



다국어 전자 증거개시제(E-discovery)
글로벌 경제의 법적 분쟁을 위한 주요 3단계



2 머리말

전자 증거개시제(E - Discovery)는 이사회 및 사업 전반에 압박이 되고 있습니다.
국제적 운영 목적 기업은 다국어를 다룰 수 있는 법적 발견전략이 필요성이 대두되고 있습니다.

3 문제의 화두: 모든 관련문서의 작성

다국어 데이터 저장은 새로운 법정 분쟁의 도전을 나타냅니다.
과연 어떻게 기업은 자사가 처한 법정분쟁 및 공방과 관련된 저장된 모든 전자 정보 문서를 찾을 수 있을까요?

5 제 1단계: 프로세싱

전자증거개시(E-Discovery)는 여러 포맷으로 된 잠재된 증거 처리 및 관리로부터 시작됩니다. 언어 식별 및 인코딩 소프트웨어는 다국어 문서를 처리할 수 있도록 처리 작업의 흐름에 통합합니다.

6 제 2단계: 리뷰 및 재검토

잠재성이 있는 증거가 식별 된 경우, 이 자료는 자동적으로 혹은 사람의 손에 매우 세세히 검토 됩니다. 또한 고급 언어 도구는 정확도를 최고로 높여 법적 분쟁 관련 직원들의 시간을 아낄 수 있도록 해줍니다.

7 제 3단계: 분석

엔티티(Entity) 추출은 다국어 전자 문서의 전체의 결합 등의 의미를 부여합니다.

머리말

많은 기업의 글로벌 경제 확장은 법적 위험요소 다뤄야 하는 상황에 직면하고 있습니다. 법정분쟁의 경계가 국경을 넘어선 시점에서 범조인 및 관계자들은 향후 결정적 증거가 되며 또 영문 뿐 아니라 제 3 국의 언어 등의 검토되어야 할 수만가지 문서의 홍수에 직면하고 있습니다. 거의 모든 기업의 수뇌부에서는 법적 분쟁에 관련해 하나의 질문이 있습니다. 그것은, “과연 회사는 언어와 상관없이 민사 또는 형사 소송의 잠재적 피해를 방지하는데 필요한 모든 문서를 생산할 수 있는가?” 하는 것 입니다.

법정에서 필요한 다국어의 검색된 문건을 식별하고, 재 검토하는 능력은 국제적 사업을 하는 기업에게는 필수 불가결한 요소입니다. 저장된 방대한 양의 정보, 사람의 한계, 알려지지 않은 언어의 전자문서, 글자의 복잡성, 각각의 다른 언어의 프로그래밍 코드, 또 결코 적지 않은 비용 등등은 검색 문서 제작에 가장 빈번하게 걸림돌이 되고있습니다. IT 리더들에게는 차세대 e-디스커버리 솔루션을 들여다보고, 향후에 모든 법적 분쟁에 도움이 되는 문서를 식별 및 분석하고 이에 따라 신속히 세세하게 그 문서들의 증거제시를 살펴볼 수 있는 능력이 요구되고 있습니다.

이는 어떻게 국제적 기업이 다국어 문서를 처리하며 Basis Technology 가 제공하는 엔티티(Entity)추출 도구를 이용하여 차세대 e-디스커버리에 대응하는지를 보여줍니다. 또한 어떤 규모와 언어든 발견한 증거로 그들을 대변하는 법적 전문인과, 기업이 법적 분쟁에 응수할 수 있도록 합니다.

“...자사(自社)는 모든 관련 문서를 다국어로 만들 수 있습니까?”

