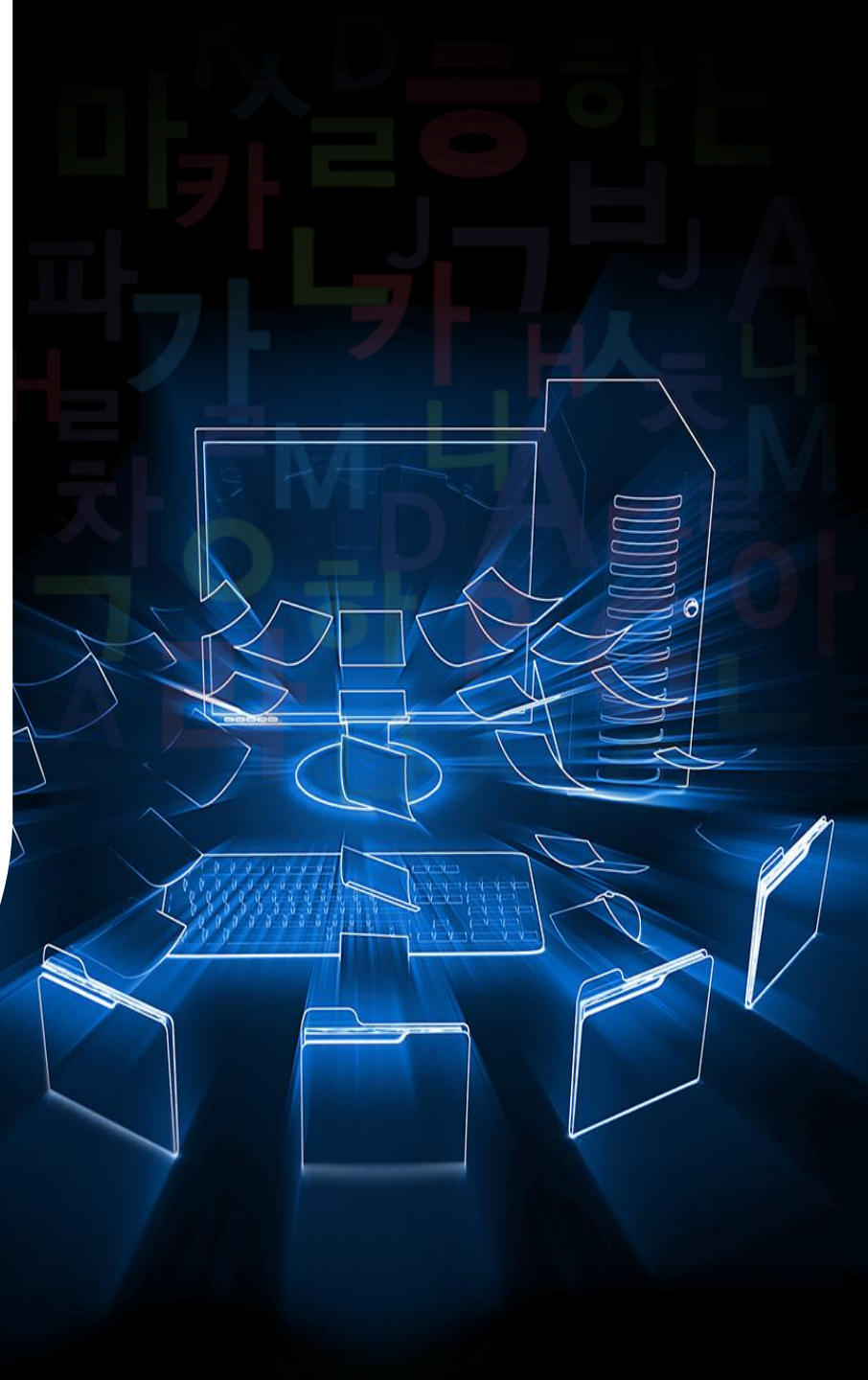
