AI기반 데이터 분석 및 AI Agent 개발 과정

『1과목 : 』
AI기반 데이터 분석

2025.09.22-10.02(9일, 62시간)

Prepared by DaeKyeong K

Ph.D.

삼성전자 구미사업부 G.제조팀

HS Global
HARDWARE SOFTWARE

목차

1. 생성형 AI와 데이터 분석
2. 조사 및 데이터 수집 방법
3. 데이터 전처리
4. 데이터 분석
5. 통계적 가설 검정 및 분석
6. 데이터 준비(Data Preparation)
7. 상관관계 및 연관성 이해
8. 인과 관계 및 예측 분석 이해
9. 머신러닝 기반 데이터 분석-지도
10. 머신러닝 기반 데이터 분석-비지도
11. 기타 데이터 마이닝
12. 텍스트 데이터 분석 텍스트 마이닝 이해

『1-12』
텍스트 데이터 분석 텍스트 마이닝

텍스트 데이터 분석 이해

형태소 분석
검색 트렌드 분석
텍스트 분류 방법
감성 분석
연관어 분석

학습 개요

학습목표

- 이 워크샵에서는 다양한 텍스트 형태의 원천 자료로부터 고품질의 정보를 도출하기 위해 입력된 텍스트를 처리하고, 구조화하여 패턴을 도출한 후 결과를 평가 및 해석하여 현실 업무에 적용할 수 있다.

눈높이 체크

- 텍스트 분석, 텍스트 변환, 형태소, 말뭉치, 단어-문서 관계, 단어사전, 텍스트 분류을 알고 계신가요?

텍스트 데이터 분석 이해

텍스트 마이닝의 정의

- 텍스트 마이닝(Text Mining)은 구조화되지 않은 텍스트 데이터로부터 유용한 정보와 지식을 추출하는 과정입니다. 이는 자연어 처리(NLP), 기계학습, 통계학적 방법론을 활용하여 텍스트에 숨겨진 패턴을 발견하는 기술입니다.
- 텍스트 마이닝(Text Mining) 또는 텍스트 분석은 비정형 텍스트 데이터로부터 유용한 정보를 추출하는 기술입니다. ICT 기술의 발달에 따라 정형·비정형의 데이터 형태로 잠재적 활용 가치가 높은 정보들이 급증하고 있으며, 특히 SNS의 활성화로 텍스트 데이터 분석의 중요성이 높아지고 있습니다.

텍스트 데이터 분석 이해

텍스트 데이터의 특징

- **구조화되지 않은 데이터**
  - 일정한 형식이나 스키마가 없음
  - 자연어로 작성된 문서, 소셜미디어 게시글, 리뷰 등

- **고차원성**
  - 단어 수만큼 차원이 증가
  - 희소성(Sparsity) 문제 발생

- **언어적 복잡성**
  - 동의어, 반의어, 중의성
  - 문맥에 따른 의미 변화
  - 신조어와 은어의 지속적 생성

텍스트 분석 절차

- 텍스트 분석을 위한 전체적인 절차는 다음 6단계로 구성됩니다.
- 요구사항 분석
  - 텍스트 분석의 첫 단계로 분석 대상에 대한 사용자의 요구사항을 이해하고 문서화하는 과정입니다. 사용자의 요구를 정확하게 분석하여 텍스트 분석 목적에 적합한 다양한 해결 방법을 검토합니다.

텍스트 분석 절차

- **텍스트 수집**
  - **업무 특성 및 목적에 적합한 데이터를 수집하는 과정입니다. 주요 수집 기술은 다음과 같다**

구분	특징	비고
Crawling	SNS, 뉴스, 웹 정보 등 웹 문서·정보 수집	웹 문서 수집
Scraping	하나의 웹사이트에 대한 정보 수집	웹 문서 수집
FTP	TCP/IP 프로토콜을 활용한 파일 송수신	FILE 수집
오픈 API	개방된 API를 통한 실시간 데이터 수집	실시간 데이터 수집
RSS	XML 기반 콘텐츠 배급 프로토콜	콘텐츠 수집

텍스트 데이터 분석 이해

텍스트 분석 절차

- **텍스트 저장 및 전처리**
  - 텍스트 분석을 위한 데이터 처리 기술 및 저장 방식을 선정하는 과정입니다. 불필요한 항목(불용어 등)을 제거하고 대상 텍스트의 품질을 향상시키는 전·후처리 기법을 적용합니다.

- **텍스트 분석**
  - 형태소 분석, 불용어 처리, 키워드 추출, 단어와 문서 관계 표현 등의 전처리 과정을 수행한 후 다음과 같은 분석 방법을 적용합니다:

  - 텍스트 분류: 텍스트를 미리 정의된 카테고리로 분류
  - 텍스트 군집: 내용이나 형태가 유사한 텍스트들을 그룹화
  - 텍스트 요약: 주요 의미를 유지하면서 텍스트 길이를 효과적으로 축약

텍스트 데이터 분석 이해

텍스트 분석 절차

- **텍스트 분석 서비스 제공**
  - 시각화를 통해 텍스트 분석 결과를 사용자가 쉽게 활용할 수 있도록 제공합니다. 태그 클라우드, 지도, 차트 등을 활용하여 결과를 표현합니다.

- **산출물 관리 및 공유**
  - 텍스트 분석 결과를 현업 구성원에게 공유하고 문서화 및 버전 관리를 수행합니다. 개인정보 처리 및 보안 관리를 통해 안전한 활용을 도모합니다.

텍스트 데이터 분석 이해

텍스트 분석 절차

- **텍스트 분석 서비스 제공**
  - 시각화를 통해 텍스트 분석 결과를 사용자가 쉽게 활용할 수 있도록 제공합니다. 태그 클라우드, 지도, 차트 등을 활용하여 결과를 표현합니다.

- **산출물 관리 및 공유**
  - 텍스트 분석 결과를 현업 구성원에게 공유하고 문서화 및 버전 관리를 수행합니다. 개인정보 처리 및 보안 관리를 통해 안전한 활용을 도모합니다.

