓰기, 문장분리 보정
  - 사전 구축
- Tokenizing
- Lexical analysis
- Syntactic analysis
- Semantic analysis

의심한그득+. + 앞으로는 사람들 많을 퇴근시간 말구
미리 가서 **사와야겠뜸**ꉂ('ù')ﾉ
맥주로 터진 입, 청포도와 **고구뇽스틱**으로 달래**봅니당**
막상 먹다보니 양이 부족하진 않았는데 괜히 ...
먹을 때 많이 먹자며 배 터지기 직전까지 **냠냐미:D**
식후땡 아슈크림ﾉ (>__ < ‘)
브라우니쿠키 먹고 출근하세요 ♡
gs편의점에서 파는 **진 - 한** 브라우니 쿠키인데 **JMTgr**

자연어 처리 6)한국어의 전처리

- 전처리
- Tokenizing
  - 자연어를 어떤 단위로 살펴볼 것인가
  - 어절 tokenizing
  - 형태소 tokenizing
  - *n*-gram tokenizing
  - WordPiece tokenizing
- Lexical analysis
  - 어휘 분석
  - 형태소 분석
  - 개체명 인식
  - 상호 참조
- Syntactic analysis
  - 구문 분석
- Semantic analysis
  - 의미 분석



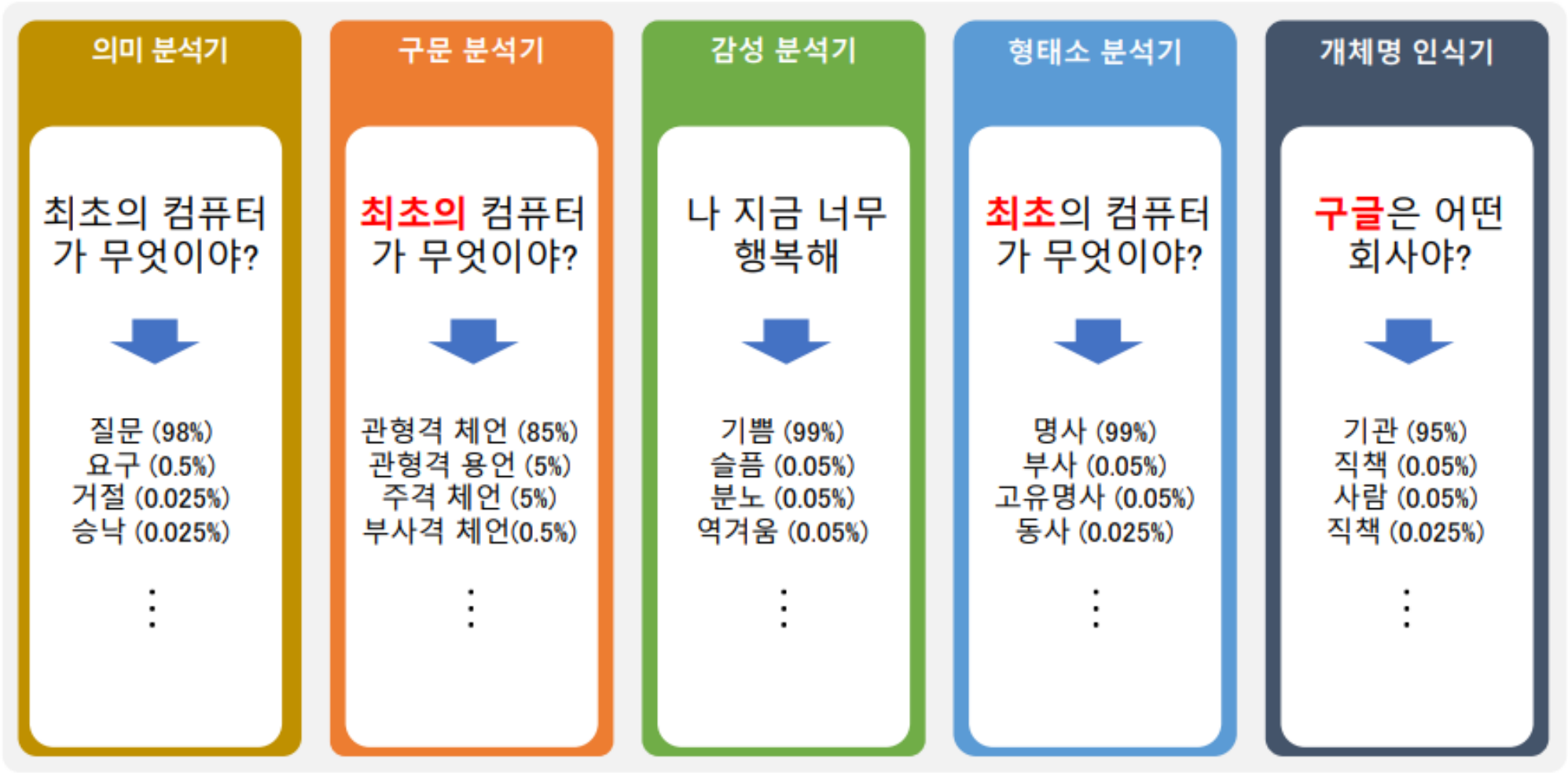
자연어 처리 7)활용분야

- 문서 분류
- 문법, 오타 교정
- 정보 추출
- 음성 인식결과 보정
- 음성 합성 텍스트 보정
- 정보 검색
- 요약문 생성
- 기계 번역
- 질의 응답