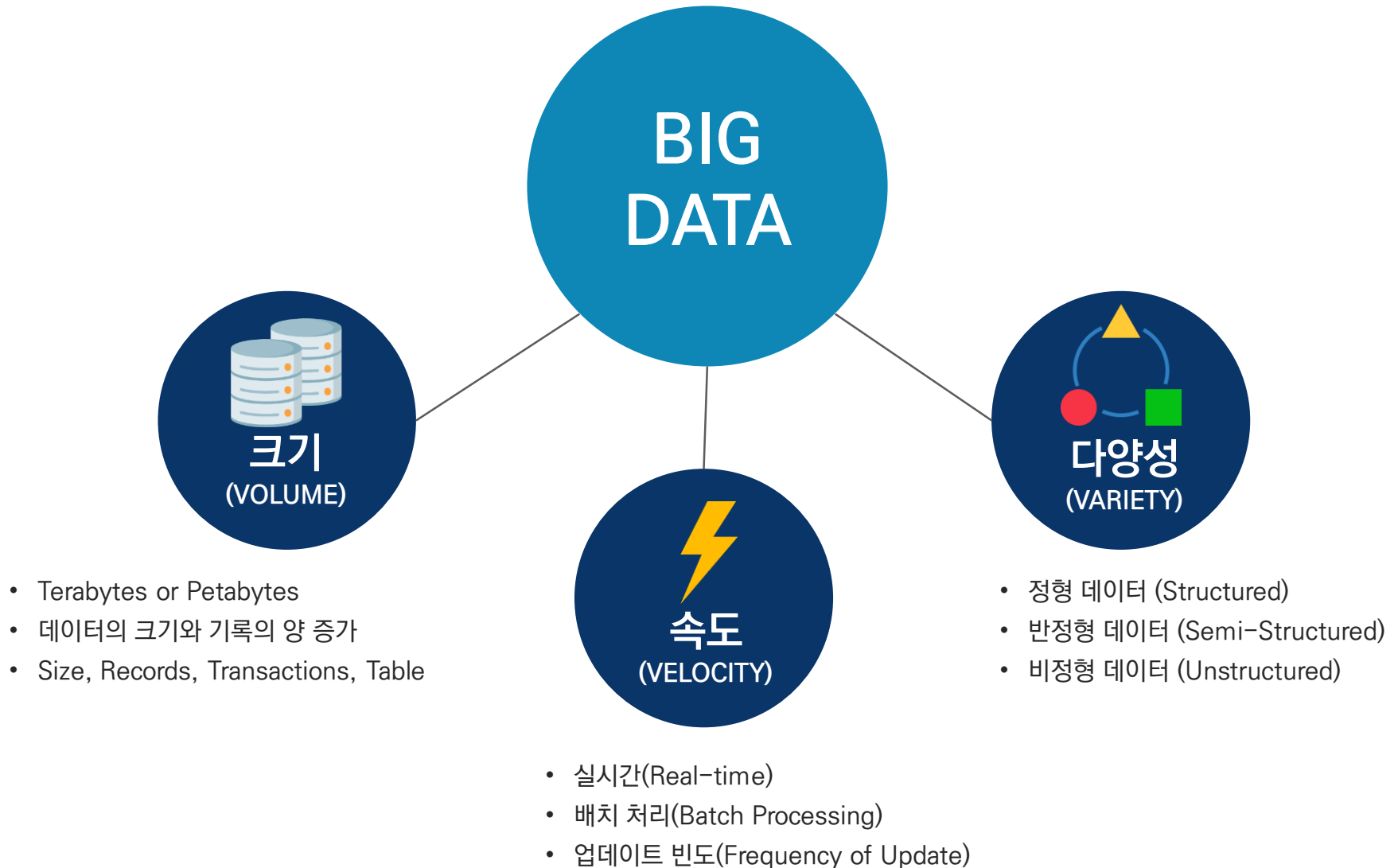


