Chapter. 03

자연어 분석 기초

자연어 모델링 I

FASTCAMPUS
ONLINE

금융공학/퀀트 I

강사. 장순용

I 키포인트

- 말뭉치.
- 불용어 vs 가용어.
- 키워드.
- 품사 태깅 (POS Tagging).
- 형태소 분석.

I 자연어 분석과 금융

- 수치형 데이터 뿐만이 아니라 자연어 데이터를 활용한 포괄적인 방법으로 시장을 이해하려 한다.

  ⇒ 자연어 데이터는 뉴스, SNS 등에서 추출할 수 있다.

- 수치형 데이터의 경우와 마찬가지로 AI를 적용할 수 있다.

  ⇒ 탐색, 시각화, 비지도학습, 지도학습, 등의 방법.

I 자연어 분석 개요

- 문서에서 특징을 추출하여 분류, 요약, 군집화, 감성분석, 등 수행.
- 문서의 내용을 통계적으로 분석함.

  ⇒ 인간이 이해하는 방식과는 차이가 있음.
- 문서의 내용은 최소 분석 단위로 분해되고 처리됨.
- 비정형 데이터인 문서의 정형화 과정이 포함됨.

모델링

I 말뭉치 (Corpus)

- 연구 재료로서 언어적 자료의 집합을 의미한다.

예). 소설, 뉴스, 등

⇒원시 말뭉치(raw corpus): 텍스트를 컴퓨터가 읽을 수 있는 형태로 저장해 놓은 것.

⇒가공된 말뭉치(tagged corpus): 형태소 분석이나 어휘, 품사 정보, 문헌, 내용 등으로 분류할 수 있도록 가공된 말뭉치.

I 전처리

- 전처리 단계:

1. 문장 부호 (마침표, 쉼표)를 기준으로 문장을 분리. 영어인 경우에는 소문자화 포함.

2. 분리된 문장에서 문장 부호, 특수 문자, 숫자 등 불필요한 요소 제거.

3. 분리된 문장에서 어절 단위로 떼어 놓는다. 어절은 띄어쓰기를 기준으로 구분.

I 불용어

- 특별한 정보를 제공하지 못하는 단어를 불용어 (stopword)라 부른다.

  → 영어에서의 대표적인 불용어는 관사와 전치사이다.

  예). “a”, “the”, “on”, “with”.

  → 한국어에서의 대표적인 불용어는 조사와 어미이다.

  예). “는”, “을”.

- 불용어는 불용어 사전을 사용해서 제거된다.

I 가용어와 키워드

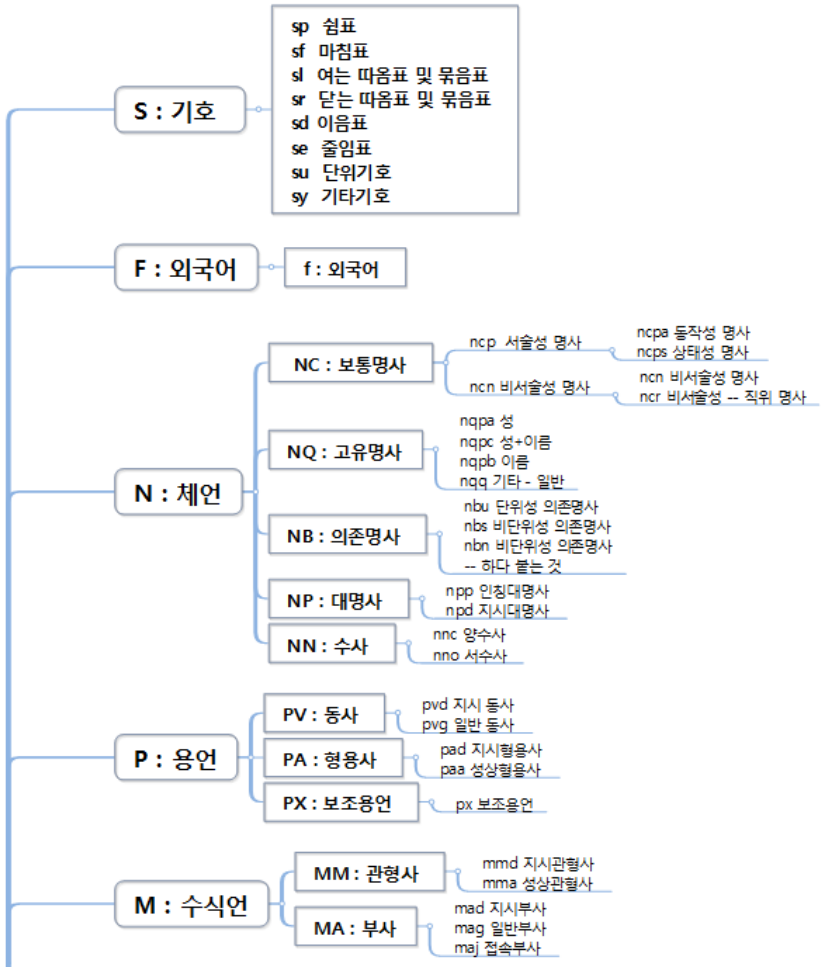
- 불용어의 반대는 가용어이다.
- 가용어 중에서 문서의 주체어를 키워드라고 부른다.