문제의 화두: 모든 관련문서의 작성

법정분쟁과 관련해 모든 관련문서를 만드는 것은 매우 간단한 사항입니다. 하지만, 이에 관련한 증거와 정보를 찾는 것은 매우 힘들며 또 기업 측에서 이를 받아들이고 법 관련해 이를 이용하지 못한다면 법정 관련 패소 및 손상된 이미지, 위약금 등의, 결과는 볼 보듯 뻔할 것입니다. 법적 분쟁에 관련된 모든 법률 문서 생산에 필요한 기술 및 인력은 기업의 수용 및 활용 가능한 언어와 포맷의 문서의 양과 매우 직접적인 관련이 있습니다. 이는 기업 문서 관련 저장소를 보면 알 수 있습니다. E-메일, 고객 관리 장부, 회계구좌 장부 그리고 모든 개인 컴퓨터 정보, 즉 모든 정형과 비정형 포맷과 서로 다른 코드와 코드페이지 등등의 것들을 예로 들 수 있습니다. 기업은 이 모든 것들의 소스를 검색하고 정보 및 증거가 될 수 있는 자료를 확보하며 향후 잠재적으로 다른 분쟁에 관련 연관성이 있는지를 판별 해야 합니다. 이 정보 검색과 문서 식별은 수백만의 문건을 찾아 법적으로 검토를 해야 하며 이는 단일 언어만 검색 혹은 식별하게 됩니다. 하지만 과연 국제기업들은 어떻게 각각 다른 다국어의 문서를 검색하고 식별하여 검토가 가능한 것일까요?

통상적으로 하나의 기업 e-메일 주소는 4.3 기가 바이트의 전자 데이터를 만들고 있다. 그리고 그 숫자는 2011년 까지 연간 6.7 기가 바이트로 성장할 것이다.

-레디카티(Radicati) 그룹-

단일언어 문서예비 식별 비용만도 수십만 달러에 육박하는 것이 일반적입니다. 그런데 다른 언어문서의 검토에 대한 프로젝트 추가비용 및 번역가 또는 기타 국가의 법률 팀에 대한 추가 비용은 상상을 초월할 것입니다. 이에 관해 언어와 상관없이 과연 어떤 것이 모든 관련 문건을 식별하는 답이 될 수 있을까요?

기업은 자동 다국어 처리와 엔티티(entity)추출 등의 기능을 이용하여 법조인들이 일일이 분석 및 검토해야 하는 방대한 문서의 양을 줄이고 모든 검색 요구사항에 대응하여야 합니다. 이 다국어 언어 수용은 포괄적인 전자증거개시(E-discovery)전략과 부합되어 사내(社內)와 대외로의 전자 증거개시(discovery)의무를 효율적으로 수행할 수 있도록 해줍니다.

“법률가는 증거와 그들의 전자증거개시(E-discovery) 전략을 문서화하고, 고객을 변호해야 할 의무가 있다. 당신은 어떻게 당신이 만든 데이터의 양을 절감할 수 있는지 변호 해야 한다. 만약 당신이 300 기가바이트에 데이터를 25 기가로 줄일 수 있다면, 당신은 그 일이 어떻게 일어 났는지 설명할 수 있게 된다. 이는 당신이 어떻게 그 모든 각각의 언어의 전자 공유정보를 당신 것으로 만들었는지 증명해야 한다는 것이다.”

Magistrate Judge Paul Grimm ·
Victor Stanley, Inc. v. Creative Pipe, Inc. 中에서
No. 06-2662 (D. Md. May 29, 2008)