## No.1 빅데이터 분류 분석 솔루션

smart **TA**



# 4차 산업 핵심 '빅데이터'의 부상



# 다양한 형태의 '빅데이터'

## 정형

데이터를 다루는 사람이라면  
흔하게 보게 되는 형식의 데이터

: 관계형 데이터베이스, 스프레드시트,  
CSV 등

## 비정형

형태가 없으며,  
연산 또한 불가능한 데이터

: 계약서, 보고서, 법률문서,  
소셜데이터, 의료기록지 등

## 반정형

연산이 불가능한 형태의 데이터

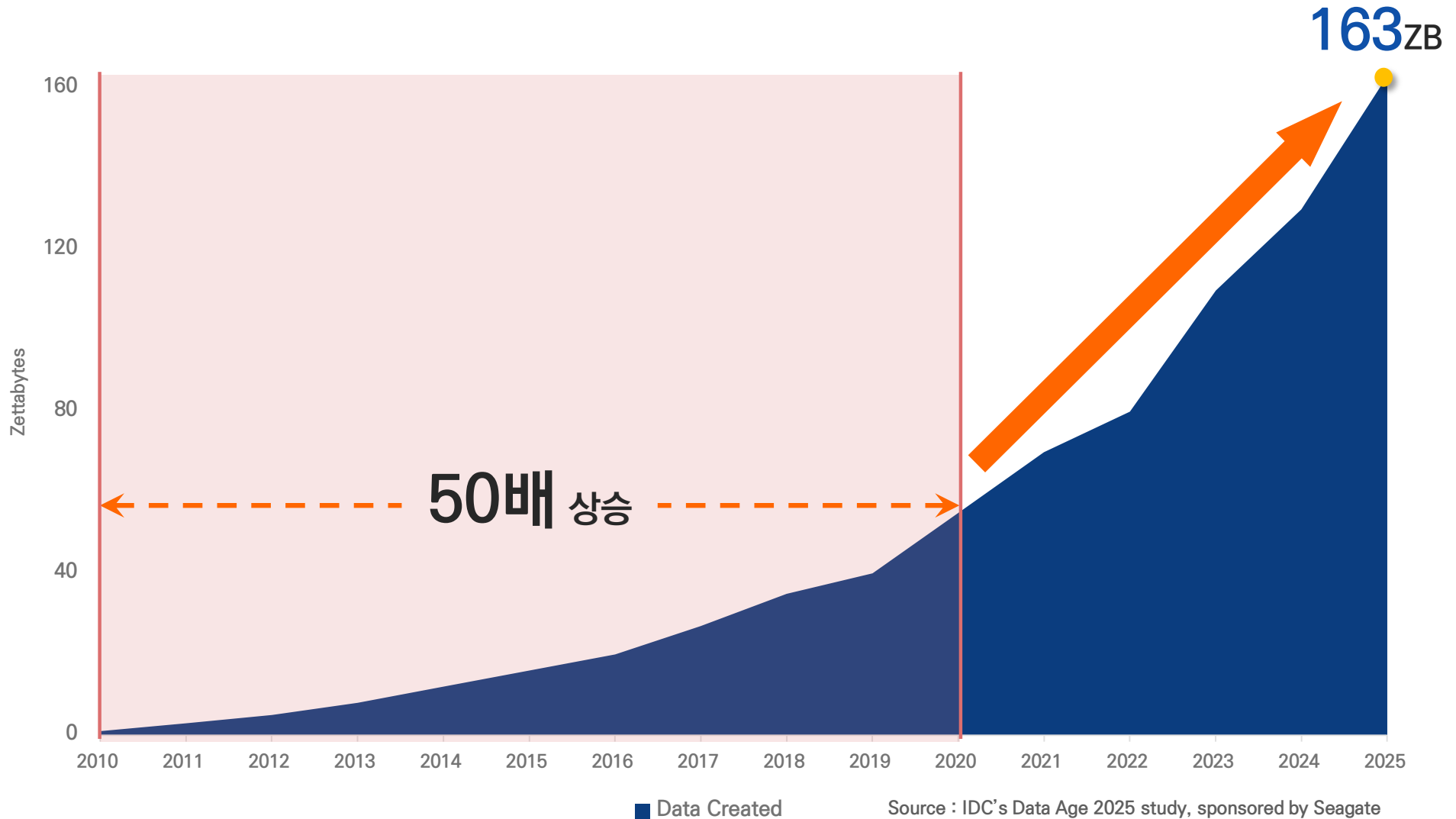
: XML, HTML, JSON, 로그 등

과일	원산지	주문번호
포도	안성	200869866
사과	영주	201986332
귤	제주	200154135
딸기	고령	201463968
복숭아	장호원	200454533



```
"title": "The JSON example",  
"descriptionText": "This is some title text",  
},  
"content": {  
  "title": "The content example text",  
  "elements": [  
    {  
      "title": "The first element",  
      "mainText": "First element main text",  
      "additionalText": "First element additional text",  
    },  
    {  
      "title": "The second element",  
      "mainText": "Second element main text",  
      "additionalText": "Second element additional text",  
    }  
  ]  
}
```

# 급증하는 텍스트 형태의 비정형 데이터



# 텍스트 형태의 비정형 데이터, 어떤 가치를 가지고 있는가?

## 이미 알고 있는 체계에 의한 정보

1. 새로 유입된 문서는 기준이 되는 항목에 얼마나 부합하는지?
2. 키워드가 다른 유사한 내용의 질문을 같은 의미로 분석하여 활용할 수 있는지?

