OWLNEST® NLP-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

OWLNEST® NLP-Mainframe System Requirements

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

향상된 언어처리 기술을 탑재한 텍스트 마이닝 메인프레임 솔루션으로써, 비정형 텍스트 문서에 존재하는 문맥 정보를 빠르게 찾을 수 있습니다.

주제와 내용을 탐지하고, 특정 용어와 부합하는 명시적 분류와 개체간의 관계를 파악
사용자 입장에서 주도적으로 수행할 수 있도록 자동 처리

What?

향상된 언어처리 기술

NLP-Mainframe

Benefits

1 *집약된 자연어처리 기술과 알고리즘*

의사결정 시간의 절약

2 *문자열 토큰 단위 통합적 언어 처리 기능*

고수준 메타정보의 통합 제공

3 *데이터 전체 집합의 구조를 간략히 파악*

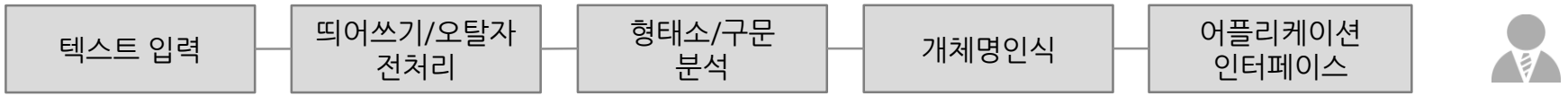
비즈니스 기회로 집중

Details

NLP기술 집약형 SW	메타정보의 통합	비즈니스적 가치
• *키워드 문법표지 부착, 개체명 인식, 구문단위 추출, 자동 띄어쓰기, 문장단위 분할*	• *개별 메타정보에 대한 기능단위 분할처리기능 제공*	• *단어 인덱싱을 지원, 이를 벡터화된 기본 정보 제공*
• *의존관계 추출 등의 고수준 기능을 자동화하여 효과적으로 제공*	• *고수준 메타정보의 통합 제공을 통해, 텍스트 분석가가 필요로 하는 기능을 편리하게 이용*	• *데이터 마이닝과 통계 모델링 기법을 적용해 인사이트를 발굴*
• *사용자는 분석업무에 집중, 가치발견의 가능성 향상*	• *웹스케일 수준의 관련 데이터에 실시간으로 접근하여 적절한 정보를 찾아내어 분석*	• *비정형 텍스트 데이터로부터 고객의 니즈에 귀 기울이고, 서비스와 제품의 요구사항을 파악*