텍스트 데이터 분석 이해

텍스트 분석 방법론

- **텍스트 분류(Text Classification)**
  - 임의의 텍스트를 미리 정의된 카테고리로 분류하는 지도 학습 기법입니다. 분류기 학습을 통해 각 카테고리의 특성 정보를 정의하고, 입력 텍스트와의 유사도를 비교하여 적합한 분류를 선정합니다.
- **텍스트 군집(Text Clustering)**
  - 텍스트의 특성을 분석하여 내용이나 형태가 유사한 텍스트들을 동적으로 군집화하는 비지도 학습 기법입니다. 미리 정의된 카테고리 없이 데이터 간의 유사도에 근거하여 부분 집합으로 분할합니다.
- **텍스트 요약(Text Summarization)**
  - 대상 텍스트의 주요 의미를 유지하면서 길이를 효과적으로 줄이는 기술입니다:

  - 추출 요약: 문서에 존재하는 문장을 변경 없이 추출
  - 생성 요약: 문서의 내용을 압축 및 재구성하여 새로운 요약문을 생성

텍스트 데이터 분석 이해

활용 분야

- 마케팅: 소비자 반응 분석, 브랜드 모니터링
- 금융: 뉴스 기반 주가 예측, 리스크 분석
- 의료: 의료 문헌 분석, 증상 패턴 분석
- 정치: 여론 분석, 정책 반응 모니터링
- 교육: 학습자 피드백 분석, 교육 콘텐츠 개선

『1-12』
텍스트 데이터 분석 텍스트 마이닝

텍스트 데이터 분석 이해

형태소 분석

검색 트렌드 분석

텍스트 분류 방법

감성 분석

연관어 분석

학습 개요

학습목표

- 이 워크샵에서는 형태소에 대해 알 수 있습니다.

눈높이 체크

- 형태소 분석을 알고 계신가요?

1. 자연어처리란?

자연어처리란?

- 자연어: 사람들의 사회생활에서 자연스럽게 발생하여 쓰이는 언어
  - 컴퓨터에게 명령을 하기 위해 제약을 더하여 만든 프로그래밍 언어와 같은 인공언어와 대비
- 자연어 처리
  - 사람들이 사용하는 자연어를 **컴퓨터를 이용하여 이해하고 생성**하도록 하는 제반의 연구

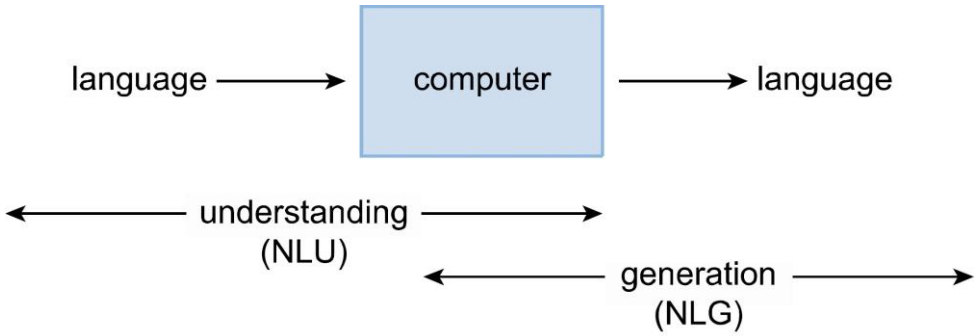

그림 1-1 전산학에서 바라본 자연어처리의 과정

- NLU(자연어 이해): 입력된 언어의 의미를 파악하여 의미 표현 형태로 변화시키는 과정
- NLG(자연어 생성): 주어진 의미를 표현하기 위하여 해당 의미를 나타내는 언어를 생성하는 과정

1. 자연어처리란?

자연어처리의 응용 분야

- 언어학적인 활용: 전산언어학에서 언어를 분석하는데 사용
- 전산학적인 활용
  - 기계번역
  - 음성인식
  - 개인비서 서비스
  - 날씨정보 요약
  - 인공지능 스피커

1. 자연어처리란?

자연어처리 연구의 패러다임 - 규칙 기반

- 언어의 문법적 규칙을 사전에 정의해두고 이에 기반하여 자연어를 처리

- 초창기의 자연어처리 연구에 많이 사용됨
  - 1954년: 러시아어-영어 번역기 (조지타운대학, IBM)
  - 1960년: SHRDLU ELIZA 대화형 시스템

- 규칙 기반으로 자연어를 처리하는 예시
  - 기계 번역: 문장을 형태소 등의 단위로 분해하고, 그 사이에서 감지된 규칙을 사용하여 번역
  - 명령 인식: 문장에서 목적어, 동사 등이 위치하는 규칙을 이용해 대상과 행동 등을 이해

- 분명한 한계점이 존재하여 더이상 사용되지 않음
  - 한국어처럼 어순이 정형화되어있지 않으면 분석에 한계가 존재
  - 규칙을 미리 지정하는 것의 부담이 큼
  - 규칙을 적용할 단위로 분해하는 작업의 정확도가 낮음

1. 자연어처리란?

자연어처리 연구의 패러다임 - 통계 기반

- 규칙 사전 정의를 통계적으로 처리
  - 언어에 어떤 규칙이 있다면 그 단어나 어구 사이에 통계적으로 유의미한 값이 도출된다는 가정

- 컴퓨터의 성능이 발전하여 대량의 데이터를 빠르게 처리할 수 있게 되면서 발전
  - 통계적인 분석을 위해서는 사전에 수집된 대량의 문장들(코퍼스)을 처리해야 함

- 조건부 확률이라는 수학적 개념이 가장 핵심적
  - 어떤 사건이 이미 일어난 것을 가정하고, 그 상황에서 다른 사건이 일어날 확률
  - 어떠한 단어가 등장할 확률을 앞뒤의 단어(들)를 기반으로 산출하는 것

- 통계적인 분석의 한계
  - 선형적인 분석이기 때문에 복잡한 규칙을 처리하기엔 어려움이 있음
  - 여전히 사람의 손이 많이 가는 통계 분석 자료 활용

1. 자연어처리란?