→ 보통은 문서 내에서 발생 빈도가 높은 단어들을 키워드로 선정한다.

→ 키워드 선정은 분석하고자 하는 목적 및 문서의 특성을 따르는 것이 원칙이다.

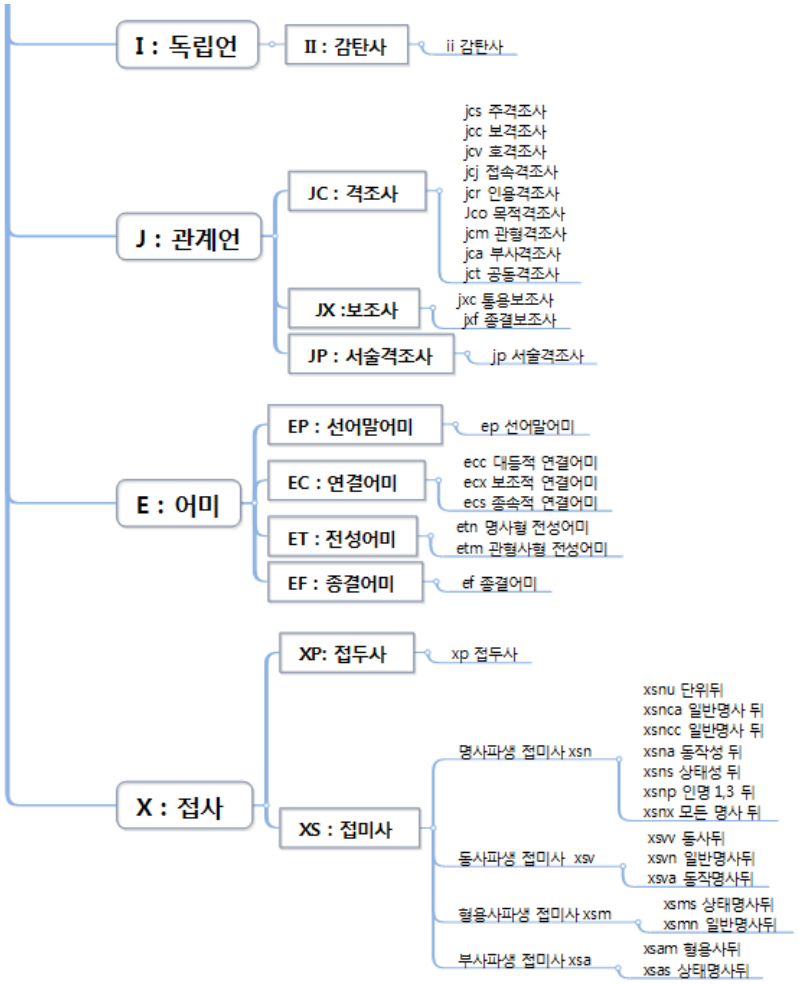
I 품사 태깅 (POS Tagging)

- 한국어 KAIST 품사 태그 셋.



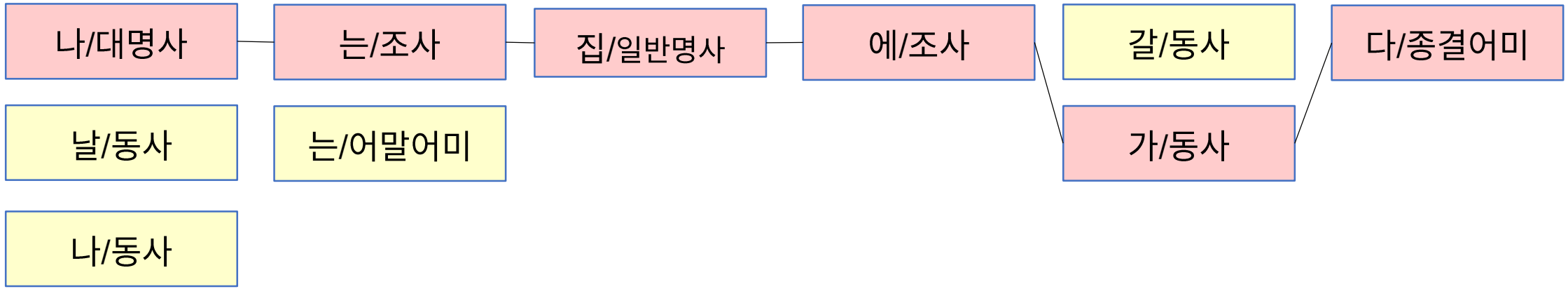
I 품사 태깅 (POS Tagging)

- 한국어 KAIST 품사 태그 셋.