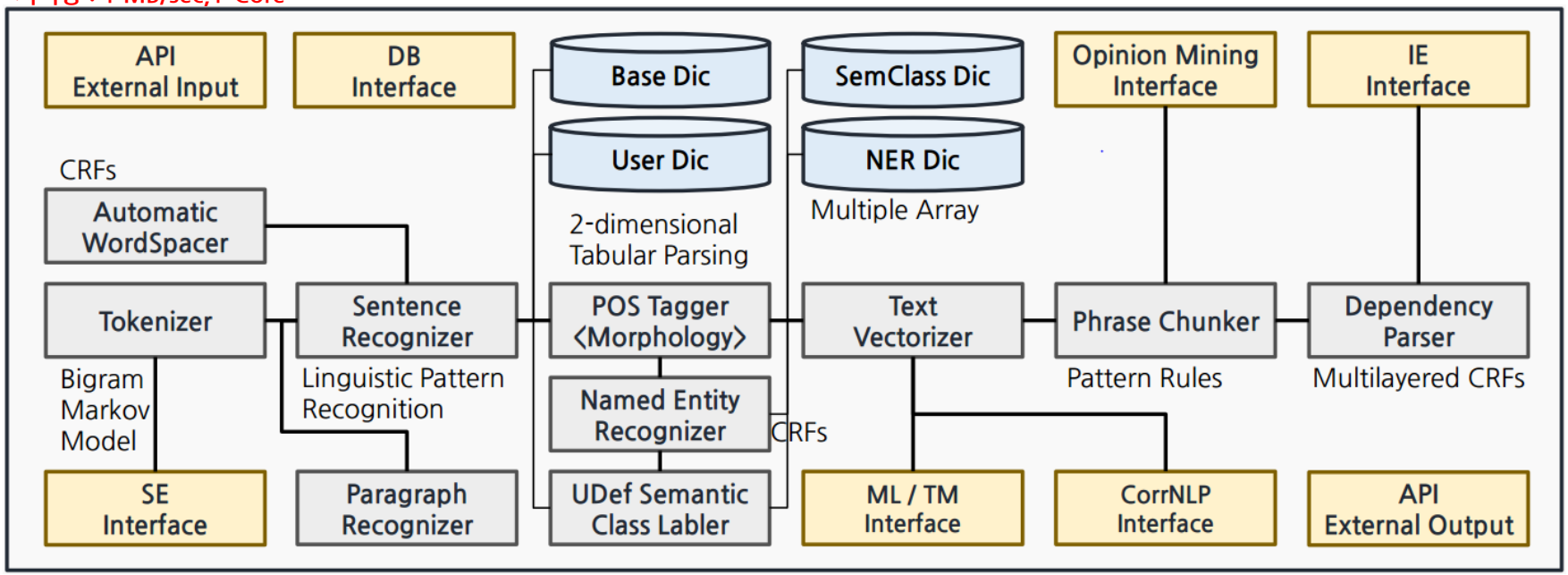
텍스트 데이터 전처리를 위한 자연어처리 통합 솔루션 NLP-Mainframe 의 기능을 기반으로 견고한 텍스트 마이닝 아키텍처를 제공합니다.

자동 띄어쓰기, 오탈자교정, 문장/문단경계 인식, 형태소분석, 구문분석, 개체명인식
인터페이스 : 검색, 연관어, 기계학습, 감성분석



텍스트 데이터 전처리 과정

처리량 : 1 MB/sec,1-Core



NLP-Mainframe 의 주요 구성요소와 처리 흐름은 아래와 같습니다.

문장/문단경계 인식 모듈, 자동 띄어쓰기 모듈
형태소 분석 모듈, 개체명 인식 모듈, 의존 구문분석 모듈

01 패턴 인식을 통해, 경계 인식과 각 문장/문단 별 아이디 발급

02 2-dimensional tabular parsing 기분석 패턴 기반 고속 처리

03 형태소간 통계적 연접정보를 기계학습 기법 / 대규모 기반사전

04 결합규칙과 구문청크 단위의 다층 의존관계를 반영, Multilayered CRFs 기계학습 기법

05 관련 기능들에 대한 관리 프로세스 제공

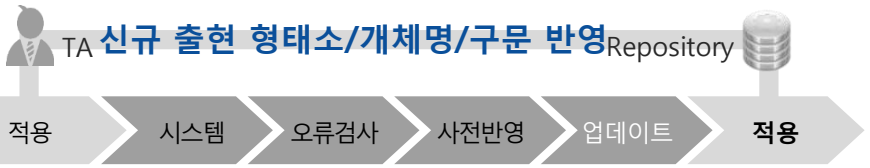
06 관리자 UI를 통한 시스템 실행 및 학습상태 조회 기능

07 Plug/Play, 사전별 관리

NLP-Mainframe 솔루션 도입 후 품질관리 체계는 아래와 같습니다.

자연어처리 솔루션 도입 후 전처리 품질을 효과적으로 유지하기 위해, 기술지원을 실시하여 신규 출현 용어에 대한 형태소, 개체명, 구문을 반영하여 용어사전을 강화하고 솔루션의 품질을 향상시킵니다.