자연어처리 연구의 패러다임 - 딥러닝 기반

- 알고리즘
  - 일반적인 '알고리즘': 어떤 상황에 어떻게 어떤 값을 계산해야 하는지 사전에 지정된 연산 흐름
  - 어떤 데이터가 들어오는지 예측이 가능하고, 프로그래머 역시 그 처리법을 명확히 알고 있어야 함

- 기계 학습
  - 입력으로 들어올 데이터를 대입시켜 알고리즘이 스스로 연산의 가중치를 학습하게 함
  - 프로그램을 작성 후 바로 사용하지 않고 학습시키는 과정이 필요
  - 학습된 가중치는 나중에 저장하였다가 다시 활용할 수 있음

- 신경망
  - 기계학습의 일종으로, 신경계의 구성 형태를 기반으로 만들어진 구조
  - 여러 입력을 가중치를 적용하여 합하고, 활성 함수에 통과시킨 후 전달
  - 입력과 출력을 제외한 층을 '은닉층' 이라고 함

1. 자연어처리란?

자연어처리 연구의 패러다임- 딥러닝 기반

- 딥러닝
  - 신경망 구조에서 은닉층 수를 매우 많이 늘린 것
  - 은닉층 수에 대한 구체적인 기준은 없으나, 연산의 흐름이나 각 가중치가 무엇을 의미하는지 개발자조차 알 수 없게 됨
  - 하지만 여러 복잡한 특징들을 처리할 수 있게 되어 각광받고 있음

- 딥러닝 기반 자연어 처리
  - 모델을 잘 구성해두는 것이 중요
  - 문장 전체를 고려하는 모델을 만들고 싶다면, 모든 단어에 걸쳐 적용되는 연결을 하나 만듦
  - 통계적 분석보다 고차원적인 분석을 할 수 있어 자연어 처리의 성능이 비약적으로 상승함

1. 자연어처리란?

딥러닝을 사용하는 자연어처리 연구 단계

- 딥러닝을 사용하는 자연어 처리의 연구 순서
  1. 어떠한 목적으로 자연어처리를 도입하는 것인지 결정
  2. 목적과 관련한 학습데이터 수집 또는 구축
  3. 학습데이터를 통해 학습시킬 모델 구조를 작성
  4. 준비한 학습데이터를 이용하여 모델을 학습
  5. 완성된 모델을 검증하고, 실전에 투입

- 성능 개선은 주로 2단계, 3단계에서 진행

1. 자연어처리란?

딥러닝을 사용하는 자연어처리 연구 단계

- 단어 임베딩
  - 자연어로 되어있는 문장을 컴퓨터가 받아들일 수 있도록 하는 문장의 전처리 과정 (모델의 일부)
  - 다양한 방법이 있으나, 단어간 연관성 등을 유지하는 벡터화 하는 방법이 많이 쓰임
  - 문법적으로만 사용되는 단어(조사, be동사 등)는 일반적으로 삭제
  - 사전에 임베딩 된 단어 사전을 사용하여 연구를 진행하는 경우가 많음

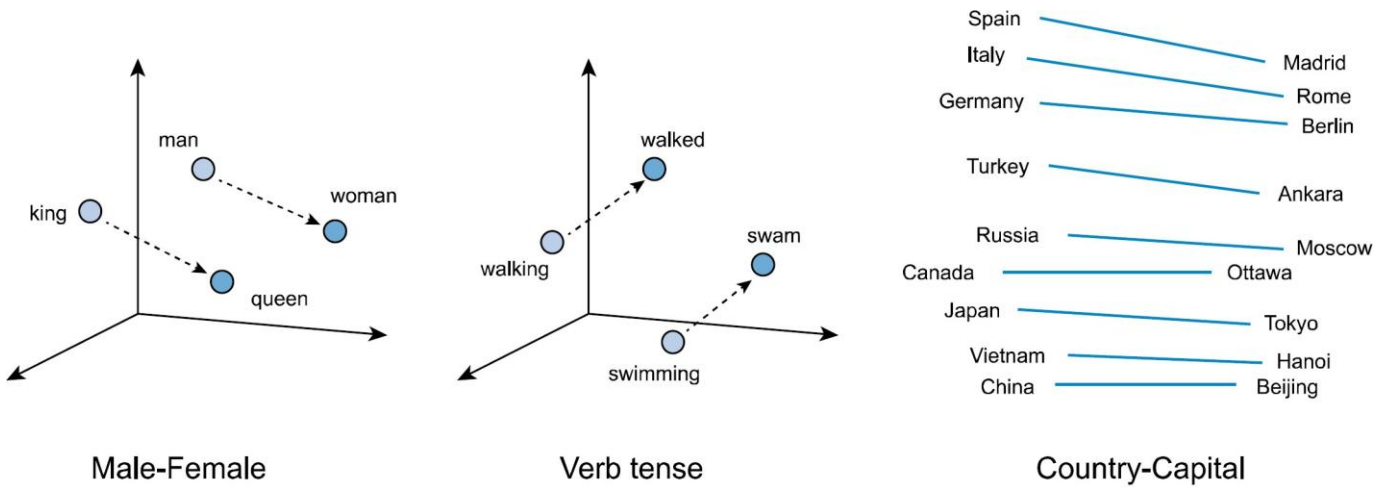

그림 1-6 영어 단어를 임베딩한 예시

1. 자연어처리란?

딥러닝을 사용하는 자연어처리 연구 단계

- 코퍼스
  - 매우 많은 수의 문장을 정제하여 모아둔 것
  - 통계/딥러닝 기반 자연어 처리에서 가장 핵심을 담당하는 자료
  - 연구 필요성에 따라서 문장에 문장 성분을 기입하거나 대응하는 번역문을 쌍을 짓는 등, 연구에 사용할 (모델이 학습해야 할) 정보를 같이 기입
- 모델
  - 딥러닝을 활용한 연구의 핵심
  - 연구의 목적에 맞추어 모델이 어떤 부분을 읽고, 어떤 형식으로 출력하는지를 결정
  - 감정 분석을 한다면 Classification, 번역을 한다면 Generation 등
  - 성능이 아주 좋은 모델은 출력단에서만 변화를 주어 다른 작업에 사용되기도 함

2. 언어학의 기본 원리

언어학 개요

언어학의 기본적인 원리

- 언어를 이루는 단위 : 음절(Syllable), 형태소(Morpheme), 어절, 품사(part-of-speech, POS)

- 언어의 구조 : 구구조(句構造, Phrase structure), 의존구조(Dependency structure)

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

음절(Syllable)

언어를 말하고 들을 때, 하나의 덩어리로 여겨지는 가장 작은 발화의 단위.