- 기계 독해
- 챗봇
- 형태소 분석
- 개체명 분석
- 구문 분석
- 감성 분석
- 관계 추출
- 의도 파악

자연어 처리 8)다양한 Applications

자연어 처리 8)다양한 Applications

- 형태소 분석, 문서 분류, 개체명 인식 등, 대부분의 자연어 처리 문제는 '분류'의 문제



자연어 처리 8)다양한 Applications

상당부분 해결

Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

성과를 내고 있음

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my *mouse*.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

여전히 어려운 문제

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

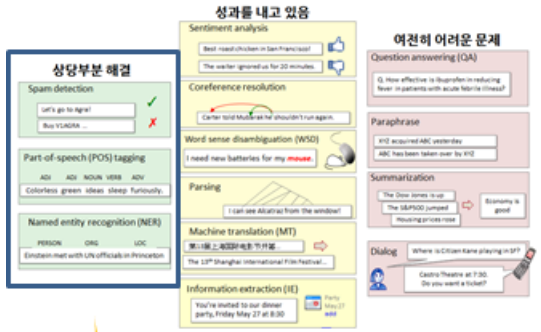
Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30.
Do you want a ticket?

출처 : https://statkclee.github.io/ds-authoring/ds_kpf.html#1

자연어 처리 9)작업 흐름



상당부분 해결

Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

단어주머니
(Bag of Words)

