Corpus : 자연언어 연구를 위해 특정한 목적을 가지고 언어의 표본을 추출한 집합

방대한 데이터를 이용하여 데이터 학습 훈련하여 문제에 대한 정의가 포함되어 있는 데이터베이스를 만들어 해답을 출력

Wordnet:

어휘 목록, 영어 단어를 'synsest'이라는 유의어 집단으로 분류, 간략하고 일반적인 정의 제공 , 어휘 목록 사이의 다양한 의미 관계를 기록

목적: 사전과 시소러스의 배합을 만들어 보다 직관적인 사용, 자동화된 본문 분석과 인공지능에 응용

텍스트 정규화(text normalization)는 미리 정의된 규칙에 맞추어 텍스트를 변환하는 작업으로, 여기에는 공백 제거, 구두점 삭제, 대문자/소문자 변환 등이 포함된다.

Lemma: 단어의 기본형 Wordform: 전체 어미 변경 가능한 일부 표면 형태이다. Cat vs Cats

Type: 어휘의 한 요소 token: 실행 중인 텍스트에서 그 유형의 인스턴스

Tokenization이란 문자열을 여러 개의 조각, 즉 여러 개의 Token (토큰)들로 쪼개는 것을 말한다.

Word Segmentation Algorithm: 문자열의 시작 부분에서 포인터 시작 ->사전에서 포인터에서 시작하는 문자열과 일치하는 가장 긴 단어 찾기 -> 문자열의 단어 위로 포인터 이동

Normalization: 모든 데이터 포인트가 동일한 정도의 스케일(중요도)로 반영되도록 해주는 게 정규화(Normalization)의 목표

구구조 문법

구성소 관계 (Constituency relation) 에 기반

단어 -> 절 -> 문장문장 전체를 트리 구조로 분석

WordNet 상위어 하위어 등위어 전체어 부분어

Lemmatization: 원형을 찾는 행위(사전) (am, is, are  be)

Stop word: 전치사, 접속사(fuction word) 속아내는 기법

Minimum Edit distance: 두 그룹의 차를 구해서 차가 큰 것  다른 언어라고 파악 Insertion, Deletion, Subsitution

Morphology: 단어를 구성하는 작은 단위 어간, 접사

Stemming: 어근을 찾는 행위

Dynamic Programming: 문제를 조그마게 자르고 그것들의 해결책을 한 후 종합하여 큰 문제를 해결하는 방법

청킹은 문장을 개별 단어보다 더 의미 있는 구로 그룹화하는 것

과정 : 문장 -> 토큰화 -> 품사(POS) 태깅 -> 청크(Chunk) 태깅

재귀 하향 파서

왼쪽에서 오른쪽으로 입력을 읽고 파생 트리를 하향식으로 작성,전위순회 방식으로 노드를 통과하는 파서

conll2000 말뭉치: 뉴스(WSJ) 텍스트의 27만개 단어로 구성된 청킹 코퍼스Trian/text set 으로 나뉘어져 있음

treebank\_chunk : 각 단어별로 태그된 treebank 말뭉치

BIO 표현

B : Begin의 약자, 개체명이 시작되는 부분

I : Inside의 약자, 개체명의 내부 부분

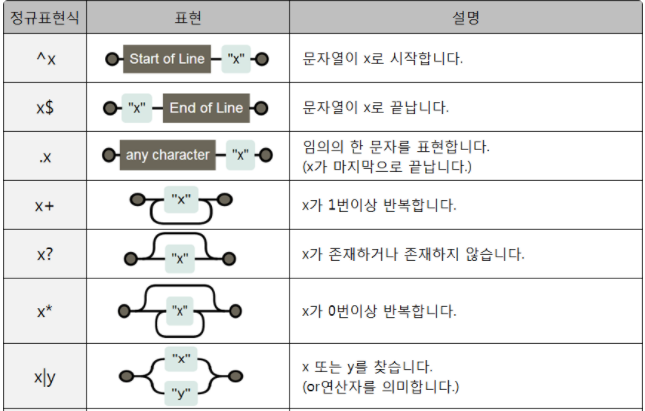
O : Outside의 약자, 개체명이 아닌 부분

NLTK 패키지 말뭉치 토큰 생성 형태소 분석 품사 태깅

텍스트이(가) 표시된 사진

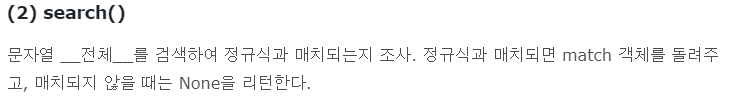
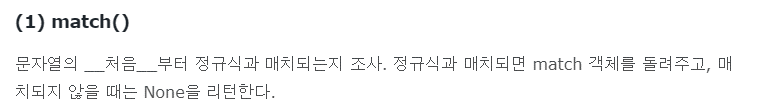
자동 생성된 설명

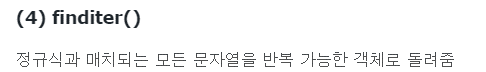
Regular Expressions



텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

False positives(Type I)

코로나 진단키트에서 거짓 양성

There, then, other를 다 발견함

False negatives(Type ll)

실제론 양성인데, 발견하지 못함

체언(명사,대명사,수사)

용언(용언의 활용,동사,형용사)

수식언(관형사,부사)

독립언(감탄사)

관계언(조사)

Noun(명사), verb(동사), adjective(형용사),preposition(전치사), adverb(부사), article(관사), interjection(감탄사),pronoun(대명사), conjunction(접속사)