한국어에서 음절은 기본적으로 초성(Onset), 중성(Nucleus), 종성(Coda)으로 이루어져 있음.
초성은 음절에서 가장 처음에 오는 소리로 자음(Consonant, C), 중성은 가운데 소리로 모음(Vowel, V), 종성은 마지막 소리로 자음이 해당.

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

음절(Syllable)

- 한국어의 음절은 모음 단독으로 이루어 질 수도 있고 모음 앞, 뒤에 자음이 하나씩 붙어 다음과 같 이 V, C+V, V+C, C+V+C와 같은 형태로 구성됨.

자음(Consonant, C), 중성은 가운데 소리로 모음(Vowel, V), 종성은 마지막 소리로 자음이 해당

V	아, 오, 이, 에	(1)
V+C	약, 얼, 옹, 인	(2)
C+V	다, 리, 무, 서	(3)
C+V+C	말, 글, 성, 혼	(4)

그림 3-1 한국어 음절 구성의 종류

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

음절(Syllable)

- 음절은 말소리의 단위이기 때문에 엄밀히 말하면 (2)와 같이 소리 나는 대로 적었을 때의 한 글자를 말함.

'이 문장에서 음절은 몇 개일까?' (1)

/이 문장에서 음저른 멷 개일까?/ (2)

그림 3-2 음절의 개수

12개

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

형태소(Morpheme)
언어에서 의미를 가지는 가장 작은 단위로 형태소를 쪼개면 더 이 상 기능이나 의미를 갖지 않게 됨.
실질적인 의미의 유무에 따라 실질 형태소(어휘)와 형식 형태소(문법 형태소)로 나뉨.
자립성의 유무에 따라 자립 형태소와 의존 형태소로 나뉘며 자립 형태소는 문장에서 홀로 쓰일 수 있으나 의존 형태소는 다른 형태
소와 결합되어 사용될 수 있음.

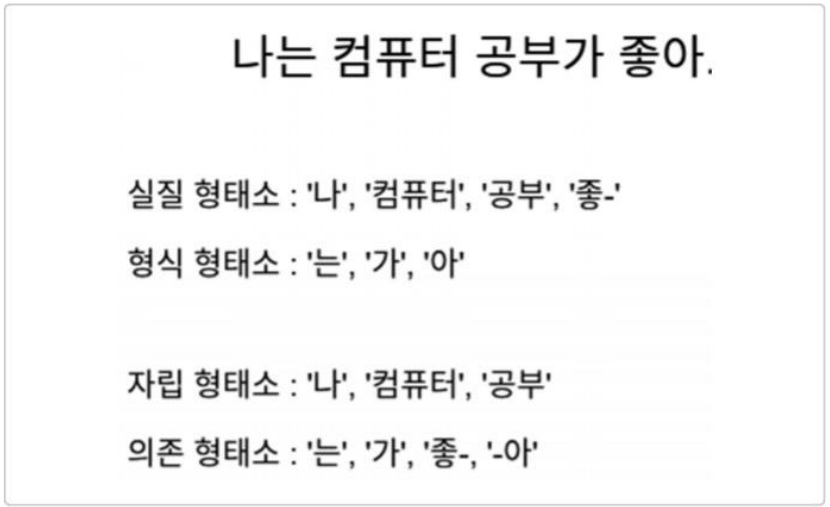

그림 3-3 형태소 분석

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

어절

- 어절은 한 개 이상의 형태소가 모여 구성된 단위로 발화 시에는 어절을 중심으로 끊어서 말하고, 글 을 쓸 때 어절은 띄어쓰기 단위와 거의 일치

바닷가에∨왔더니

바다와∨같이∨당신이∨생각만∨나는구려

바다와∨같이∨당신을∨사랑하고만∨싶구려

백석, 바다 中

그림 3-4 어절

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

품사
문장 내에서 단어가 수행하는 역할을 기준으로 체언, 수식언, 관계언, 독립언, 용언의 5언으로 나눔.
형태에 따라서는 가변어와 불변어로, 의미에 따라서는 명사, 대명사, 수사, 관형사, 부사, 조사, 감탄 사, 동사, 형용사의 9품사로 나눔.

그림 3-5 불변어와 가변어

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

품사

- 5언중에서는 용언이 가변어에 해당하여 쓰이는 형태가 변화하고, 체언, 수식언, 관계언, 독립언은 변 화가 없는 불변어

음절, 형태소, 어절, 품사

품사 - 체언

- 체언은 문장에서 몸통, 중심이 되는 역할을 하며 대개 조사가 뒤에 붙으며 9품사에서는 명사, 대명 사, 수사가 이에 속함

명사 :

보통 명사(볼펜, 종이, 구름, 희망), 고유명사(한라산, 서울), 자립 명사, 의존 명사(것, 따름, 뿐)

대명사 :

인칭 대명사(나, 그대, 이분, 누구, 어느분), 지시 대명사(이, 그것, 아무것), 의문 대명사(누구, 무엇, 어디)

수사 :

양수사(하나, 둘, 삼, 사, 대여섯, 예닐곱), 서수사(첫째, 제일, 제이, 일호, 이호, 서너째)

그림 3-6 체언의 종류

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

품사 – 수식언
수식언은 다른 말을 꾸며주는 역할을 하는 단어로 9품사에서는 관형사와 부사가 속함.
관형사는 체언 앞에 놓여 체언을 꾸며주는 역할을 하고 부사는 주로 용언, 즉 동사와 형용사 앞에서 그 내용을 꾸며주거나 혹은 문장 전체를 꾸며 줌.