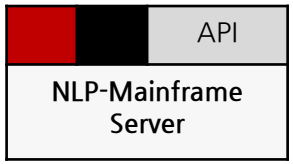
자연어처리 솔루션 품질운영 프로세스

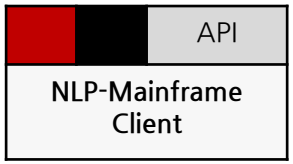


1. - STT 데이터 신규 출현 용어의 정기적 학습 Engine 반영
2. - CRFs 학습 기반 형태소, 개체명, 구문 일원화 분석
3. - 엔진 업데이트 후 기준데이터 분석 품질 테스트로 품질관리
   - 학습모델 품질 측정
4. - API와 연계로그 관리
   - 실시간 이슈 반영
5. - 엔진 업데이트를 위한 오분석건, 신규용어 등을 수집

NLP-Mainframe 의 주요 특징은 아래와 같습니다.

- *시스템을 위한 서버 설치형 소프트웨어*
- *서버의 하드웨어 성능을 극대화할 수 있는 기업형 솔루션*





- *개발자를 위한 클라이언트 API*
- *각 컴포넌트 모듈에 대한 외부 제어 가능형 사용자 솔루션*

다중형식 처리

• 다양한 형식의 작업 문서, 입력파일 자동변환 기능

취급 파일 (pdf, doc, ppt, hwp, xls, txt) 의 통합 저장, 메타정보 관리

csv, json, xml 등의 파일출력 및 DB 인터페이스 지원

↓

통합 자료 구조

통합 인터페이스

• 컴포넌트 모듈별 실행 및 상태 조회를 위한 사용자 인터페이스

컴포넌트 통합형 입출력 제어구조로 자료 관리의 편리성

사용자에게 익숙한 엑셀 테이블 형태의 자료 관리 구조

↓

솔루션 제어 기능

메타정보 분석

• 단어/구문 수준의 정보를 자동 분석하여 메타정보화

중요 키워드에 대한 일반/특수/확장 개체명 정보 자동 인식