## 이전에 몰랐던 새로운 정보의 발굴

1. 고객이 어떤 것에 관심을 갖고 어떤 것에 불만을 갖고 있는지?
2. 전문가의 소견에서 질병 및 결점 발생 확률이 어느정도 되는지?

# 비정형 텍스트 데이터, 어떻게 관리하고 계신가요?

매일 쌓여만 가는 텍스트 데이터 관리...  
무한한 잠재적 가치를 지닌 데이터를 제대로 활용하지 못하고 계시진 않으신가요?



# 이러한 비정형 텍스트 형태의 데이터를 수작업으로 분류한다면?

06 인사이트 추출

데이터 분류와 분석 단계에서  
가장 많은 시간 소요  
※ 계약서 검토 기준 약 2주

05 데이터 분석

04 데이터 분류

03 데이터 가공

02 데이터 취합

01 데이터 수집

데이터 분석 처리 과정  
6단계

텍스트 형태의 데이터를 분류 및 분석해주는  
솔루션의 역할이 매우 중요!!

# 그렇다면, 솔루션을 통해서 어떤 가치를 창출하고 싶으신가요?

마케터

SNS상의 다양한 글에서  
제가 원하는 유형의 글만 찾아서  
분류하고 속성을 분석  
하고 싶어요



계약서 검토 담당자

계약서의 조항이  
너무 많아요... 불리한 조항만  
추출해서 보고 싶어요



의료진

환자들의 기록지를 정형화 해서  
의료 연구에 활용하고 싶어요



고객센터 관리자

고객센터의 QA를 통해  
주요 이슈 사항을  
파악하고 싶어요



공무원

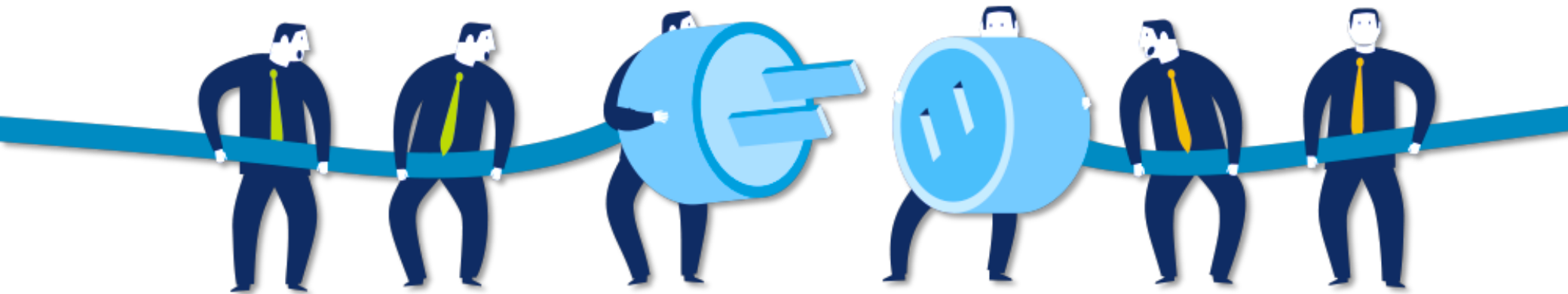
우리 지역 상권 활성화에 대한  
정책을 마련하기 위해 사람들의  
인식을 분석하고 싶어요





# SMART TA는 이 모든 것이 가능합니다!

데이터가 원유인 4차 산업 시대에  
가치 있는 정보와 사용자를 연결하는  
텍스트 분류 솔루션



# SMART TA의 4가지 차별 포인트

## 속성 분류

사용자가 지정한 속성에 대한 분류가 가능하게 합니다.



