OWLNEST® SE-Mainframe

Proactive discovery and insights from contexts

OWLNEST Corp.
주소 : 서울특별시 강남구 자곡로 174-10 강남에이스타워 G9 318호 (06373)
문의 : contact@owl-nest.com, +82 2-742-3021

What is OWLNEST® SE-Mainframe ?

OWLNEST® SE-Mainframe 은 향상된 언어처리 기술을 탑재한 통합 검색 메인프레임 솔루션입니다. 본 시스템을 통해 비정형 텍스트 문서에 존재하는 키워드와 문맥 정보를 찾을 수 있습니다. 텍스트 처리와 정보의 접근성 향상을 위해 효과적인 알고리즘을 적용합니다. 본 시스템은 PDF, TXT, PPT, EXCEL, HTML 등으로 존재하는 텍스트 데이터를 읽고 이해하기 위한 작업을 자동화함으로써 사용자의 시간과 자원을 절약합니다.

Why is OWLNEST® SE-Mainframe important?

OWLNEST® SE-Mainframe 은 대규모 비정형 텍스트 데이터를 최적의 인덱스 구조로 처리하여 고품질의 검색결과로 표시합니다. 이를 통해 시스템 사용자는 검색과 관련된 명확하고 객관적인 정보를 얻을 수 있습니다. 더 나아가서 사용자는 분석대상과 관련된 데이터 경향성을 한눈에 파악할 수 있고, 가치와 위험요소를 탐지하는 기회를 얻게 됩니다. 일련의 과정을 거치면서 사용자는 자신에게 주어진 기회와 가치에 집중할 수 있고, 수많은 텍스트로부터 손쉽게 인사이트를 얻을 수 있는 기능을 제공합니다.

For whom is OWLNEST® SE-Mainframe intended?

OWLNEST® SE-Mainframe 은 양적 데이터를 처리해야 하는 개인과 조직을 위해 고품질의 검색 결과를 빠른시간에 제공할 수 있습니다. 이를 이용해 대규모 텍스트 데이터로부터 정보를 유형화하고 경향성을 파악하려는 비즈니스 전문가와 데이터 분석가들이 일차적인 고객이 될 수 있습니다. 또한 사회적 트렌드를 반영한 상품, 서비스 기획이 필요한 마케터나 홍보담당자에게 요긴한 도구가 될 수 있습니다. 더 나아가서는 여론의 흐름을 읽고 이에 대처하는 여론분석 혹은 컨설턴트에게도 중요한 함의를 제공할 수 있습니다.

현재 기업과 조직이 주목해야할 비정형 텍스트 정보는 급격하게 증가하고 있습니다. 이에 따라 올바른 의사결정을 위해 전체 텍스트 데이터에서 중요한 논점과 관련 필수 정보를 우선적으로 파악하는 작업의 중요성은 점진적으로 강조되고 있습니다. 반면 조직의 가치와 위험요인을 발굴해야 하는 분석가가 언론보도기사, 고객의 피드백, 리뷰, 이메일, 웹문서, 블로그, 까페, SNS 데이터, 지식인, 불편사항, 분석보고서 등 모든 종류의 텍스트 데이터를 모두 읽고 파악하는 작업은 사실상 불가능에 가깝습니다.

대규모 텍스트 데이터에서 경향성을 발견하고 새로운 비즈니스 이슈를 도출하기 위해서, 전체 데이터를 대상으로 객관적 정보를 파악하는 작업이 선행되어야 합니다. 그러나 일상 언어를 포함하는 비정형 텍스트 데이터는 정형 데이터와 달리 중의성을 띠고 핵심 내용이 명시적으로 드러나지 않는 경우가 많습니다. 이를 해결하기 위해 텍스트에 잠재된 데이터를 구조적 정형 데이터와 병합하는 작업이 필요합니다.

OWLNEST® SE-Mainframe 은 사용자가 웹스케일 수준의 관련 데이터에 실시간으로 접근하여 적절한 정보를 찾아내고 분석할 수 있도록 지원합니다. 본 시스템은 전체 데이터에서 주제와 내용을 탐지하고, 특정 용어와 부합하는 명시적 분류와 개체간의 관계를 파악해내는 일련의 과정에 따라 동작합니다. 모든 과정이 사용자 입장에서 주도적으로 수행할 수 있도록 자동 처리됩니다. 또한 분석 관점에서 알고리즘을 설계하고 이를 모듈화하여 목표 시스템에 적용한다는 점에서도 본 솔루션의 차별적 가치가 드러납니다.

