<훈련결과보고 요약서>

<table>
<tr><td>성 명</td><td colspan="2">나상태</td><td>직 급</td><td colspan="2">전산주사</td></tr>
<tr><td>훈 련 국</td><td>영국</td><td>훈련기간</td><td colspan="3">2019.9.8.-2020.9.7</td></tr>
<tr><td>훈련기관</td><td colspan="2">더럼 대학
(Durham Univerity)</td><td colspan="2">보고서매수</td><td>102매</td></tr>
<tr><td>훈련과제</td><td colspan="5">빅데이터를 활용한 공정거래 감시 역량 강화방안 연구</td></tr>
<tr><td>보고서제목</td><td colspan="5">빅데이터를 활용한 공정거래 감시 역량 강화방안 연구
(A Tutoring System for Strengthening Capacity Using Big Data Analysis)</td></tr>
<tr><td>내용요약</td><td colspan="5">1. 서론
디지털 정보량이 기하급수적으로 증가함에 따라 방대한 데이터를 어떻게 활용하는지에 대한 연구가 전 세계적으로 급부상하고 있다. 즉 빅데이터를 이용한 새로운 가치 창출이 국가의 경쟁력을 강화하는 데 중요한 역할을 하고 있다.

공정거래위원회에서도 매년 많은 양의 데이터가 새롭게 수집되고 있으며, 이러한 데이터를 활용하여 공정거래 감시역량을 강화하는 방안을 고민하고 있다. 위원회가 보유하고 있는 많은 양의 데이터에 대한 단순 검색에서 벗어나, 다양한 빅데이터 분석 기법을 통해 사건처리 과정에서는 사건 맞춤형 법령이나 판례 데이터 등을 추천하고, 의사결정 과정에서는 신속하면서도 공정한 판단을 할 수 있도록 필요한 정보를 제공하는 등 경쟁법을 집행하는 과정에서 활용할 수 있는 방법들을 제안하고 있다. 또한, 빅데이터 분석을 통해 다양한 방법으로 공정거래 감시역량을 강화하는 방법들도 고민 중이다. 이 프로젝트에서는 사용자가 자신의 역량을 강화하기 위해 주어진 데이터를 연구</td></tr>
</table>

	하거나 분석할 때 튜터링(tutoring)을 할 수 있는 시스템을 제안한다. 지난 수십 년간 Information and Communication Technology (ICT)의 발전과 함께 e-leanring에 대한 대중들의 요구도 꾸준히 증가하였다. E-learning은 학습자의 시간과 공간적인 제약을 최소화할 수 있는 학습자 중심적인 교육 시스템이다. 특히, 인터넷 기술의 발전은 학습자가 쉽게 인터넷 콘텐츠에 접근할 수 있게 하여 교육환경의 변화에 많은 영향을 끼쳤다. 초기의 전형화된 e-learning 방식에서 발전된 최근 연구되고 있는 web-based adaptive tutoring 방식은 개개인의 배경 지식과 특성을 고려하여 각각의 교육에 최적화하는 방법을 제공하고자 한다. 이러한 web-based adaptive tutoring 방식은 Intelligent Tutoring System (ITS) 과 Adaptive Hypermedia System (AHS) 이 결합된 연구영역이다. ITS가 전통적인 Artificial Intelligence (AI) 기술을 기반으로 한 튜터 (tutor) 중심의 시스템이라면, AHS는 좀 더 유연한 검색 기술기반의 학습자가 중심이 되는 시스템이라고 할 수 있다. 사람마다 배우는 속도나 방식이 조금씩 다르므로 강사가 제한된 시간에 모든 학습자를 이해시키는 것은 불가능에 가깝다. ICT의 발달과 함께 이를 해결하려는 방법들이 연구되고 있으며 온라인 교육 시스템은 그 중 대표적인 방법이라 할 수 있다. 특히, 올해 일어나고 있는 COVID-19와 같은 pandemic이 발생하였을 때 온라인 교육 시스템은 중요한 대안이 되고 있다.