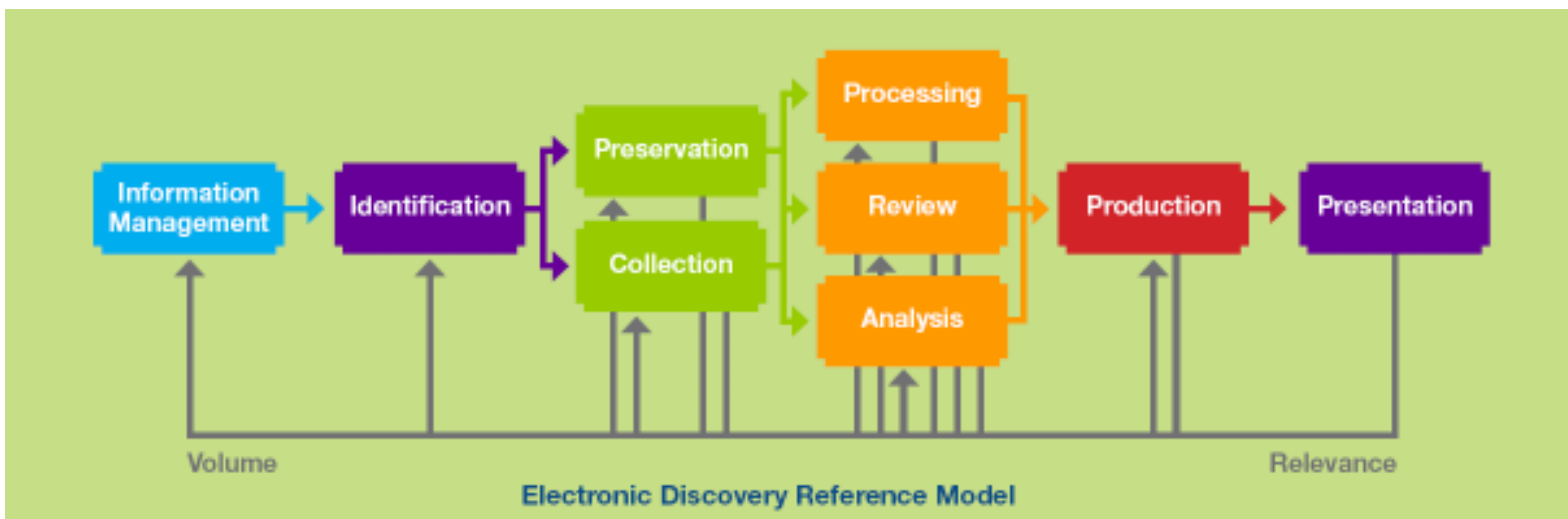
**Next Generation e-Discovery
Multi-language Processing & Entity Extraction**

다국어의 e-디스커버리 만들기

디스커버리 프로세스에서는 많은 단계가 존재하며 또 각각의 단계는 수백에서 수천의 시간할애를 요구하기도 합니다. 실사 e-디스커버리에 전형적 모델이 없다고 할지라도, e-디스커버리 솔루션은 기업에게 많은 노동력을 필요하는 작업을 간편화하며, 작업 효율을 향상시키며, 비용을 저감할 수 있습니다. 한번 필요에 의해 만들어진 모델의 전자적 증거의 저장은 모든 잠재적 요청의 관련된 모든 문서를 식별할 것입니다. 최초의 식별 단계의 목적은 저장된 전자정보에서 잠재된 소스를 찾고 또 그것의 영역, 폭, 깊이 등의 범위를 정하는 것입니다. 이는 기업이 내재적으로 더 많은 정보를 저장할 때 간편한 키워드 검색으로 수천 수만의 문서 중 선별된 문서를 찾아줄 것입니다.

기업의 정형과 비정형의 데이터 다룰 때 자료 선별작업 하나만으로도 엄청난 작업량을 나타냅니다. 더 많은 자료에서 분류해야 할 잠재적 증거자료와 더 많은 법률 팀의 검토를 요구하며 이는 엄청난 비용과 시간으로 환산됩니다. 하지만, 이런 잠재적 증거자료 식별능력의 부족은 법률 팀에 더 많은 비용할애를 남기는 결과를 가져옵니다.

결국 정답은 가장 알맞은 어플리케이션을 이용하여 초반 식별작업부터 각각의 다른 e-디스커버리의 전체단계 전반을 돕는 것 입니다. 하지만 다수의 다른 언어를 다룰 때는 오직 소수의 솔루션을 e-디스커버리 작업 흐름 전반에 잘 부합되어야 이를 통해 도움을 줄 수 있습니다. 다중언어 e-디스커버리의 이 주요단계를 프로세싱 (Processing), 리뷰(Review) 그리고 분석(Analysis)이라고 하며 그리고 이는 아래의 참조모델에 잘 나타나 있습니다.



Information management: 정보 매니지먼트 Identification: 식별 및 구분
 Preservation: 보존 및 저장, Collection: 수집
 Processing: 프로세싱, Review:리뷰, Analysis:분석 Production: 제조하기, Presentation: 제출 및 프레젠테이션
 Volume: 부피 ,Relevance: 연관성 Electronic Discovery Reference model: 전자증거개시(E-discovery)의 참조모델

제 1 단계: 프로세싱

잠재적 증거가 식별 및 저장 그리고 수집된 뒤에는 정보검색, 텍스트 마이닝 그리고 비슷한 어플리케이션으로 분류 및 처리되며, 문서에 쓰여진 언어는 다국어 전자 문서에 꼭 필요한 식별 및 분절화 과정을 거치게 됩니다. 디스커버리 어플리케이션은 이와 같은 사전 작업 없이는, 정형과 비정형 데이터를 정확히 처리할 수 없습니다.