## 사전 및 패턴 적용

분석 대상의 자원을 기반으로 사전화 및 패턴 적용



## 자동 예측

어순의 유형을 자동으로 예측합니다.



## 원인 파악

오 분류된 문서의 원인 파악이 가능합니다.



# 일반적인 텍스트 분석 솔루션은...

삼성 청소기로 갈아 탔어요~ 무게는 좀 무겁지만 흡입력은 만족하고 있어요.

가격이 사악한 것 빼고는 전반적으로 만족합니다.

청정 스테이션이랑 물걸레 기능은 정말 편하네요!



일반적인 텍스트 분류 결과

다중 속성 분류 불가능

‘긍정’ 글로 분류

# 그러나 SMART TA는 다릅니다.

삼성 청소기로 갈아 탔어요~ 무게는 좀 무겁지만 흡입력은 만족하고 있어요.

가격이 사악한 것 빼고는 전반적으로 만족합니다.

청정 스테이션이랑 물걸레 기능은 정말 편하네요!



## SMARTTA 텍스트 분류 결과

다중 속성 분류 가능



회사	제품	속성	감성
삼성	청소기	무게	부정
삼성	청소기	흡입력	긍정
삼성	청소기	가격	부정
삼성	청소기	청정스테이션	긍정
삼성	청소기	물걸레	긍정

# SMART TA를 통해 잠자고 있는 데이터를 분석하여 새로운 가치를 창출하세요!

## 수익화

기업 내외부의 모든 비정형 텍스트 데이터를  
완벽한 경제적 가치로 전환

## 활성화

데이터 자산을 통합 및 조정 함으로써  
조직 내 모든 시스템에 실행 가능한  
통찰력이 확보 될 수 있도록 분석 극대화

## 품질 향상

현명한 데이터 관리를 위해 데이터를  
메타데이터 기반으로 분류하고 목록화



# 주요기능

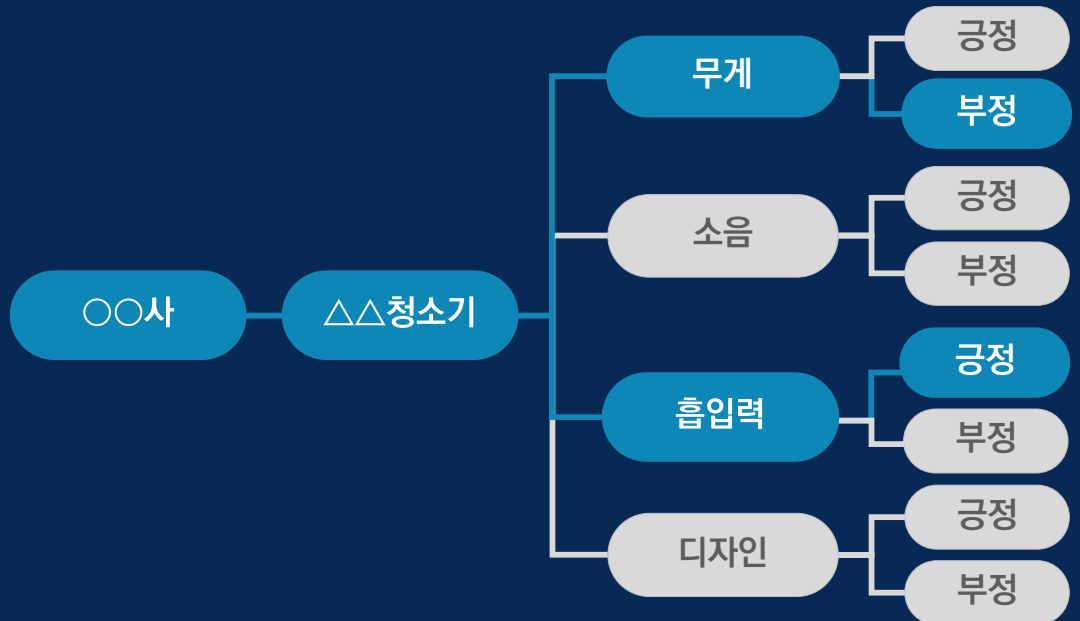
# SMART TA는 사용자가 원하는 분석 방향으로 구성됩니다.