관형사	부사
성상 관형사 : 새, 옛, 윗, 웃, 헌, 헛	성상 부사 : 잘, 매우, 촐랑촐랑
지시 관형사 : 이, 그, 다른, 무슨 등	지시 부사 : 내일, 저리, 이미
수 관형사 : 한, 두, 열, 첫째, 몇, 모든 등	부정 부사 : 안, 못
	양태 부사 : 반드시, 제발, 글쎄
	접속 부사 : 또는, 그리고, 왜냐하면, 즉

그림 3-7 수식언

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

품사 - 수식언
성상 관형사 : 성질이나 상태를 꾸며주고 지시 관형사는 어떤 대상을 가리켜 지시함.

수 관형사 : 사물의 양이나 수를 나타냄.
성상 부사 : '매우', '잘'과 같이 그 여하를 나타냄.

지시 부사 : 시간과 처소 또는 특정한 대상을 가리킴.
부정 부사는 용언의 뜻을 부정하는 역할을 함.
양태 부사는 말하는 이의 태도를 표현.

접속 부사는 단어와 단어를 이어주거나 문장과 문장을 이어주는 역할을 함.

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

품사 - 관계언

관계언은 문장에서 자립형태소에 붙어 문법적 관계를 나타내는 의존형태소로, '조사'가 있음.
조사는 체언 또는 용언의 명사형 등의 뒤에 붙어 말의 뜻을 더해주는 품사로 조사에는 격조사, 접속 조사, 보조사가 있음.

격조사 : 격을 나타내는 조사, 체언 등의 뒤에 붙어서 다른 말에 대한 그 말의 자격을 나타냄 주격, 서술격, 목적격, 보격, 관형격, 부사격, 호격 등이 있음.

접속 조사 : 두 단어를 이어주는 역할, '-와', '-과' 등이 있음.

보조사 : 여러 성분에 두루 붙어서 특별한 뜻을 더해주는 역할을 함

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

품사 - 독립언

독립언은 독립적으로 쓰이는 품사로 다른 품사를 수식하지도 않고 받지도 않음.
독립언에 해당하는 품사로는 감탄사가 있으며 놀람이나 감정 등을 나타내는 감정 감탄사, 말하는 사람의 뜻을 나타내는 의지 감탄사, 부름이나 대답을 나타내는 호응 감탄사 등이 있음.

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

품사 - 용언

용언은 독립된 뜻을 가지고 어미를 활용하여 문장 성분으로서 서술어의 기능을 하는 말.

동작이나 성질, 상태 등을 나타내는 단어가 용언으로 동사, 형용사가 이에 해당.

용언은 어간(語幹)과 어미로 이루어져 있음.

2. 언어학의 기본 원리

음절, 형태소, 어절, 품사

품사 - 용언
동사는 사물의 동작이나 작용을 나타내는 단어.
목적어의 필요성 유무에 따라 자동사와 타동사
행동의 자발성 여부에 따라 능동사와 피동사
행동의 주체가 누구냐에 따라 주동사와 사동사
쓰임에 따라 본동사와 보조동사
활용 형태에 따라 규칙 동사와 불규칙 동사로 나뉨.

음절, 형태소, 어절, 품사

품사 - 용언

- 나는 학교에 간다. (자동사)
- 동생이 피자를 먹는다. (타동사)

- 삼촌이 나에게 아기를 맡기다. (맡다 + 접사 '-기-' : 사동사)
- 범인이 경찰에게 잡히다. (잡다 + 접사 '-히-' : 피동사)

- 접다 : 접고, 접지, 접어, 접으니 (규칙 동사)
- 묻다 : 묻고, 묻지, 물어, 물으면 (불규칙 동사)

- 꼬마가 공을 들고 간다. (보조 동사 '들다' + 본동사 '가다')
- 밥을 먹지 아니하다. (본동사 '먹다' + 보조 형용사 '아니하다')

그림 3-8 여러 종류의 동사

3. 형태소 분석

형태소 및 형태소 분석 개념

- 형태소는 의미가 있는 최소의 단위로서 더 이상 분리가 불가능한 가장 작은 의미 요소입니다. 형태소 분석은 주어진 단어 또는 어절을 구성하는 각 형태소를 분리한 후 분리된 형태소의 기본형 및 품사 정보를 추출하는 것입니다.
- 형태소 분석(Morphological Analysis)은 자연어의 최소 의미 단위인 형태소를 식별하고 분류하는 과정입니다. 한국어와 같은 교착어에서 특히 중요한 전처리 단계입니다.
- 문장, 어절, 단어, 형태소의 계층구조
  - 어절: 띄어쓰기를 기준으로 구분되는 단위
  - 단어: 하나의 어절을 구성하는 단위
  - 형태소: 의미가 있는 최소의 단위

3. 형태소 분석

형태소의 종류

- **자립 형태소**
  - 명사: 사람, 장소, 사물을 나타냄
  - 대명사: 명사를 대신하는 말
  - 수사: 수량이나 순서를 나타냄
  - 감탄사: 감정이나 의성어, 의태어

- **의존 형태소**
  - 조사: 문법적 관계를 나타냄 (은/는, 이/가, 을/를)
  - 어미: 동사나 형용사의 활용을 나타냄
  - 접사: 단어의 의미를 변화시킴

3. 형태소 분석

한국어 형태소 분석의 특징

- **교착어적 특성**
  - 어근에 접사가 붙어 의미가 확장됨
  - 하나의 어절에 여러 형태소가 결합

- **불규칙 활용**
  - 동사/형용사의 불규칙 변화
  - 예: 듣다 → 들어, 푸르다 → 푸른

- **띄어쓰기 오류**
  - 실제 텍스트에서 띄어쓰기 규칙 미준수
  - 형태소 분석 정확도에 영향

3. 형태소 분석

한국어 형태소 분석의 특징

- **텍스트 전처리(Pre-processing)**
  - 텍스트 분석을 위해 문장 분리, 불필요한 문장 성분을 제거하는 과정입니다.
  - 입력 텍스트로부터 문장 부호를 기준으로 문장 분리
  - 문장 부호, 특수 문자, 숫자 등 불필요한 성분 제거
  - 띄어쓰기를 기준으로 어절 분리

- **품사 태깅(POS Tagging)**
  - 하나의 단어가 여러 품사를 가질 수 있는 모호성을 제거하는 과정입니다. 은닉 마르코프 모델(HMM) 등의 통계적 모델이나 규칙 기반 방법을 사용합니다.