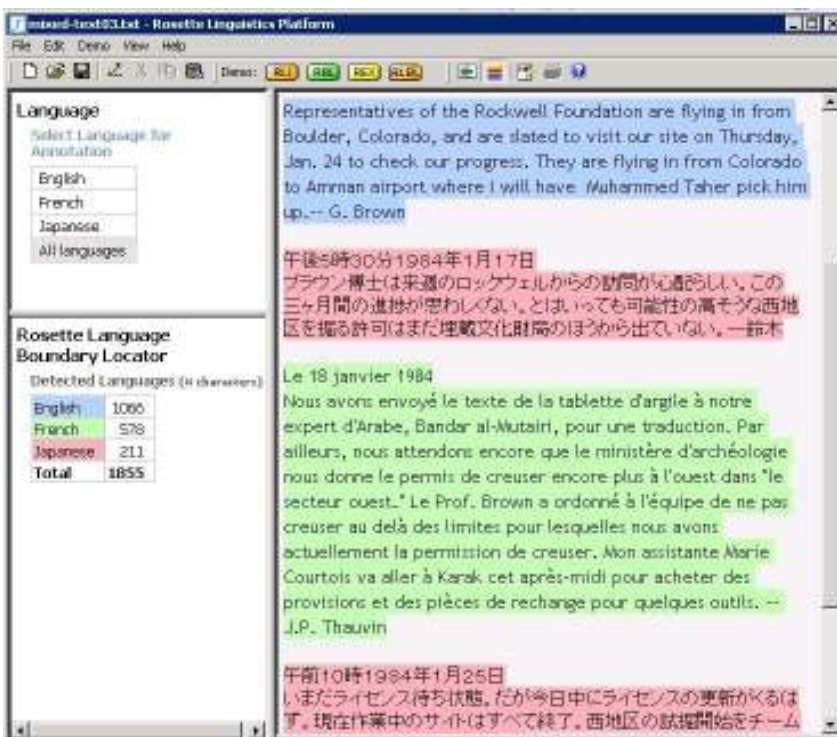
어플리케이션은 언어 식별 서비스를 이용하여 알려지지 않은 텍스트에 자동접근하여 높은 정확성으로 이 텍스트의 언어 및 인코딩을 식별합니다. 이 서비스는 문서에 있는 단일 언어를 식별하고, 그리고 다중언어와 그 경계를 구분하며, 넓은 범위의 아시아언어와, 그리고 유럽언어, 그리고 중동언어 등을 지원하고 식별합니다.

로젯 언어·인코딩 판별 시스템 (Rosette® Language Identifier)

로젯(RLI)는

55여개의 다른 언어의 문서를 식별할 수 있습니다. 로젯 언어·인코딩 판별 시스템은 언어 증거 식별로 프로세싱을 한층더 강화합니다.

- 단일 언어식별
- 다중언어를 포함한 문서 식별
- 다중언어 문서의 각각의 언어 처음과 끝의 경계 식별
- 어떤 언어를 포함했는지 그리고 각각의 언어의 전체에 대한 퍼센트 식별



로젯 언어 경계 식별기-RLBL (The Rosette Language Boundary Locator) - 로젯 언어 식별기의 컴포넌트는 하위문서로 단위로 분류 및 분절화 하여 언어를 식별합니다. 이 어플리케이션은 각각의 언어의 경계 서비스를 이용하여 자동으로 언어를 경계구분 및 분석하며 사람의 분석과 같은 장점과 정확한 분석적 통계를 산출하며 그리고 수집된 문서를 참조를 할 수 있도록 합니다.

로젯 언어 경계구분기(Rosette Language Boundary Locator)
다중언어 문서의 각각의 언어 경계를 보여주고 있습니다.