OWLNEST® SE-Mainframe 을 이용한 검색과 텍스트 분석에 대한 가치판단의 예로써 미디어 콘텐츠 관련 사례를 들 수 있습니다. 경험재인 미디어 콘텐츠는 비정형 텍스트로써 투자규모에 따른 위험 회피(hedge)가 어려운 재화입니다. 콘텐츠 제작에 투입되는 자원에 따라 투자 회수 여부가 예측되지 않기 때문입니다. 이에 따라 미디어 콘텐츠 성과에 대한 예측 능력은 텍스트 분석의 핵심 역량을 필요로 합니다. 인터넷 기사, SNS 등을 통해 특정 미디어 콘텐츠에 대한 언급, 평가 등의 데이터를 수집할 수 있고, 이를 응용해 최적의 대응 시스템을 구현할 수 있습니다.

Key Benefits

고속 검색을 통한 의사결정 시간의 절약
집약된 자연어처리 기술과 검색 알고리즘을 통해, 검색 사용자의 입력에 대한 고민을 최소화 합니다. 키워드의 문법표지 부착, 개체명 인식, 구문단위 추출, 자동 띄어쓰기, 문장단위 분할, 컨텍스트 패러프레이즈 등의 기능을 자동화하여 효과적으로 제공합니다. 이를 통해 사용자는 분석업무에 집중할 수 있고 가치발견의 가능성을 더욱 높일 수 있습니다.

키워드별 고수준 메타정보의 통합 제공
문자열 토큰 단위 통합적 언어처리기능을 통해 시스템 사용자는 각 키워드별로 문법표지, 개체명분류, 의존관계 등 고수준 메타정보를 한번에 파악할 수 있습니다. 목적별 정보 취합, 개별 메타정보에 대한 기능단위 분할처리기능을 제공합니다. 고수준 메타정보의 통합적인 제공을 통해, 텍스트 분석가 입장에서 필요로 하는 기능을 편리하게 이용할 수 있습니다.

경향성의 파악과 비즈니스 기회로 집중
OWLNEST® SE-Mainframe 은 텍스트 데이터 전체 집합의 구조를 간략히 파악할 수 있도록 단어 인덱싱을 지원하고 이를 벡터화된 기본 정보로 나타냅니다. 여기에 데이터 마이닝과 통계 모델링 기법을 적용해 인사이트를 발굴할 수 있습니다. 즉, 사용자는 비정형 텍스트 데이터로부터 고객의 니즈에 귀 기울이고, 서비스와 제품의 요구사항을 파악함으로써 비즈니스 기회를 극대화할 수 있습니다.

기계학습 인터페이스
OWLNEST® SE-Mainframe 은 텍스트 분석시 기계학습을 통한 문서 자동 분류에 대한 니즈를 해결해야할 경우, 단어 벡터화 및 통계정보 처리, 기계학습 예제 템플릿을 지원함으로써, 정보 검색의 범위를 자연스럽게 기계학습으로 확장할 수 있는 연결점을 제공합니다. 즉, 텍스트 분석의 요구사항 전반을 SE-Mainframe 의 아키텍처 범위 안에서 수용할 수 있습니다.

정보추출 인터페이스
검색엔진의 핵심은 높은 재현율로 유의미한 키워드 토큰을 확보하는 일이 핵심입니다. OWLNEST® SE-Mainframe 이 제공하는 고속 토크나이저는 검색엔진을 위한 최적화된 정보추출 인터페이스로써 정보검색에서 중요한 역할을 담당합니다.

OWLNEST® SE-Mainframe System Requirements

To learn more about OWLNEST® SE-Mainframe, please visit http://owl-nest.com/technology, test the applications and see the references.