3. 형태소 분석

키워드 추출

- **가용어, 불용어, 키워드 개념**
  - 불용어: 문서의 정보를 표현하지 못하는 단어 (한국어: 조사, 영어: 관사, 전치사)
  - 가용어: 불용어가 아닌 단어들
  - 키워드: 가용어 중에서 문서의 중심이 되는 주제어

- **키워드 추출 절차**
  - 형태소 분석 수행
  - 불용어 처리를 통한 가용어 추출
  - 가중치 계산을 통한 키워드 선정
  - 추출된 키워드를 텍스트 분석에 활용

3. 형태소 분석

말뭉치 작성

- 말뭉치 개념
  - 말뭉치는 언어 연구를 위한 자료의 집합으로, 일정 규모 이상의 크기를 갖추고 내용상으로 다양성과 균형성이 확보된 자료의 집합체입니다.
  - 말뭉치 종류

– 가공 방법에 따른 분류

- 원시 말뭉치: 텍스트를 컴퓨터 가독형 자료로 만든 말뭉치
- 가공된 말뭉치: 형태소 분석이나 품사 정보 등으로 가공한 말뭉치

– 작성 방법에 따른 분류

- 샘플 말뭉치 vs 모니터 말뭉치
- 범용 말뭉치 vs 특수목적 말뭉치
- 공시 말뭉치 vs 통시 말뭉치
- 문자언어 말뭉치 vs 음성언어 말뭉치

3. 형태소 분석

단어와 문서 관계 표현

- **단어-문서 행렬**
  - 여러 개의 단어가 모여 하나의 문서를 구성하는 관계를 행렬로 표현합니다. "단어 집합(Bag of Words)" 형태의 추상화된 텍스트 모델로, 문서 간의 유사도 측정이 용이합니다.
- **TF-IDF 가중치**
  - 문서 내에서 단어의 중요도를 측정하는 방법입니다:
  - TF (Term Frequency): 특정 단어가 문서 내에서 등장하는 빈도
  - IDF (Inverse Document Frequency): 단어를 포함한 문서 수의 역수
  - TF-IDF: TF와 IDF를 곱한 값으로 단어의 가중치 결정

- **공식:**
  - TF = (문서 내 단어 수) / (문서 내 모든 단어 수)
  - IDF = log(전체 문서 수 / 단어를 포함한 문서 수)
  - TF-IDF = TF × IDF

4. 단어사전 구축하기

분류 체계 구축

- **카테고리 키워드 개념**
  - 문서 분류의 품질을 좌우하는 핵심 요소로, 각 카테고리의 의미를 잘 표현하는 키워드 집합을 선정하고 정의하는 것입니다. 도메인 전문가 활용 방법과 문헌 참조 방법이 있습니다.

- **텍스트 분석 기반 카테고리 구축**
  - 입력 텍스트 수집 (위키피디아 등 활용)
  - 자연어 처리 (문장 분리, 형태소 분석, 불용어 처리, 개체명 인식)
  - 핵심 키워드 추출 (단어-문서 가중치 분석)
  - 유의어 분석 및 의미 확장
  - 키워드 분류 엔진 적용
  - 도메인 전문가 검증

4. 단어사전 구축하기

텍스트 분석 사전 구축

- **유의어 사전**
  - 서로 의미상 유사한 관계를 맺고 있으면서 동일한 문장 안에서 대체가 가능한 어휘를 분류한 자료집입니다.

  - 유의어 사전 구축 규칙
    - 교체 규칙: 문맥 속에서 단어를 후보 단어로 교체하여 동의어 식별
    - 배열 규칙: 동의성 정도가 모호한 단어들을 배열하여 의미 차이 파악

- **감성 분석 사전**
  - 긍정 및 부정 판단을 위한 감성어 사전으로, 다음 3단계로 구축됩니다:
    - 데이터 수집: 개인 블로그, 게시판, SNS, 온라인 상품 리뷰 등에서 감성 표현 데이터 수집
    - 주관성 탐지: 감성이 포함된 주관적 텍스트와 객관적 사실 진술 구분
    - 극성 탐지: 긍정, 부정, 중립 판단 및 극성 정도 정량화

5. 주요 분석 도구

KoNLPy 라이브러리

- Hannanum: 한나눔 형태소 분석기
- Kkma: 꼬꼬마 형태소 분석기
- Komoran: 코모란 형태소 분석기
- Okt: Open Korean Text

● 분석 결과 예시

입력: "자연어 처리는 정말 흥미로운 분야입니다."
결과: [('자연어', 'Noun'), ('처리', 'Noun'), ('는', 'Josa'),
('정말', 'Adverb'), ('흥미롭', 'Adjective'), ('은', 'Eomi'),
('분야', 'Noun'), ('이', 'Josa'), ('ㅂ니다', 'Eomi')]

5. 주요 분석 도구

분석 결과 활용

- 키워드 추출: 명사, 형용사 중심의 핵심 단어 식별
- 불용어 제거: 조사, 어미 등 의미가 적은 형태소 제거
- 정규화: 동일한 의미의 다양한 표현 통일

『1-12』
텍스트 데이터 분석 텍스트 마이닝

텍스트 데이터 분석 이해

형태소 분석

검색 트렌드 분석

텍스트 분류 방법

감성 분석

연관어 분석

학습 개요

학습목표

- 이 워크샵에서는 검색 트랜드에 대해 알 수 있습니다.

눈높이 체크

- 검색 트랜드를 알고 계신가요?

검색 트렌드 분석

검색 트렌드 분석의 개념

- 검색 트렌드 분석은 검색엔진에서 사용자들이 입력하는 검색어의 빈도와 패턴을 분석하여 사회적 관심사와 트렌드를 파악하는 기법입니다.