제 2 단계: 리뷰 (Review)

프로세싱 단계가 끝나면, 모든 문서는 **재검토**하게 됩니다. 문서 전체검색 기술은 통상적으로 사용되며, 이것의 핵심기술은 컴퓨터 언어학(*computational linguistics*)으로 자동으로 디지털 텍스트를 분석하고 이를 저장 및 검색 또한 활용 가능하게 해줍니다. 강력한 언어 기술은 **범률 분석**의 필수조건이며 분절화, 언어분류, 분할 및 연결 태그, 문장 경계 탐지 및 명사구 추출과 같은 주요한 언어적 서비스를 지원합니다. e-디스커버리에서 이 같은 서비스의 중요성은 인간언어 프로세싱에 도전과도 같습니다.

인간 언어의 복잡성

인간의 언어는 심도 있고 복잡한 주제며 형태, 발음 및 문법 등의 인간 언어학의 본질에 대한 연구입니다. 그래서 e-디스커버리에 기술과 언어가 정점에 집중되면 범률 전문가들이 자신들이 최고의 언어 기반 솔루션으로 무장했으며 각각의 다른 언어 문서 안 증거들에 숨겨진 위험요소를 최소화 한다는 것을 확신하게됩니다.

인간의 언어와 컴퓨터 언어의 조우

분석과 분절 그리고 태깅 및 텍스트 리뷰는 유니코드, 코드 페이지 등등의 인간 언어와 컴퓨터 언어의 조합과 그에 수반된 복잡성 때문에 아직도 행해지는 어려운 도전이라 할 수 있습니다. 중국어 같이 복잡한 언어는 4 만 7 천여 개의 글자가 있으며 이중 일본어와 겹치는 부분을 포함하여 합쳐진 코드의 다양성은 상상을 초월하며 이는 선구적 언어 솔루션의 필요성을 나타냅니다.

로제트 형태소 분석 시스템 (Rosette Base Linguistics)

다국어 텍스트 분석

로제트 형태소 분석 시스템 (RBL)은 많은 메이저 언어의 분절화, 분류하기 또는 품사분석 등의 중요한 **분석** 기능을 사용합니다. 이 기능들은 언어의 가장 원형에 집중하여 더 많은 텍스트의 **분석**를 가능케 하며 더 적은 허위-긍정 증거를 결과로 산출 할 수 있게합니다. RBL 을 사용하여 정교한 형태를 분석하여 아랍 언어, 아시아 언어 그리고 유럽언어 같이 주어진 언어의 특별한 형태의 텍스트를 분절화하고 이를 연결하여 줍니다. 이 형태학적 접근은 통계적인 언어 다루는 법과는 한층 다르게 언어의 깊은 이해도와 더 정확한 결과를 산출할 수 있도록 해줍니다.

텍스트 리뷰의 필수사항

■ 분절화(Tokenization)

분절화(Tokenization)는 토큰(Token)이라 불리는 개체로 텍스트를 나누는 프로세스이며 각각의 토큰은 통상적으로 모든 단어에 부합되지만 간혹 중국어와, 아랍어 같이 이에 부합되기 어려운 언어가 있습니다.

■ 분류하기(Lemmatization)

표제어(Lemma)는 사전에 있는 단어의 기본양식입니다. 분류하기(Lemmatization)는 단어의 기본양식을 줄이며 또한 단어의 각각의 다른 변화된 양식을 그룹으로 만듭니다. 즉 이것은 "arbitraging"라는 어구를 찾으면 "arbitrage"라는 어구를 포함한 모든 문서를 찾는 것입니다.

■ 분해하기(Decompounding)

합성어는 독일어와 같이 몇 개의 단어가 연결되어 하나의 새로운 단어를 만드는 언어에서 찾을 수 있습니다. 예를 들어 "Job" 과 "Börse"를 연결해 새로운 "Jobbörson" (고용교환프로그램)이라는 새로운 단어를 만듭니다.