말뭉치

TDM, DTM

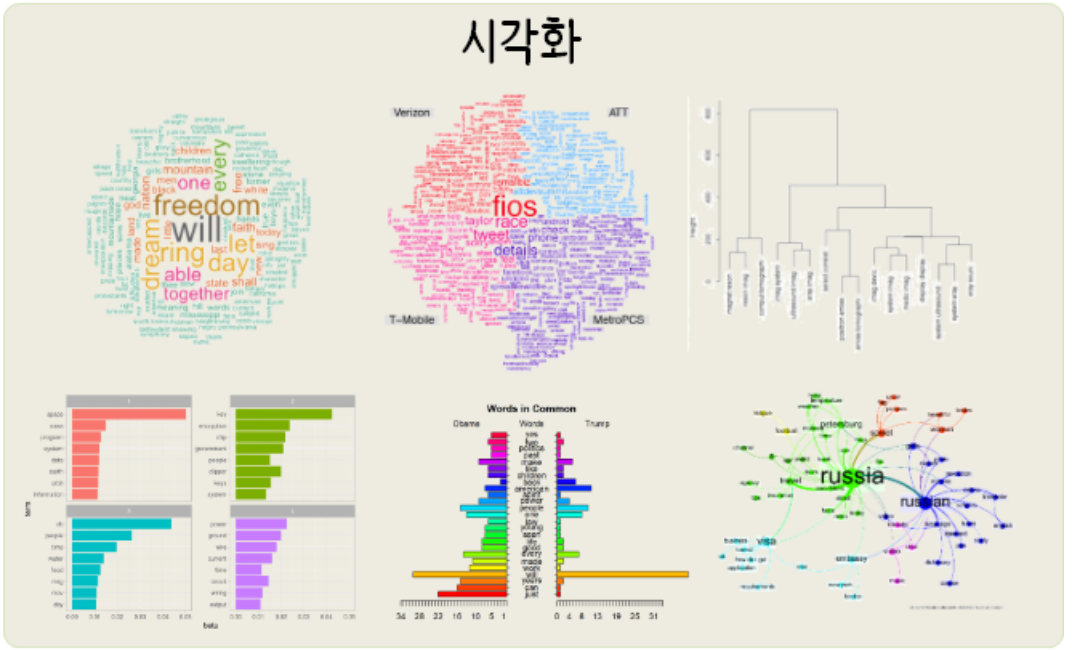
VCorpus

PlainTextDocument
메타정보 (meta) | 텍스트 (content)
...
PlainTextDocument
메타정보 (meta) | 텍스트 (content)

데이터프레임
(dataframe)

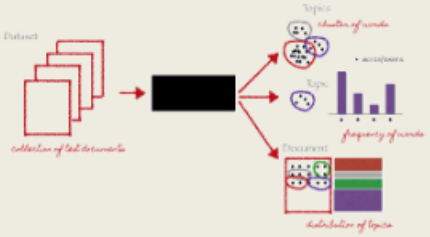
TF-IDF
(Term Frequency - Inverse Document Frequency)

시각화



NLP 주제

감성분석



토픽 모형

출처 : https://statkclee.github.io/ds-authoring/ds_kpf.html#1

2. 데이터 전처리

토큰화(Tokenization)

- 말뭉치 또는 코퍼스(Corpus)는 자연언어 연구를 위해 특정한 목적을 가지고 언어의 표본을 추출한 집합.

- 텍스트 토큰화란 말뭉치로부터 토큰을 분리하는 작업.

  ex) 'There is an apple' = Corpus

  → 'There', 'is', 'an', 'apple' 으로 나뉨.

- 문장토큰화와 단어 토큰화로 나뉨.

불용어(Stopword)

- 불용어(Stopword)는 분석에 큰 의미가 없는 단어를 지칭.

  ex) the,a,an,is,I,my 등과 같이 문장을 구성하는 필수요소이지만 문맥적으로 큰 의미가 없는 단어가 이에 속함.

- 텍스트에 빈번하게 나타나기 때문에 중요한 단어로 인지될 수 있음.

- 실질적으로 중요한 단어가 아니므로 사전에 제거를 해줘야 함.

정제 및 정규화

- **정제(Clearning) :** 갖고 있는 코퍼스로부터 노이즈 데이터를 제거
- **정규화(Normalization) :** 표현 방법이 다른 단어들을 통합시켜서 같은 단어로 만들어줌.

ex) 1. 규칙에 기반한 표기가 다른 단어들의 통합

2. 대, 소문자 통합

3. 불필요한 단어의 제거

4. 정규표현식을 이용.