검색 트렌드 분석

주요 데이터 소스

- **Google Trends**
  - 전 세계 검색 데이터 제공
  - 지역별, 시기별 검색 패턴 분석
  - 연관 검색어 및 상승 검색어 정보

- **네이버 데이터랩**
  - 한국 검색 시장 점유율 1위
  - 연령별, 성별 검색 패턴 분석
  - 쇼핑 인사이트 데이터 제공

- **소셜미디어 해시태그**
  - 트위터, 인스타그램 트렌딩 해시태그
  - 실시간 이슈 파악 가능

검색 트렌드 분석

분석 방법론

- **시계열 분석**
  - 검색량의 시간별 변화 패턴 분석
  - 계절성, 주기성, 트렌드 성분 분해
  - 특이 이벤트에 따른 급증/급감 패턴 식별

- **비교 분석**
  - 여러 키워드 간의 검색량 비교
  - 경쟁사/경쟁 제품 간 관심도 비교
  - 지역별 검색 패턴 비교

- **연관어 분석**
  - 함께 검색되는 키워드 패턴 분석
  - 사용자 검색 의도 파악
  - 새로운 트렌드 키워드 발굴

검색 트렌드 분석

활용 사례

- **마케팅 전략 수립**
  - 제품 출시 타이밍 결정
  - 광고 키워드 선정
  - 타겟 고객층 식별

- **콘텐츠 기획**
  - 인기 토픽 기반 콘텐츠 제작
  - SEO 최적화 키워드 선정
  - 에디토리얼 캘린더 수립

- **사회 현상 분석**
  - 사회 이슈에 대한 관심도 측정
  - 정책 효과 모니터링
  - 위기 상황 초기 감지

검색 트렌드 분석

분석 시 고려사항

- 검색량의 상대적 지수 이해
- 외부 요인(이벤트, 뉴스)의 영향 고려
- 검색어의 다양한 표기법 고려
- 데이터의 시차 및 업데이트 주기 파악

『1-12』
텍스트 데이터 분석 텍스트 마이닝

텍스트 데이터 분석 이해
형태소 분석
검색 트렌드 분석
텍스트 분류 방법
감성 분석
연관어 분석

학습 개요

학습목표

- 이 워크샵에서는 텍스트 분류 방법에 대해 알 수 있습니다.

눈높이 체크

- 텍스트 분류 방법을 알고 계신가요?

K-NN(K-Nearest Neighbor) 분류

K-NN(K-Nearest Neighbor) 분류정의

- 분류되지 않은 입력 텍스트에 대하여 미리 분류된 K개의 텍스트 데이터와 유사도를 측정하여 가장 비슷한 카테고리로 분류하는 방법입니다.

- 장점
  - 간단하고 효과적인 분류
  - 데이터 분포 가정 불필요
  - 빠른 학습 과정

- 단점
  - 높은 비교 비용
  - 긴 분류 시간

의사결정트리(Decision Tree)

의사결정트리(Decision Tree) 분류

- 주어진 데이터를 분류하기 위해 훈련 데이터를 이용하여 트리 모델을 생성하는 지도 학습 방법입니다. 엔트로피를 기반으로 최적의 분류 속성을 선정합니다.

- 장점
  - 결과의 직관적 이해
  - 수치 및 범주 자료 모두 적용 가능
  - 대규모 데이터에 안정적 동작

- 단점
  - 부분 최적값 문제
  - 복잡한 트리 구조 가능성
  - 훈련 데이터 의존성

SVM(Support Vector Machine)

SVM(Support Vector Machine) 분류

- 카테고리 분류 경계로부터 각 카테고리 사이의 거리(margin)가 가장 크도록 분류 경계를 결정하는 방법입니다.

- 장점
  - 다차원 공간에서 효과적
  - 서포트 벡터 기반으로 빠른 분류
  - 고차원 데이터에 적합

- 단점
  - 대용량 데이터의 긴 학습 시간

나이브 베이즈(Naive Bayes)

나이브 베이즈(Naive Bayes) 분류

- 속성들 사이의 독립을 가정하는 베이즈 정리 기반의 확률 분류기입니다.
- 공식: $P(C|w_1...w_n) = P(C) \times \prod P(w_i|C) / P(w_1...w_n)$

- 장점
  - 대용량 데이터에 효율적 학습
  - 적은 훈련 데이터로도 높은 정확도
  - 빠른 처리 속도

- 단점
  - 미등장 입력 값의 확률 0 문제
  - Underflow 현상 가능성

『1-12』
텍스트 데이터 분석 텍스트 마이닝

텍스트 데이터 분석 이해
형태소 분석
검색 트렌드 분석
텍스트 분류 방법
감성 분석
연관어 분석

학습 개요

학습목표

- 이 워크샵에서는 감성 분석에 대해 알 수 있습니다.

눈높이 체크

- 감성 분석을 알고 계신가요?

감성 분석

감성 분석의 정의

- 감성 분석(Sentiment Analysis) 또는 의견 마이닝(Opinion Mining)은 텍스트에 표현된 주관적 정보를 추출하여 긍정, 부정, 중립 등의 감정을 분류하는 자연어 처리 기법입니다.