제 3 단계: 분석 및 엔티티(Entity) 추출 (키워드 추출)

e-디스커버리 계획수립의 마지막 단계는 잠재적 증거로 식별된 문서의 분석입니다. 이것은 매우 중요한 작업으로 이는 마지막으로 포석을 깔아 문서 안의 증거를 확보하고 이를 이용하여 소송관련 법원에 제출 할 수 있는 문서를 작성하며 모든 법률 고문을 관리 및 감독할 수 있도록 합니다. 자동 분석은 디스커버리 검색을 지원하며 이는 검색어와 관련해 법률 팀이 놓칠 수 있는 모든 관련 단어까지도 검색 가능케 합니다. 이 같은 디스커버리 검색을 가능케 하는 이 기술은 엔티티(Entity) 추출 기술이라 합니다. 엔티티(Entity)추출은 자동적으로 법률 전문가의 이용가능성이 있는 동일 문서 데이터를 기초하여 중요한 검색어등을 찾습니다. 정형 및 비정형 문서에서 이 같은 엔티티(entity)를 찾거나 혹은 같은 의미의 유닛을 찾는 것은 디스커버리 검색이 특정화된 법률 관련 단어나 혹은 알려지지 않은 관련 단어를 찾는 것을 가능케 해주며 이 기술은 현 e-디스커버리 검색 기술이 할 수 없는 엔티티(entity)를 식별 및 추출합니다.

사람은 엔티티(entity)를 단어와 문맥을 보고 구별을 합니다. 귀하가 문서에서 모든 사람의 이름을 검색한다고 가정하였을 때, (혹 예상치 못했던 이름 까지도)또 그 문서에는 “Mr. John Xyzzy spoke…” 와 같은 인용구도 포함되어있습니다. 설사 귀하가 Xyzzy 라는 이름을 단 한번도 들어본 적이 없어도 귀하는 그것이 사람의 이름이라는 것을 추측합니다. 귀하는 단어의 대문자 사용 및 그리고 “Mr. [대문자 명사] [대문자 명사] [동사]…”와 같은 문맥 패턴을 통해 이와 같은 사실을 추론합니다.

엔티티(Entity)추출은 이 같은 사람의 생각형태와 같은 프로세스를 대입하고 문맥의 정황을 파악하여 컴퓨터 프로세스를 재창조합니다. 만약 date라는 단어가 문서에 있다면 이는 시간의 한 축이라는 엔티티(entity)로 나타내거나 혹은 완전히 다른 food라는 엔티티(entity)에 더 할 수 있을까요? date라는 단어는 수십 개의 문학적 의미해석이 가능하지만 date라는 단어는 food보다는 time이란 의미 쪽으로 좀 더 치우쳐 있습니다. 엔티티 추출 기술은 이와 같이 언어학 요소를 살펴 알맞은 문맥을 찾아 좀더 나은 정확도의 디스커버리 검색을 가능케 해줍니다.

엔티티(Entities)

고유용어 또는 문구: 이름, 장소, 날짜, 그리고 텍스트에 의미를 부여할 수 있는 식별자(identifiers).

엔티티 추출의 작용

대부분의 단어의 뜻은 보이는 문맥 상의 흐름으로 파악 할 수 있으며 같은 단어라고 할 지라도 다양한 내포된 암시에 따라 다른 의미를 가질 수 있습니다. 이와 반대로, 같은 의미일지라도 서로 다른 단어로 표현이 가능하며 또한 시각적 암시 및 내포되어 있는 의미에 따라 적지 않은 시각 및 용어적 해석이 가능합니다.

이 암시 및 문맥상의 흐름:

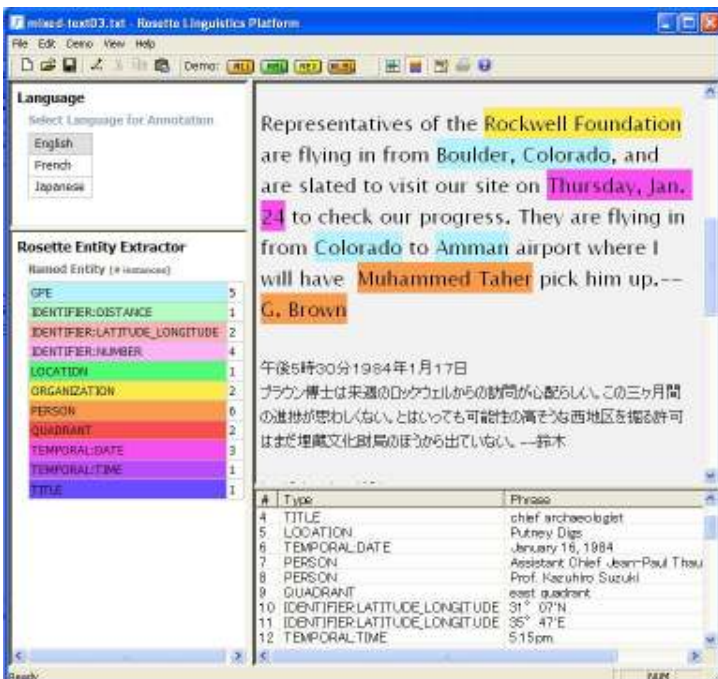
- 다른 단어로의 근접
- 단어의 서면양식 (예, 약어, 대문자 등)
- 품사(예, 제목, 목적어, 대명사 등등)
- 구두(句讀)점

로젯 엔티티 추출 시스템 Rosette® Entity Extractor

로젯 엔티티 추출시스템(REX)은 엔티티 추출 기술을 기초하여 분류, 관리, 분석 그리고 텍스트상의 정보추출 등의 통합 어플리케이션으로 디자인되어 있습니다. 이 기술들은 단순한 키워드 검색에서 제공 할 수 없는 레벨의 정확도의 다국어 문맥 환경을 만들어 줍니다. 로젯 엔티티 추출시스템(REX)은 중국어, 일본어, 한국어, 아랍어, 페르시아어, 우르두어, 네덜란드어, 영어, 프랑스어, 이탈리아어, 독일어, 스페인어등을 지원하며 더 많은 언어를 개발 및 지원할 예정입니다.

로젯 엔티티 추출기(REX)는 부통령“Vice President”이나 수익예측“earnings estimates” 과 같은 포괄적인 단어를 찾으며, 더 나아가 매우 세세한 Barack 오바마 대통령“President Barack Obama”이라던가 2009 년 5 월 22 일 “May 22, 2009” 같은 단어를 찾습니다.

이것은 리뷰(재검토)와 문서로부터 중요 정보 추출 등의 프로세싱의 주요한 단계로 추가구성 되어 다른 응용 프로그램이나 법률 팀에 분석되어야 할 정보를 준비를 하는 것 입니다.



영향 및 파장

2009년에는 그 어떤 해 보다 기업간의 법적 분쟁 및 소송이 잦았습니다. 중역 의사 결정자들은 증거 제출 문건을 만들거나 준비할 때 그들의 법률팀과 혹은 프로세스 및 기술이 밖으로 절대 노출되지 않도록 해야 합니다.

e-디스커버리는 비즈니스 운영이 글로벌화되고, 다국어적 민사 및 형사 소송이 늘어남에 따라 점차 그 영역이 확대되고 있습니다. 이에 따라 기업의 IT 팀과 법률 팀은 기업의 법적 분쟁의 위험부담을 최소화 하기 위해 뛰어난 언어학적 기술을 바탕으로 매우 높은 정확성의 e-디스커버리를 시행해야 할 필요가 있습니다.

로젯의 장점과 e-디스커버리 솔루션

회사가 국제적으로 성장함에 따라 모든 문서를 식별하고 저장하여 법적 디스커버리 및 포괄적 전략적 움직임은 글로벌 경영의 필수 조건이 되고 있습니다. 이에 따라 다국어 문서를 식별하고 프로세스하며 리뷰(재검토) 및 분석을 하는 능력은 이에 결정적 기술이라 할 수 있습니다. Basis Technology 의 로젯 솔루션은 다국어 텍스트를 식별할수 있으며, 유니코드에 표준 정형을 만들며, 그리고 이름과 지명 혹은 다른 법정 공방에 키워드가 될 수 있는 모든 단어를 찾아줍니다. 법률 팀과 IT 팀 모두 로젯 솔루션을 이용하여 효율적으로 현재 및 보류중인 소송과 관련된 모든 언어의 문서도 대처할 수 있게 됩니다.



Basis Technology Corporation

전화) + 81.3.3511.2947

팩스) + 81.3.3511.2948

e-메일) info@basistech.jp