우리 회사 제품을 사용하는 고객들이 어떤 점을 좋아하는지, 어떤 점은 개선이 필요한지 알고 싶은데.. 방법이 없을까?

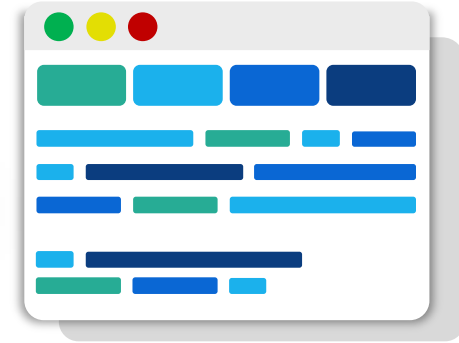


○○회사에 △△청소기로 갈아탔어!  
흡입력 진짜 최고야~  
너무 무거운 것 빼고는 다 만족!!



# SMART TA는 문서 내의 언어 규칙을 기반으로 동작합니다.

명확하게 정의된 언어 규칙을 기반으로 동작하는 시스템



△△전자 이번에 나온 신상 건조기 엄청 보송보송 잘 말라요

△△전자 건조기 성능의 긍정 글

○○사에서 이번에 에어컨 샀는데 AS도 잘 안되고...

○○사 에어컨 서비스의 부정 글

이번에 청소기를 ◇◇사 걸로 바꿨는데, 흡입력 진짜 최고야~

◇◇사 청소기 흡입력의 긍정 글



# SMART TA는 지식을 기반으로 동작합니다.

이번에 우리 신제품에는 디자인에 공을 많이 들였어.  
난 기능 뿐만 아니라 '디자인'에 대해 사람들은  
어떻게 생각할지 궁금해.



우리 회사 청소기의 차별화된 기능은  
우선 '흡입력', '물걸레' 기능이 있어!  
우린 이런 기능들에 대해서 사람들이  
어떻게 생각하고 있는지 궁금해.

“  
사용자들의 지식을 바탕으로  
속성에 대한 기초 데이터 마련  
”



# 자동으로 예측한 언어 규칙을 한 눈에!

**SMART TA**      결과      설정      나가기

어순 자동 예측      머신러닝 파라미터 값 설정       학습 데이터 셋       학습

선행 키워드	결과 키워드	함수	거리/빈도	정확도	적합도	지지도
미소정보기술	텍스트분석	DIST	7	100%	100%	100%
텍스트 데이터	텍스트마이닝	ORDDIST	13	90%	93%	95%

**자동 예측 시각화**

```
graph TD; A[텍스트 데이터] -.-> B[SMART TA]; A -.-> C[텍스트마이닝]; C -.-> B; B -.-> D[텍스트분석]; D -.-> E[미소정보기술]; D -.-> F[솔루션]; E -.-> F;
```