감성 분석

감성 분석의 단계별 분류

- **문서 수준 감성 분석**
  - 전체 문서의 전반적인 감성 판단
  - 리뷰, 평점 등에 주로 활용

- **문장 수준 감성 분석**
  - 각 문장별로 감성 분류
  - 문서 내 세부적인 감성 변화 파악

- **개체 수준 감성 분석**
  - 특정 대상에 대한 감성 분석
  - 제품의 여러 속성별 감성 구분

감성 분석 접근법

- 사전 기반 접근법 (Lexicon-based)
  - 감성 사전을 활용한 단어별 감성 점수 계산
  - 장점: 구현 용이, 해석 가능
  - 단점: 문맥 고려 부족, 도메인 의존성

- 기계학습 기반 접근법
  - 지도학습을 통한 감성 분류 모델 구축
  - 특성: 나이브 베이즈, SVM, 랜덤 포레스트 등
  - 레이블링된 학습 데이터 필요

- 딥러닝 기반 접근법
  - RNN, LSTM, BERT 등 신경망 모델 활용
  - 문맥과 순서 정보를 효과적으로 학습
  - 높은 성능, 대용량 데이터와 컴퓨팅 자원 필요

감성 분석

한국어 감성 분석의 특징

- **언어적 특수성**
  - 교착어적 특성으로 인한 복잡한 어미 변화
  - 높임법과 존댓말의 영향
  - 함축적 표현과 은유의 빈번한 사용

- **문화적 맥락**
  - 직접적 표현보다 간접적 표현 선호
  - 사회적 관계를 고려한 표현 방식
  - 세대별, 지역별 언어 사용 패턴 차이

감성 분석

평가 지표

- 정확도 (Accuracy)
  - 전체 예측 중 올바른 예측의 비율
  - 클래스 불균형 시 한계

- 정밀도 (Precision)와 재현율 (Recall)
  - 정밀도: 긍정 예측 중 실제 긍정의 비율
  - 재현율: 실제 긍정 중 올바르게 예측한 비율

- F1-Score
  - 정밀도와 재현율의 조화평균
  - 클래스 불균형 상황에서 유용

감성 분석

활용 분야

- **마케팅 및 브랜드 관리**
  - 제품 리뷰 모니터링
  - 브랜드 평판 관리
  - 고객 만족도 측정

- **금융 분야**
  - 뉴스 기반 주가 예측
  - 투자자 심리 분석
  - 리스크 관리

- **정치 및 사회**
  - 정책에 대한 여론 분석
  - 선거 예측
  - 사회 이슈 모니터링

『1-12』
텍스트 데이터 분석 텍스트 마이닝

텍스트 데이터 분석 이해
형태소 분석
검색 트렌드 분석
텍스트 분류 방법
감성 분석
연관어 분석

학습 개요

학습목표

- 이 워크샵에서는 연관어 분석에 대해 알 수 있습니다.

눈높이 체크

- 연관어 분석을 알고 계신가요?

연관어 분석

연관어 분석의 개념

- 연관어 분석은 텍스트 데이터에서 특정 단어와 함께 자주 등장하거나 의미적으로 관련된 단어들을 찾아내는 분석 기법입니다. 이를 통해 단어 간의 관계와 패턴을 파악할 수 있습니다.

연관어 분석

연관어 분석 방법

- **동시 출현 분석 (Co-occurrence Analysis)**
  - 일정 범위 내에서 함께 나타나는 단어들의 빈도 계산
  - 문서, 문장, 또는 n-gram 단위로 분석
  - 공출현 행렬(Co-occurrence Matrix) 생성

- **상호정보량 (Mutual Information)**
  - 두 단어가 독립적으로 나타날 확률 대비 함께 나타날 확률
  - PMI(Pointwise Mutual Information) = log(P(x,y) / (P(x)P(y)))
  - 높은 값일수록 강한 연관성

- **Lift 계수**
  - 한 단어가 나타났을 때 다른 단어가 나타날 가능성의 증가율
  - Lift = P(Y|X) / P(Y)
  - 1보다 클 때 양의 연관성

연관어 분석

네트워크 분석

- **단어 네트워크 구축**
  - 단어를 노드(Node), 연관성을 엣지(Edge)로 표현
  - 네트워크 시각화를 통한 관계 파악
  - 중심성 분석을 통한 핵심 키워드 식별

- **중심성 지표**
  - 연결 중심성: 직접 연결된 노드의 수
  - 근접 중심성: 다른 노드들과의 거리 기반
  - 매개 중심성: 다른 노드 간 최단 경로에 위치하는 정도
  - 고유벡터 중심성: 연결된 노드의 중요도 고려

연관어 분석

토픽 모델링

- 잠재 디리클레 할당 (LDA, Latent Dirichlet Allocation)
  - 문서 집합에서 잠재된 주제를 발견
  - 각 문서를 여러 주제의 혼합으로 가정
  - 주제별 연관어 추출 가능

- Word2Vec과 Doc2Vec
  - 단어의 분산 표현(Distributed Representation) 학습
  - 의미적으로 유사한 단어들의 벡터 공간상 근접성
  - 코사인 유사도를 통한 연관어 추출

연관어 분석

분석 결과 해석

- **연관어 강도 평가**
  - 통계적 유의성 검증
  - 임계값 설정을 통한 노이즈 제거
  - 도메인 전문가의 검토

- **시간적 변화 분석**
  - 시기별 연관어 패턴 변화
  - 트렌드 변화에 따른 연관성 진화
  - 이벤트 전후 연관어 비교

연관어 분석

활용 사례

- **추천 시스템**
  - 사용자 관심사 기반 콘텐츠 추천
  - 상품 연관 구매 패턴 분석
  - 검색어 자동완성 기능

- **마케팅 인사이트**
  - 브랜드 연상 이미지 분석
  - 경쟁사와의 연관어 비교
  - 신제품 포지셔닝 전략 수립

- **학술 연구**
  - 연구 분야 트렌드 분석
  - 학술 논문 분류 및 추천
  - 연구자 간 협업 네트워크 분석

실습34 : 텍스트 데이터 분석 및 텍스트 마이닝

문제
[booking.com_hotel_review.csv] 업로드
koreanize_matplotlib-0.1.1-py3-none-any.whl, NanumBarunGothic.ttf 업로드한 라이브러리를 설치하고 Matplotlib 한글 사용 환경을 설정 한 다음 나눔체로 한글을 표현해 줘. 이 데이터를 탐색해줘

THANK YOU.

앞으로의 엔지니어는 단순한 '코더'나 '기계 조작자'가 아니라 뇌-기계 인터페이스를 통해 지식과 능력을 즉각 확장하는 존재(뉴로-인터페이스: Neuro Interface)가 될 수 있습니다.

🎯 목표 달성을 위한 여정이 시작됩니다.
✨ 궁금한 점이 있으시면 언제든 문의해주세요!
🚀 함께 코더와 프롬프트 전문가로 성장해 나갑시다!