OWLNEST® SE-Mainframe

Proactive discovery and insights from contexts

OWLNEST Corp.
주소 : 대전광역시 유성구 문지로 193 KAIST ICC 학부동 F736 (305-732)
문의 : contact@owl-nest.com, +82 2-742-3021

Solution Overview

OWLNEST® SE-Mainframe 은 비정형 데이터 검색 및 분석 모델링 도구의 집합체로서 텍스트 집합의 문맥적 의미를 발견하고 가치있는 정보를 추출하는 기능을 제공합니다. 단어와 구문 수준의 메타정보와 표현 수준에서 통계적 특징과 패턴에 관계된 명시적 정보를 자동처리 함으로써 사용자의 생산성 향상에 기여합니다. 본 솔루션은 비정형 텍스트 데이터의 검색, 양적 분석, 텍스트 마이닝, 사례기반 예측분석 등과 같이 대규모의 문서 집합에 포함된 정보를 손쉽게 찾아내기 위한 고성능의 검색 알고리즘을 지원합니다. 텍스트 처리에 관련된 작업은 대규모 연산을 요구하므로 신속한 데이터 처리를 위한 병렬처리 및 다중코어 작업을 위한 제어기능을 기본 제공합니다. 본 솔루션은 소프트웨어 자원과 하드웨어 제어 기능을 동시에 지원하기 때문에 사용자는 손쉽게 텍스트 분석의 요구사항을 달성하고 숨은 가치를 찾아낼 수 있습니다. 오늘날 하루가 다르게 증가하는 비정형 데이터의 양을 고려했을 때, 잠재가치의 발굴과 위험요인의 분석을 위해서는 정량적 텍스트 처리를 통한 정성적 분석에 대한 필요성은 더 이상 간과할 수 없는 현실적 문제입니다. 본 솔루션을 통해 비정형 텍스트로부터 의미있는 데이터를 추출함으로써 컨텍스트 정보를 이용하여 서도정의

Key Features

다양한 형식의 문서 포맷 처리 기능과 내부 파일시스템
- 취급 파일 (pdf, doc, ppt, hwp, xls, txt) 의 통합 저장, 메타정보 관리
- csv, json, xml 등의 파일출력 및 DB 인터페이스 지원

차별화된 검색 기능 지원 (키워드, 자연어, 컨텍스트)
- 키워드, 자연어 검색 및 컨텍스트 검색으로 유사 문장/문서 자동 추출
- 검색결과의 카테고리 분류를 위한 최적의 솔루션

개체명 및 의존관계 메타정보, 키워드 빈도, 연관어 분석
- 중요 키워드에 대한 일반/특수/확장 개체명 정보 자동 인식
- 개체명-표현 간의 의존관계의 자동 추출을 통한 메타정보화
- 키워드 빈도 및 연관어의 통계적 분석 결과 제공

유니코드 인코딩 적용으로 다국어 일괄 처리
- 유니코드 처리를 통한 텍스트 데이터 저장/관리
- 언어별 텍스트 관리/저장/추출을 위한 손쉬운 사용자 설정 기능

웹스케일 데이터의 처리, 고속 병렬 토크나이저
- 웹스케일의 입력 텍스트의 처리를 위한 고속 병렬 문자열 토큰화 기능
- 토큰별 인덱스 생성과 메타정보의 연결을 통한 빠른 정보처리

FileSystem / InMemory Indexer
- 파일시스템을 기본으로 하는 인덱싱 최적화 솔루션
- 인메모리 인덱싱을 지원하여 검색 속도 최적화
- 인덱스 압축/암호화를 통한 보안이슈의 솔루션 제공

SE-Mainframe Client/Server
- 시스템 통합을 위한 Cross-Language Development Driver 지원

1 문장간 문맥적 유사도를 계산하여 검색
- 두 키워드는 전혀 다른 의미를 가지지만, 유사한 컨텍스트를 가진 경우 출현
- 서로 다른 키워드로 두 문장을 한번에 검색 가능
- 컨텍스트 정보를 이용한 문장 패러프레이즈 검색

2 특정 키워드 중심 재검색
- 검색결과를 기반으로 특정 키워드에 관련된 문맥을 재검색
- 특정 키워드에 대한 가중치 부여를 통해 포커싱한 문맥정보의 검출

3 입력 검색어 길이 무제한 지원
- 검색어 문자열의 길이를 무제한 입력처리
- 파일시스템/인메모리 최적 제어를 통한 빠른 검색 속도

4 연관어, 개체분류, 독음/동형이음 검색
- 연관어 통계 분석을 통한 자동 연관어 추천/검색 지원
- 개체분류를 통한 메타 정보 검색
- 독음/동형이음 검색 기능

일반 검색엔진

사용자 요구사항	입력 키워드를 포함하는 문서를 조회
입력 쿼리	**키워드, 조건 (AND, OR)**
검색결과 특징	입력 키워드가 포함된 문서 화면노출
랭킹 조건	날짜, PageRank
특징	컨텍스트 유사문서 검색 불가