테이블이(가) 표시된 사진

자동 생성된 설명

VBZ 동사 현재형 3인칭 단수

CC 등위 접속사

DT 한정사

IN 전치사, 종속 접속사

JJ 형용사

NNP 고유명사

PRP 대명사

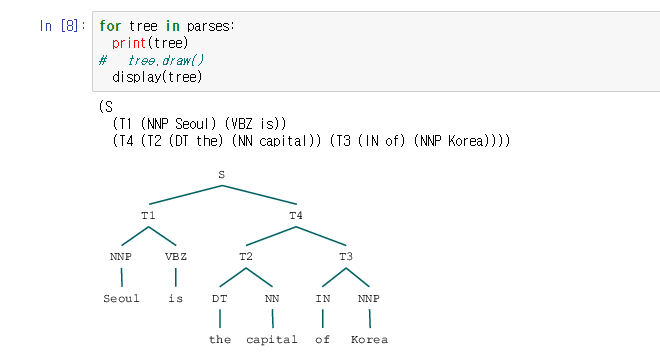


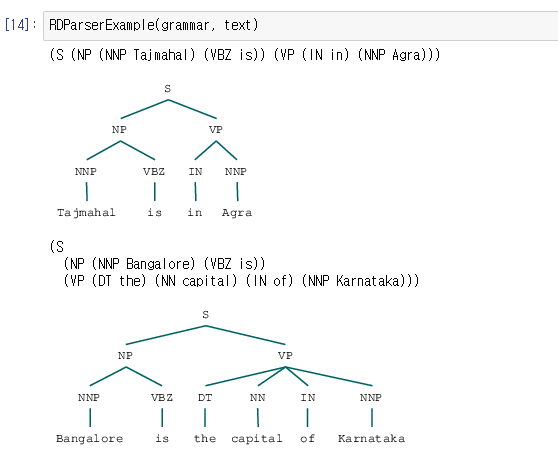
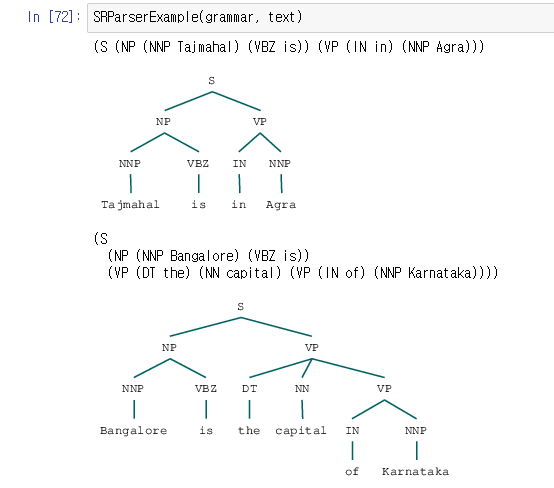
텍스트이(가) 표시된 사진

자동 생성된 설명learnDefaultTagger [(‘전부’,’NN’), ~]

learnRETagger [('Mysore', None), ('is', 'VERB'), ('an', 'INDEFINITE-ARTICLE'), ('amazing', 'ADJECTIVE'), ('place', None), ('on', 'PREPOSITION'), ~전부 None]]

learnLookupTagger [('Mysore', None), ('is', 'VBZ'), ('an', 'DT'), ('amazing', 'JJ'), ('place', 'NN'), ('on', 'IN'), ('earth', 'NN'), ('.', '.'), ('I', None), ('have', None), ('visited', None), ('Mysore', None), ('10', None),('times',None),('.','.')]텍스트이(가) 표시된 사진

자동 생성된 설명

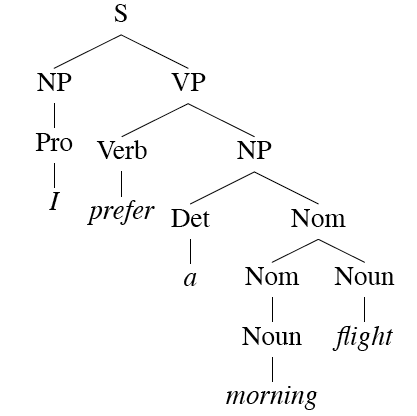
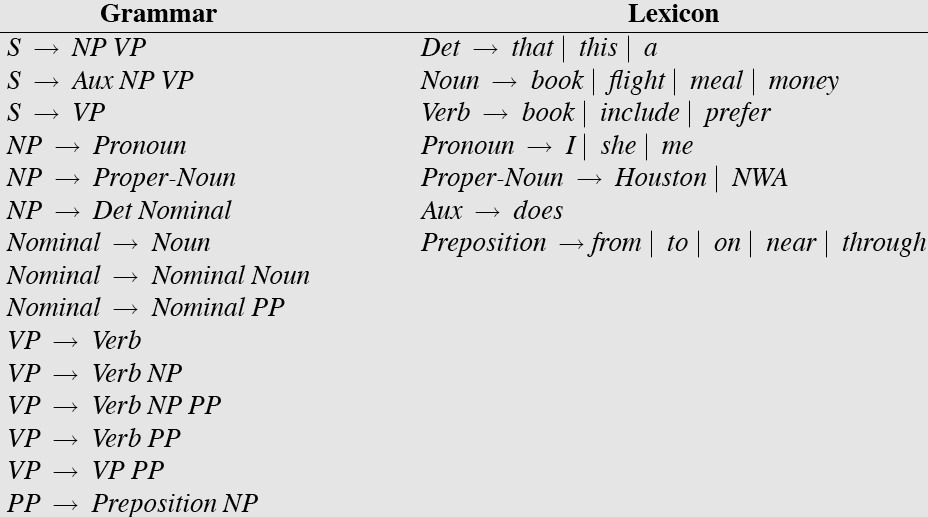


ShiftReduceParse

상향식, bottom up

Recursive Descent Parser

하향식 Top down

labeledDepParse.tiff

구구조 문법 (Phrase Structure Grammar)

의존 문법 (Dependency Grammar)