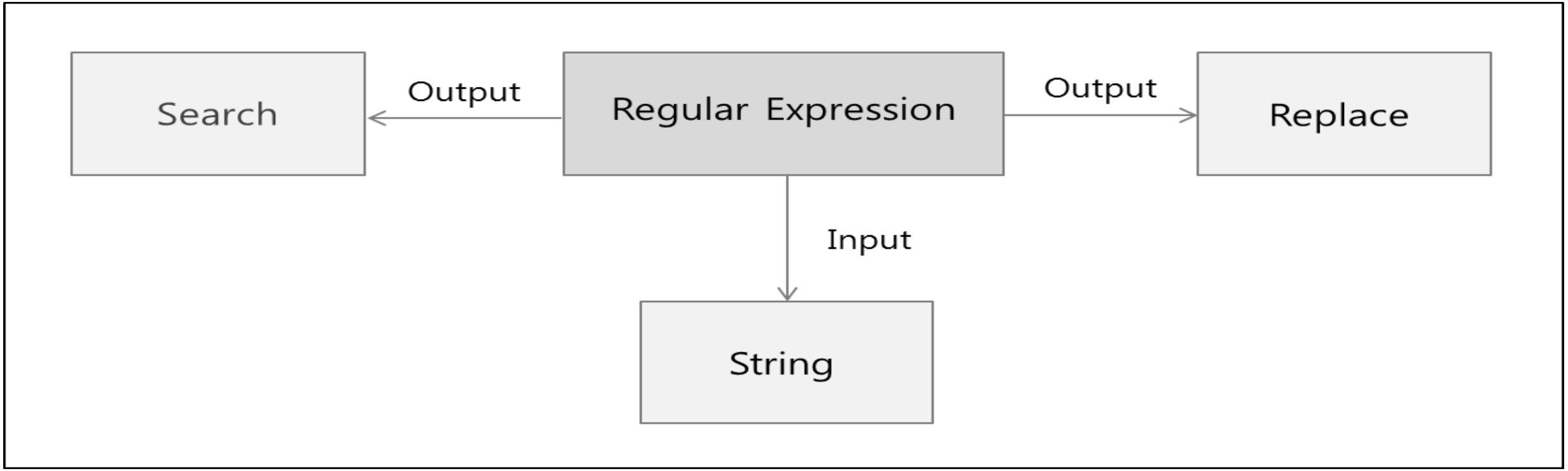
어간추출(Stemming)과 표제어(Lemmatization) 추출

- **어간추출(Stemming) :** 단어로부터 어간(Stem)을 추출하는 작업.

- **표제어 추출(lemmatization) :** 어간추출보다 더 정확히 어근 단어를 찾아줌. 품사와 같은 문법적인 요소와 더 의미적인 부분을 감안하기 때문에, 어간 추출보다 시간이 더 오래걸림.

정규표현식(Regular Expression)

- 특정한 규칙을 가진 문자열의 집합을 표현하는데 사용하는 형식 언어
- 문자열의 검색과 치환을 위한 용도로 쓰임



출처 : https://www.nextree.co.kr/p4327/

정수 인코딩(Integer Encoding)

- 컴퓨터는 숫자로만 처리가 가능하기 때문에, 자연어를 숫자로 바꾸는 방법 등을 고려해야 됨.
- 각 단어를 고유한 정수에 매핑(Mapping)시키는 전처리 작업이 필요 할 때가 있음.
- 방법: 1) Dictionary 사용
  2) Counter 함수 사용
  3) NLTK의 FreqDist 사용

품사 태깅(Pos Tagging)

- 문장을 형태소 단위로 분리 한 후, 해당 형태소의 품사를 태깅하는 것

- 품사는 명사, 대명사, 수사, 조사, 동사, 형용사, 관형사, 부사, 감탄사와 같이 공통된 성징을 지닌 낱말끼리 모아 놓은 낱말의 갈래를 의미[네이버 지식백과]

- 품사 태깅을 적용한 후, 명사만 추출하거나, 주요 품사만 추출해 데이터로 사용할 수 있음.

3. 언어 모델

언어 모델(Language Model)

- **주어진 단어들로부터 다음에 등장할 단어의 확률을 예측하는 모델**

- **크게 통계를 이용한 방법과 인공신경망을 이용한 방법으로 구분 가능**

- **활용:** a) 기계 번역(Machine Translation)
  b) 오타 교정(Spelling Correction)
  c) 음성 인식(Speech Recognition)

언어 모델 1) 문장의 확률 예측

- **언어 모델은 문장의 확률 또는 단어 등장 확률을 예측하는 일**

$$P(W) = P(w_1, w_2, w_3, w_4, w_5, \ldots, w_n) = \prod_{n=1}^{n} P(w_n)$$

언어 모델 1) 문장의 확률 예측

- **언어 모델은 문장의 확률 또는 단어 등장 확률을 예측하는 일**

$$P(W) = P(w_1, w_2, w_3, w_4, w_5, \ldots w_n) = \prod_{n=1}^{n} P(w_n | w_1, \ldots, w_{n-1})$$