컨텍스트 검색엔진

사용자 요구사항	입력 문장의 컨텍스트 유사문서 조회
입력 쿼리	**일반 자연어 문장/문서**
검색결과 특징	입력 문장/문서와 유사한 문서 화면노출
랭킹 조건	날짜, 컨텍스트 유사도
특징	키워드 불포함 문서도 추출

NAVER 검색엔진	
입력	원래 모든지 잘먹던 아빠인데 당뇨가 생긴뒤로는 밥을 포함한 모든 음식을 거의 안먹어요 그래서 지금 엄청 말랐어요ㅠㅠ 그리고 요즘 주변분들이 당뇨로인해서 합병증도오고,병원에 입원하는경우를 많이봐서 많이 걱정이되요 원래 당뇨가 있으면 입맛이 없는건가요?
1	그래서 사람의 몸 속에서도 수분 및 영양분을 조절하여... 떨어뜨려주어서 당뇨에도 상당히 좋다. (참고 - S.B.S 잘.먹.고.잘.사.는.법 : 칡은 만능의 약재이며 음식이다.)... 치매를 포함한 모든 형태의 치매 위험이 평균 25% 낮은 것으로 ...
2	그리고 오리팬은 피지를 깨끗이 제거해줌으로, 몸 전체... 호두는 그냥 먹어도 맛있지만 호두에는 체내에 잘 흡수되지 않는 성분이 포함되어 있으므로 기름으로 짜서 먹거나... # 모든 병의 근본적인 치료는 음식에 있습니다 ...
3	그리고 콩류 음식 많이 드시면 좋습니다. 콩 중에서도... 호두는 그냥 먹어도 맛있지만 호두에는 체내에 잘 흡수되지 않는 성분이 포함되어 있으므로 기름으로 짜서... 죽염= 모든 생물이 썩지 않는 것은 염성의 힘 때문인데 ...

컨텍스트 검색엔진	
입력	원래 모든지 잘먹던 아빠인데 당뇨가 생긴뒤로는 밥을 포함한 모든 음식을 거의 안먹어요 그래서 지금 엄청 말랐어요ㅠㅠ 그리고 요즘 주변분들이 당뇨로인해서 합병증도오고,병원에 입원하는경우를 많이봐서 많이 걱정이되요 원래 당뇨가 있으면 입맛이 없는건가요?
1	당뇨로입원중이신아버지가있는데병원밥이입맛에맞지않는지잘먹지도못하고그래서할머니께서김치랑젓갈무쳐놨다고가져다주라는데 아무리밥을못먹어도당뇨면김치면몰라도젓갈은짠음식이니까먹으면안되죠? ...
2	당뇨로 인해 급격한 체중변화가 있은후부터 이런 증상이 있었던거 같은데 다리근육이 뭉치는 느낌이 들면 양다리를 맘대로 못 가누게 되어 일어서지 못하고 꼼짝을 못하는 경험을 하게 되는데(이런 증상이 나타날땐 어김없이 당이 많이 함유된 ...
3	고혈압은 원래 있었고,병원에 가서 당뇨 수치를 봤더니 360~370이 나왔데요 (엄청높다능) 그래서 걱정이 되서요 저는 아빠 없으면 못살아요 (죽는병은 아니지만) 그래서 아빠가 식단을 채소 위주로 바꾸고,운동을 하기로 했어요 ...

그림 1: OWLNEST® SE-Mainframe 을 이용한 핵심 검색 기능 예시 (컨텍스트 패러프레이즈 검색)

OWLNEST® SE-Mainframe

Proactive discovery and insights from contexts

OWLNEST Corp.
주소 : 대전광역시 유성구 문지로 193 KAIST ICC 학부동 F736 (305-732)
문의 : contact@owl-nest.com, +82 2-742-3021

Technology Application Cases

빅데이터 분석 플랫폼 삼성생명 SAMSUNG

OWLNEST® SE-Mainframe 의 토큰별 벡터화 기능과 통계정보를 반영할 경우, 자연어 검색과 연관어 검색이 가능합니다. 일반검색의 경우 질의와 텍스트가 일치할 경우에만 검색이 이루어지지만, 자연어 검색을 통해 동의미(유의미)-이형태 문자열로 구성된 텍스트 데이터를 검색할 수 있습니다. 문서/문단/문서 단위의 검색이 가능해짐으로써, 연 100GB에 달하는 삼성생명 BDA 콜센터 녹취콜로부터 고객 VoC의 경향을 빠르게 파악할 수 있습니다.