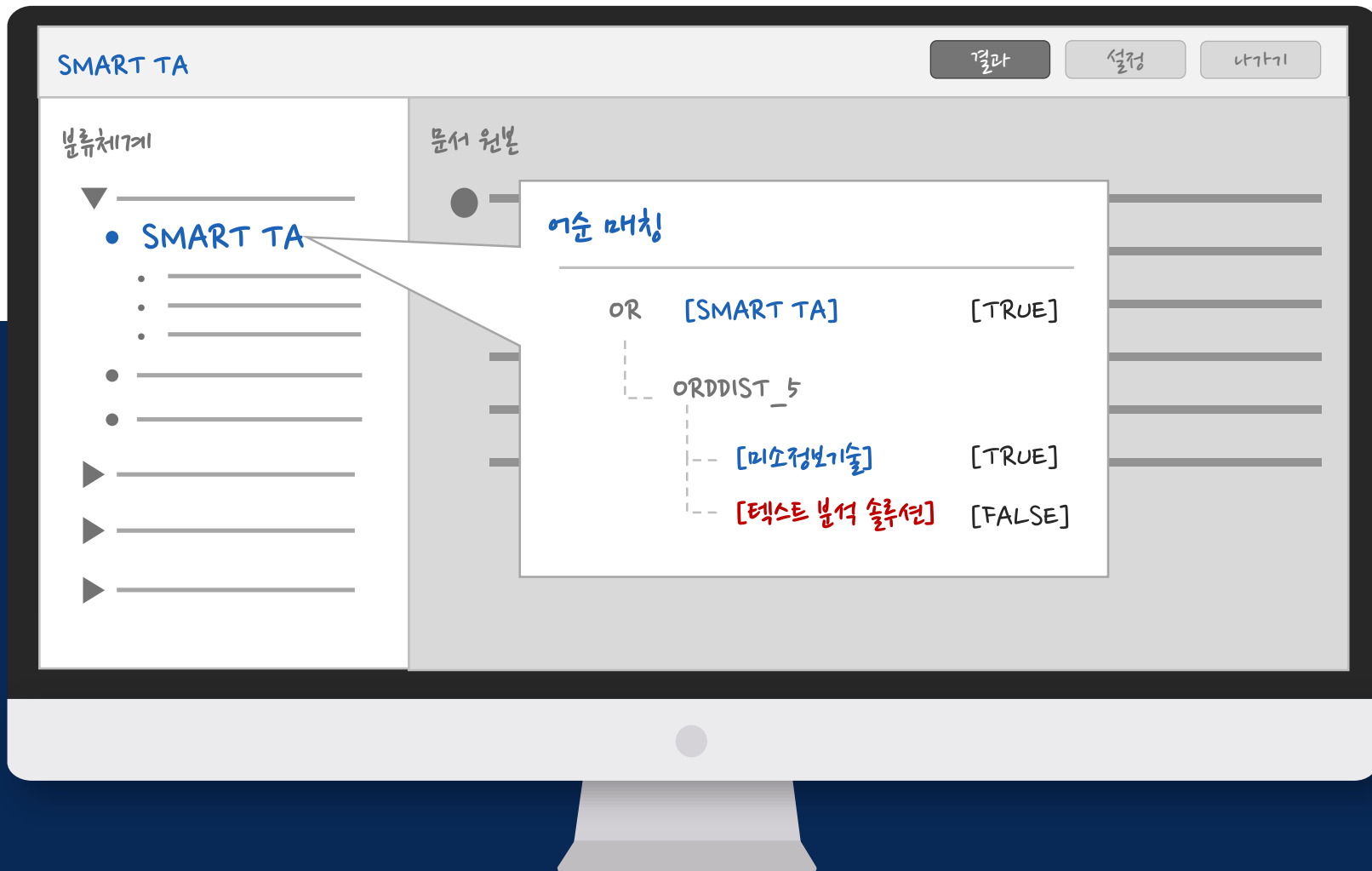
**자동 예측 키워드 보기**

미소정보기술의 대표적인 텍스트 분석 솔루션인 'SMART TA'는 비정형 텍스트 데이터를 이용하여 텍스트마이닝을 통해 유의미한 결과로 재 분류 및 분석하는 솔루션입니다... SMART TA를 이용하여 사용자가 원하는 분류로 맞춤 분석하세요!

# 보다 정확한 분석결과를 한번에!



# 잘못 분류된 문서에 대한 원인 파악 까지!



# 생성된 분류 모델을 서버에 배포하여 분류를 자동화 할 수 있습니다.



# 주요 레퍼런스

# SMART TA는 다양한 비즈니스와 함께합니다.

- **대우건설**

계약서 및 ITB문서 분석

- **통일부**

새터민 진술서 분석

- **SK텔레콤**

VoC 분석

- **한터 글로벌**

소셜 데이터 분석

- **국민건강보험공단 일산병원**

영상의학과 : CCTA

진단검사의학과 : 골수검사, 말초혈액도말검사

병리과 : Kidney Bx

골밀도 검사

- **서울아산병원, 울산대병원**

심장내과 : CAG, Thallium, CCTA

- **서울대병원**

소화기내과 : CFS

- **세종병원**

응급의학과 : X-ray, CT, CAG

- **연세대학교 의료원**

병리과 : 병리 결과지(폐암, 자궁경부암, 유방암, 대장암, 위암)

영상의학과 : 영상 판독지(직장암)

외과 : 병리 결과지(유방암, 위암)

영상 판독지(유방암, 위암)

수술 기록지(유방암, 간이식)

- **아주대학교 의료원**

갑상선내분비외과 : 갑상선암(병리소견, 영상판독지)