개체명-표현 간의 의존관계의 자동 추출을 통한 메타정보화

↓

구조적 정보

유니코드 인코딩

• 다국어 및 기호 문자열의 처리를 위한 최적의 인코딩

유니코드 처리를 통한 텍스트 데이터 저장/관리

언어별 텍스트 인코딩 통합 관리/저장/추출

↓

최적 인코딩

웹스케일 처리

• 대용량 파일의 처리를 위한 멀티 프로세싱 기능 지원

웹스케일의 입력 텍스트 처리를 위한 고속 병렬 문자열 토큰화 기능

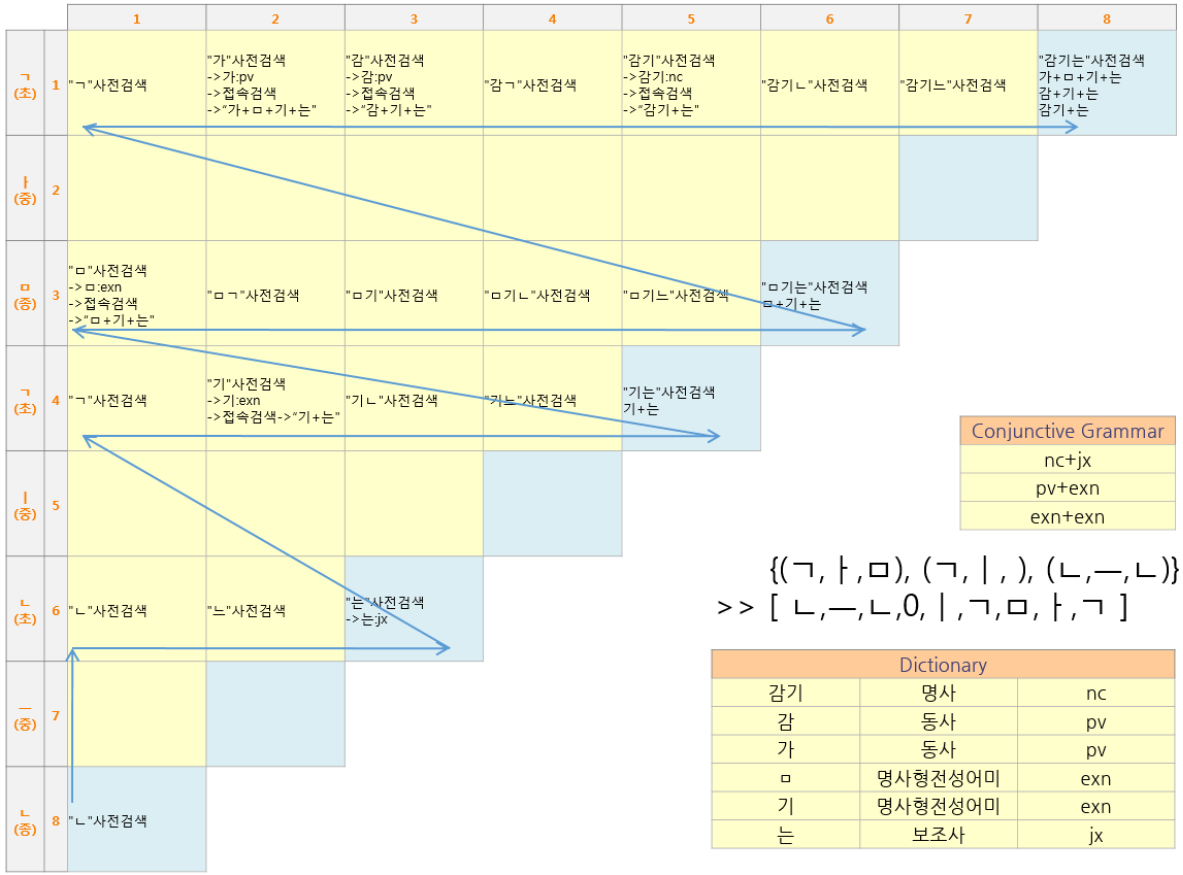
토큰별 인덱스 생성과 메타정보의 연결을 통한 빠른 정보처리

↓

고속 병렬 처리

NLP-Mainframe 은 한국어에 최적화된 형태소 분석을 위해 개발된 2-dimensional Tabular Parsing 알고리즘, 접속 패턴을 활용한 트리구조가 적용되어 있습니다.

Tabular Parsing 알고리즘 상세



Conjunctive Grammar
nc+jx
pv+exn
exn+exn

{(ㄱ, ㅏ, ㅁ), (ㄱ, ㅣ,), (ㄴ, ㅡ, ㄴ)}
>> [ㄴ, ㅡ, ㄴ, 0, ㅣ, ㄱ, ㅁ, ㅏ, ㄱ]

Dictionary		
감기	명사	nc
감	동사	pv
가	동사	pv
ㅁ	명사형전성어미	exn
기	명사형전성어미	exn
는	보조사	jx

주요 특징 / 내용

1 2-dimensional Tabular Parsing

- Nondeterministic push-down automata
- 한국어의 특성을 고려한 Right-to-Left 알고리즘
- 인식된 입력 부분 문자열과 기저장/기분석된 데이터를 비교