여론의 경향성 파악

mk 매일경제 mbn

OWLNEST® SE-Mainframe 의 CorrNLP Interface 를 이용하여, 자동 토픽 분석을 연동할 경우, 기사 및 SNS 등으로 대표되는 텍스트 데이터에서 동적으로 변화하는 토픽을 추출할 수 있습니다. 이를 통해서 전체적인 여론의 경향성을 파악할 수 있고, 특정 토픽의 변화양상 또한 손쉽게 인지할 수 있습니다. 이와 같은 자료는 매일경제신문의 정치부에서 정치에 대한 대중의 인식과 정치인의 발언에 관련된 양상을 분석하여 차별화된 분석 프레임을 확보하는 기회를 제공합니다.

위험징후 조기 경보

국립재난안전연구원 NATIONAL DISASTER MANAGEMENT INSTITUTE

OWLNEST® SE-Mainframe의 ML/TM, IE Interface 를 통해 재난 이슈의 발생을 조기 감지할 수 있습니다. 기사나 SNS에서 특정 사안이 언급된 빈도와 대응 시나리오에 맞는 문서를 손쉽게 자동 추출할 수 있습니다. 더 나아가서 특정 사건과 동시에 출현하는 단어, 개념들을 통계적으로 검증하는 연관성 분석을 통해 특정 재난 이슈에 대한 팩트를 객관적으로 프로파일링할 수 있습니다. 이는 재난 이슈의 자동 감지와 이슈를 자동 트래킹하는 객관적인 조기 경보 시스템이 될 수 있습니다. 또한 사회적 이슈와 관련된 여론의 이상징후를 조기에 탐지할 수 있습니다.

Main Components

FileSystem / InMemory Indexer

고객의 목소리를 담은 VOC 문서, 사실 정보를 전하는 신문기사, 온라인 소셜 네트워크 서비스의 글 등에 대한 텍스트 분석을 위해서는 담화 (Discourse) 수준의 접근이 필요합니다. 따라서 문장/분단 별 단위로 자동 정규처리를 하는 것은 정확한 분석을 통해 올바른 의사결정을 도출하기 위한 과정입니다. 각 문서의 분할단위 처리가 가능해짐으로써 다양한 관점의 의견을 분석하고 상세 특징을 추출할 수 있습니다. 언어적 패턴 인식 (Linguistic Pattern Recognition) 을 통해, 정확한 경계 인식과 각 문장/문단 별 아이디가 발급됩니다.

Automatic Word Spacer

SNS 문서나 VOC STT (Speech-to-text) 문서와 같이 띄어쓰기 교정이 필요한 유형의 데이터 처리에 사용시 고품질의 정규화 텍스트 문서를 얻을 수 있는 한국어 자동 띄어쓰기 기능입니다. Conditional Random Fields (CRFs) 기계학습 기법을 이용해 n-gram 토큰 간 확률을 이용하여 단어간 띄어쓰기 여부를 결정함으로써 다양한 산업 현장의 텍스트 처리 문제에 적용할 수 있는 견고한 성능과 빠른 처리속도를 제공합니다.

Multilingual Query Handler

한국어 비정형 텍스트 데이터 처리를 위해서는 조사, 어미, 활용형 등을 처리하기 위해서 형태소를 추출하는 과정이 필수적입니다. 형태소 분석은 각 키워드를 실제 문서상의 다양한 표현형으로부터 사전상의 등록형으로 일관성 있게 처리하여 추출하는 과정입니다. 이를 통해, 텍스트 분석에 필요한 키워드의 최대 재현율을 확보할 수 있습니다. 형태소 단위로 키워드를 관리함으로써 검색엔진, 개체명 인식, 구문분석과 같은 고급 자연어처리 기능에 모두 활용할 수 있는 데이터베이스를 확보할 수 있습니다. 문자열간 2-dimensional tabular parsing 과 기분석 패턴을 통해 초고속 준실시간 처리가 가능합니다.