N-gram 언어 모델(N-gram Language Model)

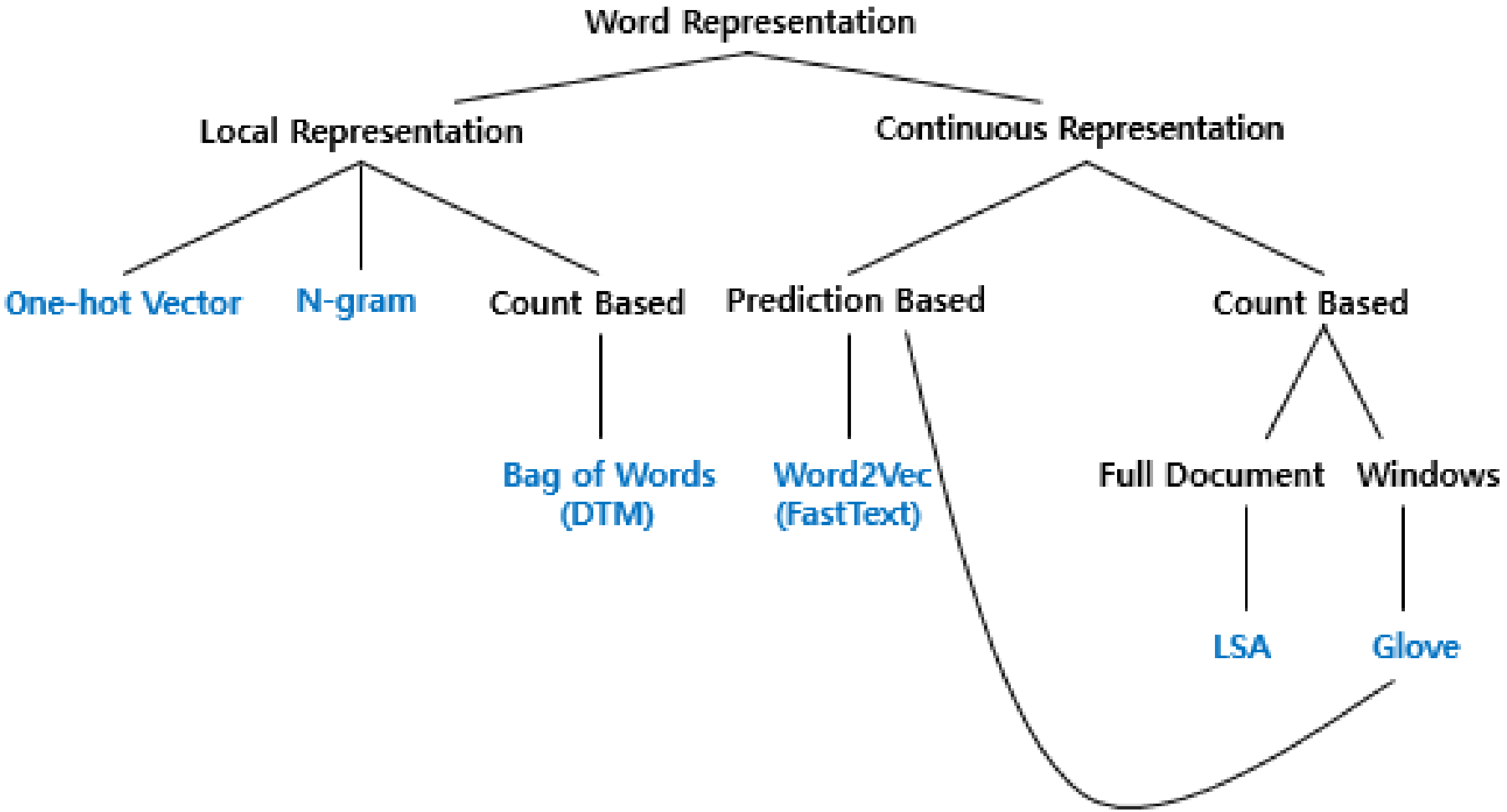
- 카운트 기반한 통계적 접근
- 주변에 n개 단어를 뭉쳐서 보는 것. 뭉쳐진 n개의 단어들을 gram이라고 함.
- 단어 개수에 따라 부르는 명칭이 다른데 2개의 단어를 묶어서 사용하면 bi-gram, 3개면 tri-gram이라고 부름.
- N-gram의 한계