I 품사 태깅 (POS Tagging): 예

예). “나는 집에 가다”



I 형태소 분석

- 형태소는 의미가 있는 최소의 단위로서 더 이상 분리가 불가능한 요소.

⇒ 문법적, 관계적 뜻을 나타내는 단어 또는 단어의 부분이다.

⇒ 어간과 어미 단위도 형태소로 간주한다.

⇒ 어간을 분리해내는 텍스트 처리기술을 stemming 이라고 한다.

예). “intelligent”, “intelligence”, “intelligently” ⇒ “intelligen”

예). “잡히고”, “잡히다” ⇒ “잡히”

즉, 철자법과는 무관하다.

I 형태소 분석

- 형태소 분석이란 주어진 단어 또는 어절의 형태소에서 기본형 및 품사 정보를 추출하는 것.

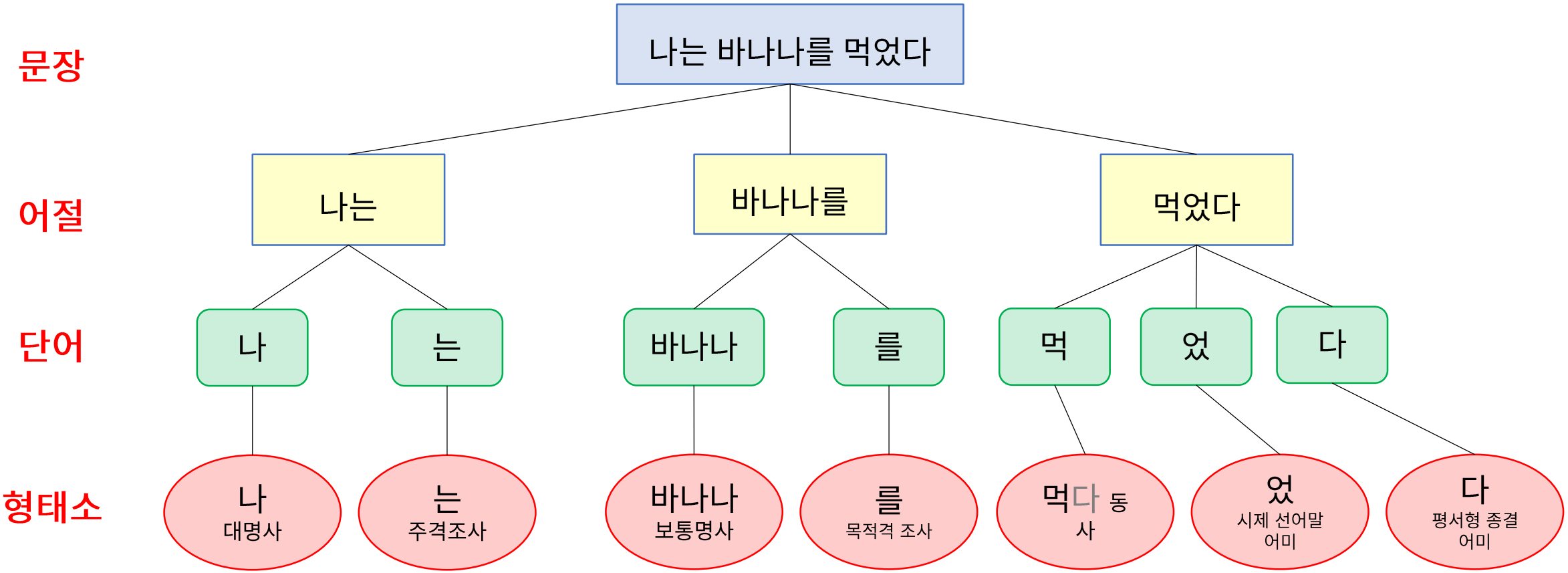
⇒ **어절**: 한국어에서는 띄어쓰기를 기준으로 어절을 구분한다.

예). “나는 바나나를 먹는다” = “나는” + “바나나를” + “먹는다”.

⇒ **단어**: 어절을 구성하는 단위로 어절은 하나 또는 두 개 이상의 단어로 구성됨.

예). “나는” = “나” + “는”.

I 형태소 분석: 도식화



FAST CAMPUS
ONLINE
장순용 강사.

끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.