Searcher - Optimized Ranking

개체명이란 인물, 장소, 시간, 기관 등과 같은 일반 개체명과 상품명, 직책명, 부서명 등과 같이 사용자 요구정의가 필요한 특수 개체명, 그리고 다중 개체명 분류를 지원하는 확장 개체명으로 나뉘어집니다. SE-Mainframe 은 일반 개체명의 경우, 기본 기능으로 제공되고, 특수 개체명의 경우 사용자 정의를 지원합니다. 확장 개체명은 일반/특수 개체명에 추가적으로 부여됨으로써, 일종의 개체명간 상하위 관계를 통한 택소노미 구조를 반영한 개체명 정보가 부착된 단어 단위의 추출이 가능합니다. 형태소간 통계적 연접정보를 반영한 Conditional Random Fields (CRFs) 기계학습 기법과 200만개에 달하는 대규모의 기반사전을 통해 개체명 인식의 성능을 보장합니다.

Controller Interafce for Search Cluster

의존 구문분석이란 특정 개체명과 표현단위의 관계를 파악하여 상품과 의견 간의 연관성 추출 분석 등에 활용될 수 있는 도구입니다. 특정 제품에 대한 소비자 반응을 다양하게 파악하고자 할 경우, 개체명 인식과 의존 구문분석을 활용하여 빠르고 정확한 텍스트 분석의 관점을 확보할 수 있습니다. 의존 구문분석은 형태소간 결합규칙과 구문청크 단위의 다층 의존관계를 반영할 수 있도록 Multilayered CRFs 기계학습 기법을 통한 견고한 [illegible] 합니다.

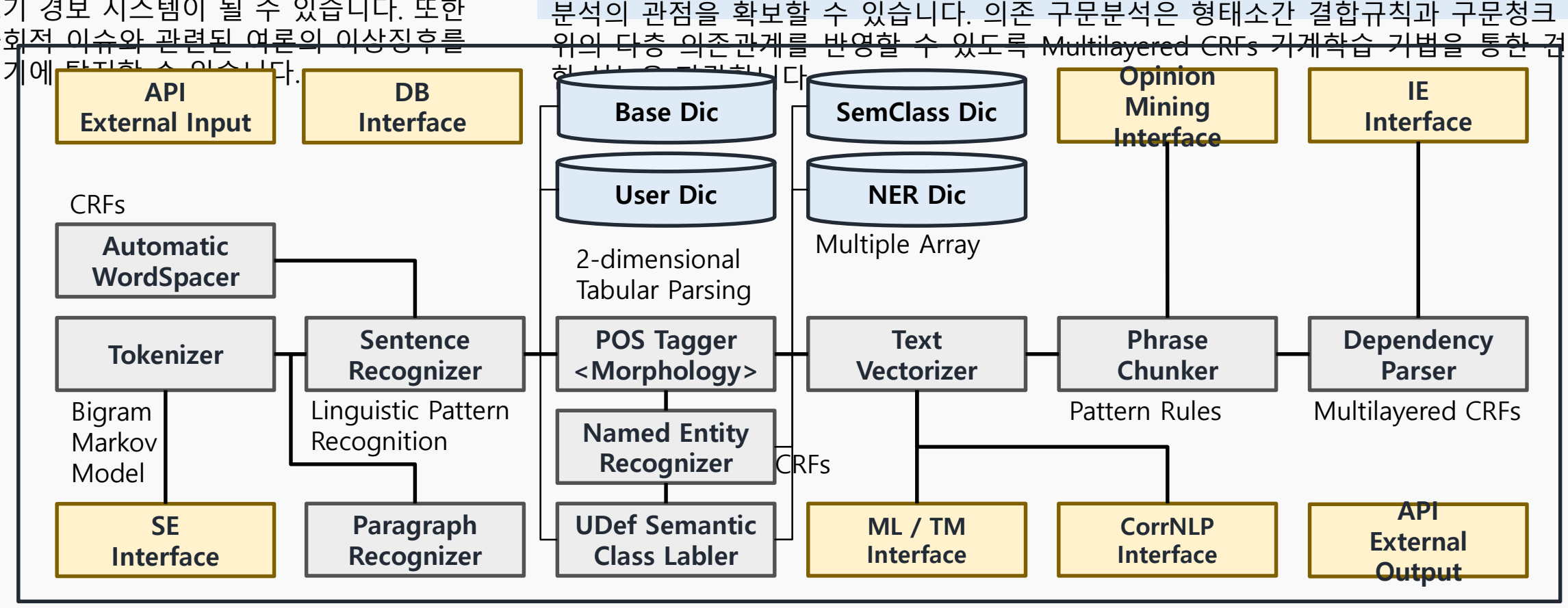

그림 2: OWLNEST® SE-Mainframe, Core Architecture