2 형태소간 통계적 연접정보 기계학습 및 대규모 기반사전

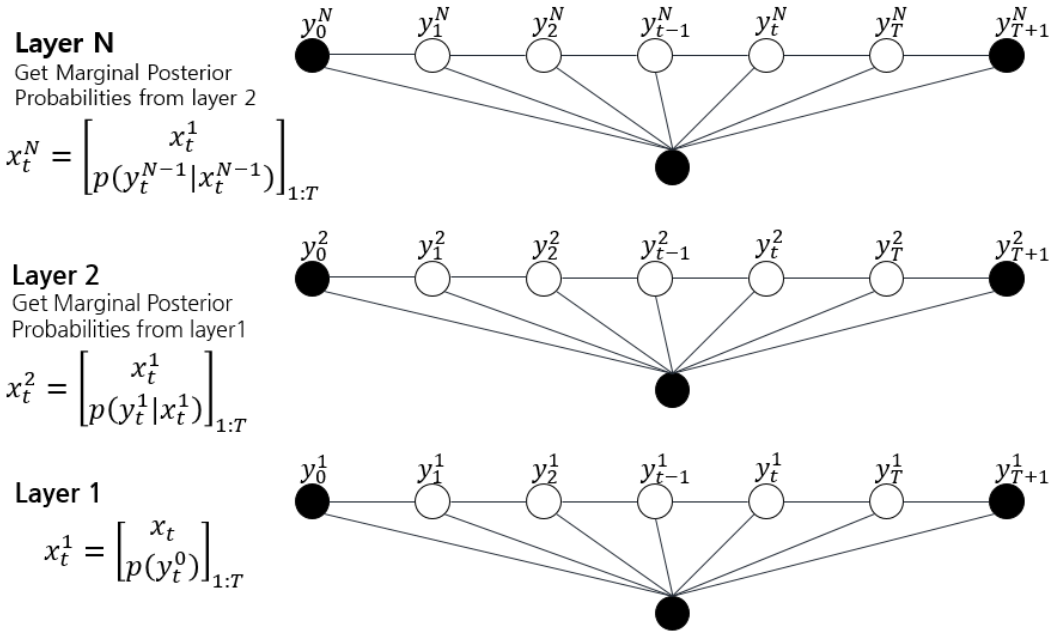
- LR Grammar 를 통한 연접정보 기분석
- 검증된 말뭉치로부터 통계적 연접정보를 추출하여 기계학습
- 형태소/내용어/기능어/개체명 별 기반사전

3 접속 패턴 기반 고속 처리

- 접속 Data Index 를 이용한 고속 문자열 처리
- CPU/RAM 사용량 최소화를 통한 시스템 성능 극대화

NLP-Mainframe 은 자동 띄어쓰기 및 의존구문분석을 위해 연속형 데이터에 대한 알고리즘인 Multilayered CRFs 기계학습이 적용되어 있습니다. 이를 통해 Disambiguation 문제를 해소합니다.