소화기내과 : 간암(병리소견)

가정의학과 : 초진 기록지(free text)

인체유전체검사(유방암, 폐암, 대장암, 갑상선암, 위암, 천식, 간암)

- **가이드스타**

소셜 분석

- **수원시**

AI기반 상권분석 맞춤형 정보 시스템

- **AI 바우처 지원 사업**

국세법령정보시스템 질의 응대 시스템

보험상품설명서 확인 RegTech AI 시스템

관세 정보 검색 포털 시스템

- **데이터 바우처**

건강기능식품 통합 데이터 정형화

- **진짜 맛있는 과일**

AI큐레이션 빅데이터 가공 영역

- **해양수산부**

어선안전조업관리시스템

# 자주 묻는 질문





머신러닝 기반의 접근방식을 활용하지 않고,  
왜 언어 규칙 기반으로 동작하나요?

언어규칙방식과 ML접근방식 모두 마찬가지로 언어모델의 부재로 인해 의미 체계의 규칙을 기계가 구별하기 원하는 모든 요소에 대해 프로그래밍 해야합니다.

가장 큰 차이점은 언어 규칙 기반 엔진을 사용하면 언어 규칙에 정의된 내용에 의해 첫번째 시도에서 얻는 정확도가 ML보다 일반적으로 더 높습니다.

또한 정의하는 규칙은 모두 독립적이므로 규칙이 구성되면 기계는 엔진이 지원하는 모든 언어를 즉시 이해하고 의도가 잘못 분류된 경우에 쉽게 변경할 수 있는 '완전 검증 가능 모델'입니다.

**따라서 저희는 SMART TA를 활용하여 완전 검증을 통해  
신뢰도 높은 데이터를 확보합니다**





요즘은 딥러닝 기술을 적용하면  
학습 데이터 없이도 바로 분류된다던데,  
그런 기술을 적용하면 안되나요?

네, 물론 가능합니다.  
딥러닝은 알고리즘을 계층으로 구성하여 자체적으로 배우고 지능적인 결정을  
내릴 수 있는 "인공 신경망"을 만듭니다.  
하지만 딥러닝 알고리즘은 분류는 대개 도메인에 특화된 비즈니스일수록  
학습 모델의 부재로 인해 결과의 정확성 및 신뢰도를 보장할 수 없습니다.



**하지만, 저희는 도메인 전문가로부터 받은 지식을  
바탕으로 신뢰도 높은 데이터만을 선별하여 활용합니다.**



저희는 정리된 키워드나 학습 데이터가 없는데  
그럼 SMART TA를 사용하지 못하는 건가요?

아니오 그렇지 않습니다.  
제대로 된 인공지능 시스템 구축의 첫 발판을 마련하고,  
학습 데이터 생산과 키워드 추출에 소요되는 시간을 절약하기 위해  
저희는 여러 알고리즘(TF-IDF, TextRank 등)을 활용하여 중요 키워드를  
추출 후, 도메인 전문가에게 제공합니다.

이를 통해 추출된 초기 데이터와 키워드를  
도메인 전문가에게 추천하여 컨펌된 데이터를 활용합니다.





그렇다면, 추후에 생성되는 신조어나  
미처 입력하지 못했던 키워드는 사람이  
수작업으로 생성하고 관리해야 하나요?

아니오 그렇지 않습니다.

SMART TA 는 입력한 지식 기반으로만 동작 합니다.

하지만 저희는 자연 언어 처리(NLP) AI의 최신 딥러닝 모델을 사용하여  
기존 SMART TA 에 입력한 사전의 키워드를 기반으로 알고리즘을 통해  
신규 생성되는 키워드 추천을 통해 SMART TA 내의 사전을 고도화 하실 수  
있도록 추천된 키워드 목록 제공합니다.

**이미 도메인 전문가가 알려준 지식을 기반으로  
학습을 하는 방식이기 때문에  
정합성과 신뢰도를 보장할 수 있습니다.**



No.1 빅데이터 분류 분석 솔루션



(주)미소정보기술  
서울시 강남구 테헤란로 34길 21-5 빌딩 S  
우)06222  
Tel. 02. 2205. 0551  
Fax. 02. 2205. 0552  
Email. hanisol@misoinfo.co.kr