  1) 희소 문제(Sparsity Problem)

  2) n을 선택하는 것은 trade_off 문제

4. 단어 표현 방법

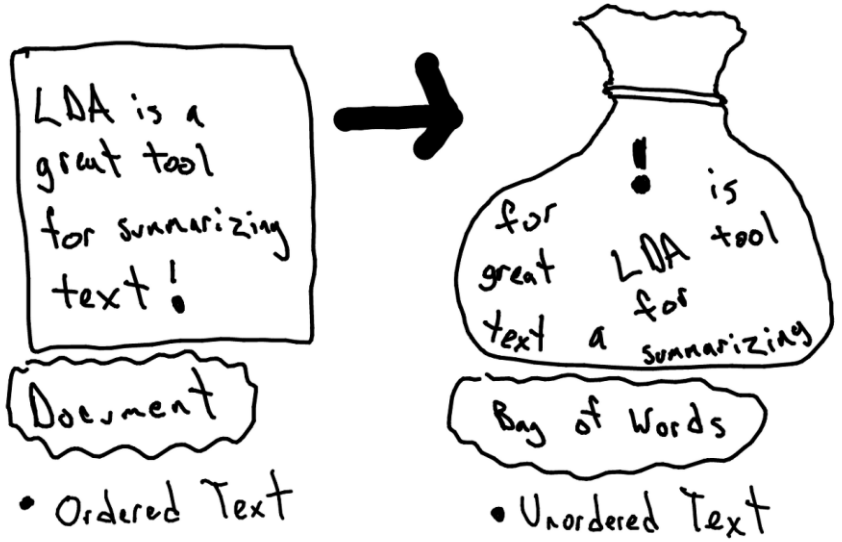
단어 표현의 카테고리화



출처 : https://wikidocs.net/31767

BOW(Bag of words)

- 단어들의 문맥이나 순서를 무시하고, 단어들에 대해 빈도 값(frequenct)을 부여해 피쳐 값을 만드는 모델
- 문서 내 모든 단어를 한꺼번에 가방(Bag)안에 넣은 뒤에 흔들어서 섞는다는 의미로 Bag of words(BOW) 모델이라고 함.



출처 : https://ldabook.com/word-representations.html

문서 단어 행렬(Document-Term Matrix,DTM)

- **다수의 문서에서 등장하는 각 단어들의 빈도를 행렬로 표현 한 것.**

Ex) 4개의 문서가 있다고 하자

문서1 : 먹고 싶은 사과

문서2: 먹고 싶은 바나나

문서3: 길고 노란 바나나 바나나

문서4: 저는 과일이 좋아요.

-	과일이	길고	노란	먹고	바나나	사과	싶은	저는	좋아요
문서1	0	0	0	1	0	1	1	0	0
문서2	0	0	0	1	1	0	1	0	0
문서3	0	1	1	0	2	0	0	0	0
문서4	1	0	0	0	0	0	0	1	1

TF-IDF(Term Frequency-Inverse Document Frequency)

- 카운트 기반 벡터화는 카운트 값이 높을수록 중요한 단어로 인식함.

- 단어의 빈도만 고려한다면 모든 문서에서 자주 쓰일 수 밖에 없는 단어들이 중요하다고 인식될 수 있음.

TF-IDF

- **IF-IDF는 개별 문서에서 자주 등장하는 단어에 높은 가중치를 주되, 모든 문서에서 전반적으로 자주 등장하는 단어에 대해서는 패널티를 주는 방식으로 값 부여.**

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

출처 : woongheelee.com

Thank you.

텍스트 분석 / 류영표 강사
youngpyoryu@dongguk.edu