Multilayered CRFs 알고리즘 상세

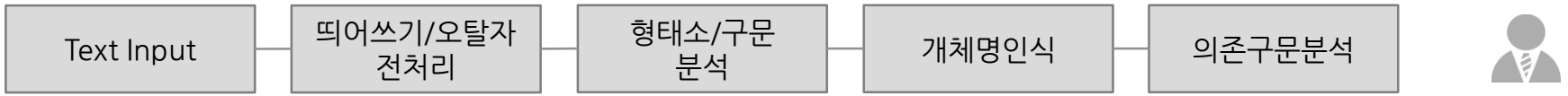


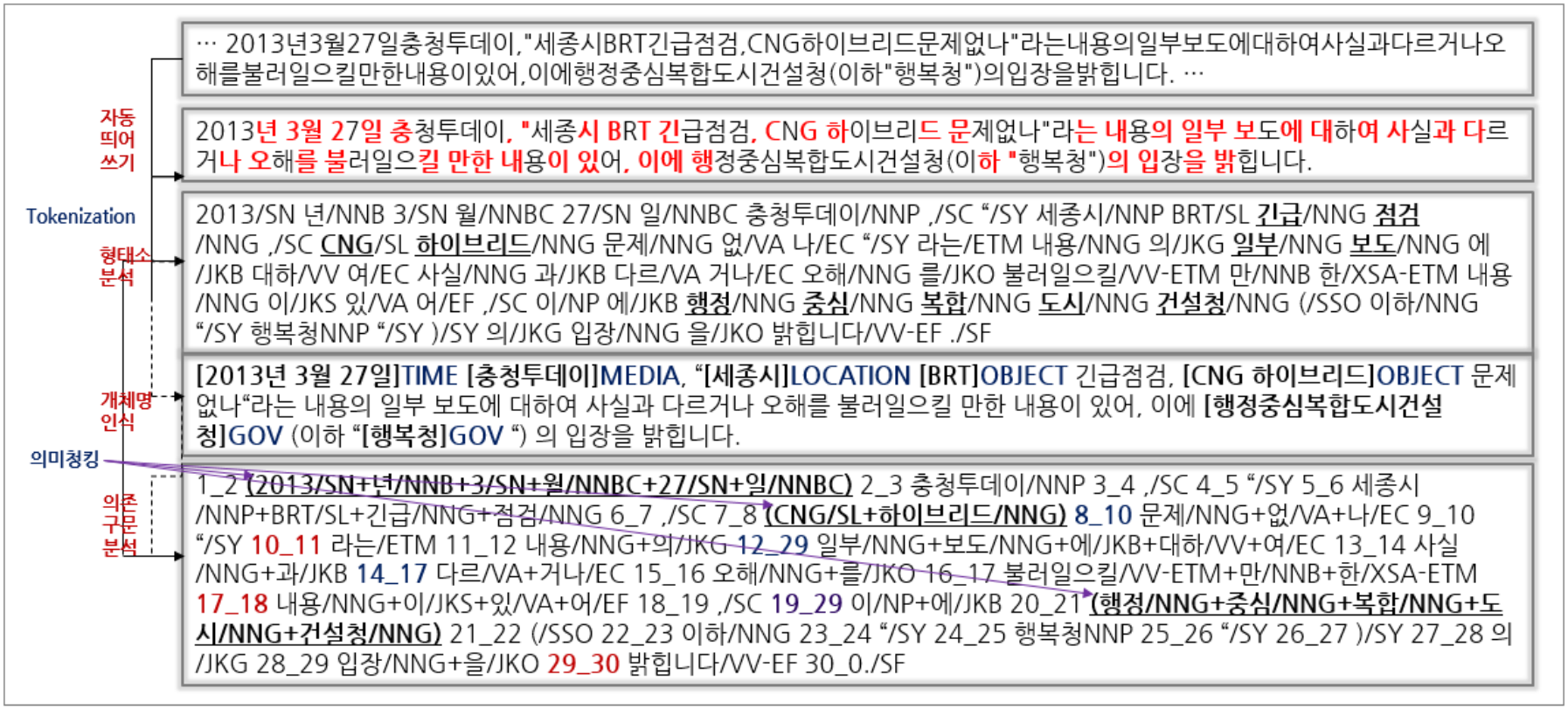
의존구문분석 예시

1_2	부대/nc+내부/nc+는/jx
2_3	물론/a
3_4	자동차/nc
4_5	한/mn
5_6	대/nbu+만/jx
6_7	지나가/pv+아도/ecs
7_8	먼지/nc+가/jc
8_9	일/pv+어/ecs
9_10	앞/nc+이/jc
10_11	안/a+보이/pv+는/exm
11_12	건조/ncs+하/xpa+ㄴ/exm
12_13	비포장도로/nc+를/jc
13_14	10/nnn+km/f+나/jj
14_15	비/nc+로/jca
15_16	쓸/pv+라는/exm
16_17	윗사람/nc+의/jcm
17_18	지시/nc+가/jc
18_19	내려오/pv+ㄴ/exm+것/nb
19_0	이/jcp+었/efp+다/ef+./s.

NLP-Mainframe 은 언어처리 및 분석 모델링 도구의 집합체로서 텍스트 집합의 문맥적 의미를 발견하고 가치있는 정보를 추출합니다.

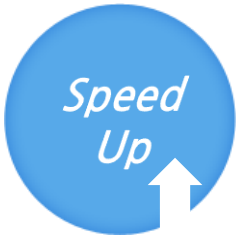
단어와 구문 수준, 메타정보와 표현 수준에서 통계적 특징과 패턴에 따른 명시적 정보를 자동처리





NLP-Mainframe 은 기능유형으로 구분된 각 사전유형을 모듈구조로 관리합니다.

모듈형 사전구조의 효과



도입 시스템의 안정화 주기 단축
빠른 시스템 안정화

손쉬운 사전관리 및 유지보수
유지보수 효율성 향상

요구사항에 따라 사전 재조합
사전생성 효율화

일원화 구조 모듈의 위험요소
오류등록 Risk 최소화

기능유형	사전유형								사전갯수
내용어	일반명사	고유명사	대명사	신조어	한자어	복합명사	개체명	차용어	12
기능어	부사	조사	어미	어근	접미사	접두사	수사	감탄사	20
용언	동사	형용사	보조용언	관형사	지정사				6
기타	부호	외국어	숫자	한자	사용자정의				5
총계									41

NLP-Mainframe 의 시스템 도입 시, 손쉬운 시스템 운영 및 유지보수를 위한 관리 프로세스를 제공합니다.

텍스트 입력 → 띄어쓰기/오탈자 전처리 → 형태소/구문 분석 → 개체명인식 → 어플리케이션 인터페이스

신규 어휘 등록 | 데이터 관리 / 컴파일 | 시스템 운영

변경 필요 시

추가어휘 확보 → 추가어휘 등록 → 데이터 추가/수정 → 컴파일 수행 모델 생성 → 상태 저장 시스템 적용 → 운영/ 모니터링

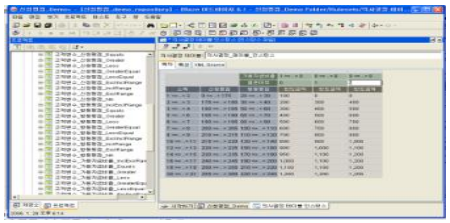
관리 완료 시 → 모델 반영

- 신규데이터 분석을 통해 추가 대상 선정
  - 신규 추가어휘 대상 선정요소
  - 신규 추가어휘 후보 선정
  - 신구 츄가어휘 적용 전략 수립
- 신규 추가어휘 테이블 작성

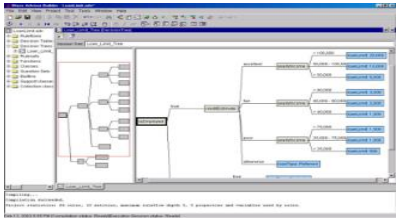
- 데이터 추가/수정
- 사전 등록형태 오류 검증

- 신규 등록 데이터 포함 컴파일 수행
- 신규 모델 생성 후 시스템 적용

추가 어휘 확보



사전구성 테이블



테이블 구조 조회

등록 완료

2012년 이후, 여러 고객사에서 프로젝트의 성공적인 목표 달성을 위해 NLP-Mainframe 을 도입했습니다.

프로젝트 명	발주처	기간	역할	형태
삼성전자 DS부문 정보보호 관리업무 자동화 컨설팅	SDS	'17.10. ~ '18.01.	분석 컨설팅, 솔루션 납품	분석 컨설팅
글로벌 재난안전 리스크 정보 탐색 및 모니터링 기술 개발	국립재난안전연구원	'17.06. ~ '17.12.	재난정보 분석 시스템 개발	시스템 개발
삼성페이 서비스 고도화를 위한 분석 시스템	삼성전자	'16.03. ~ '16.06.	분석 컨설팅, 솔루션 납품	분석 컨설팅
재난안전 맞춤형 텍스트 탐색 및 실시간 분석 UI 개발	국립재난안전연구원	'16.06. ~ '16.12.	재난정보 분석 시스템 개발	시스템 개발
삼성생명 컨설팅 영업지원 시스템	삼성생명	'15. 6. ~ '15.12.	머신러닝 시스템 개발	컨소시엄 개발
삼성생명 BDA 경영자원화 구축	삼성생명	'15.06. ~ '15.12.	솔루션 납품, 모듈 커스터마이징	컨소시엄 개발
텍스트마이닝을 이용한 교통정보체계 감사토픽 분석	감사원	'15.02. ~ '15.04.	시스템 개발	Term License
미래위험 변화예측을 위한 사회환경탐색기술 개발	국립재난안전연구원	'14.05. ~ '14.08.	텍스트마이닝, 토픽분석	시스템 개발
리얼상품평 프로젝트 언어처리 모듈 구축	GS홈쇼핑	'12.02. ~ '12.08.	리뷰 자동 분석 시스템 개발	솔루션 도입