	이 프로젝트에서 제안하고 있는 빅데이터 분석 기반의 튜터링 시스템은 학습자가 학습을 진행할 때 학습효과를 높이기 위해, 학습에 제공된 데이터를 분석하여, 키워드를 추출한 후, 추출된 키워드를 이용하여 학습자에게 튜터링을 할 수 있는 시스템이다. 제안한 시스템이 학습자의 역량 강화에 이용될 수 있는지에 대한 가능성을 알아보고자 한다. 2. 본론 학습자의 역량을 강화하기 위한 다양한 튜터링 시스템이 존재하지만, 주어진 데이터를 분석하여 추출한 키워드를 이용하여 학습자를 튜터링하는 방법은 비교적 새로운 접근 방법이다. 이 프로젝트에서는 영문데이터들을 이용하여 테스트하였다. 먼저 실제 대학에서 컴퓨터 학과 학생들을 대상으로 강의 되고 있는 소프트웨어 공학 과목에서 사용되었던 강의 노트를 이용하여 전체 시스템을 테스트하였고, 다음으로 공정거래위원회 영문사이트의 Publications에 업로드된 문서를 이용하여 키워드 추출을 테스트하였다. 제안된 시스템은 크게 4개의 컴포넌트로 이루어져 있다. 사용자와 시스템 간의 상호작용과 사용자 데이터 분석을 위한 데이터 수집에 이용되는 사용자 인터페이스 컴포넌트와 주어진 데이터로부터 키워드를 추출하기 위한 키워드 추출 컴포넌트, 수집된 사용자와 추출된 키워드를 분석하기 위한 데이터 분석 컴포넌트와 마지막으로 사용자와 키워드 데이터 저장을 위한 데이터베이스 컴포넌트로 이루어져 있다.

	제안된 시스템 구현을 위해 사용된 프로그래밍 언어는 데이터 분석에 가장 많이 사용되는 파이썬 (Python)을 이용하였다. 파이썬은 데이터 분석에 필요한 대부분의 라이브러리 (Library)를 제공하고 있어 초보자도 쉽게 배울 수 있는 프로그래밍 언어이다. 이 프로젝트에서는 데이터 분석을 위해 가장 많이 사용되는 pandas라이브러리와 자연어 형태인 영문을 컴퓨터가 이해하기 쉽게 변환하기 위한 사전처리에 필요한 NLTK라이브러리, 키워드 추출과 클러스터링을 구현하기 위한 sklearn라이브러리, 그리고 사용자 인터페이스 구현에 사용된 Tkinter라이브러리 등 다양한 라이브러리를 이용하였다. 데이터 분석을 통해 사용자 역량을 강화하기 위해 제안한 튜터링 시스템을 구현하기 위해서는 가장 먼저 주어진 데이터로부터 키워드를 추출하여야 한다. 일반적으로 주어지는 데이터는 컴퓨터가 이해하기 어려운 자연어 형태이기 때문에 이를 컴퓨터가 이해하기 쉽게 구조화하는 사전처리가 필요하다. 사전처리는 크게 불용어 제거, 토큰화 과정, 어간 추출과정으로 나눌 수 있다. 첫 번째로 영문데이터의 경우 ‘the’, ‘he’, ‘she’와 같은 데이터 분석에 불필요한 불용어(stopwords)를 제거한다. 또한, 상황과 목적에 맞게 불필요한 단어를 불용어리스트에 추가하여 제거할 수도 있다. 예를 들어, 공정위에서 제공된 문서의 경우 ‘kftc’가 자주 등장하지만 데이터분석에 큰 의미가 없으면 ‘kftc’를 불용어 리스트에 추가할 수 있다. 다음으로 단어의 나열로 이루어진 문장을 토큰(token)이라 불리는 단위로 나누는 토큰화 과정이 필요하다. 컴퓨터가 이해하기 어려운 긴 문장을 컴퓨터가 이해하기 쉽게 토큰단위로 나누는 과정이다. 사전

	처리의 마지막 과정은 어간 추출 또는 표제어 추출과정이다. 이 과정은 서로 다른 단어지만 하나의 단어로 일반화시킬 수 있는 경우 하나의 단어로 일반화시키는 과정이다. 어간 추출의 경우 일정한 규칙을 기반으로 각 단어의 어간을 추출하는 방법으로 쉽게 구현할 수 있지만, 사전에 존재하지 않는 형태의 어간이 추출될 수도 있다. 표제어 추출의 경우에는 미리 준비된 표제어를 기반으로 하기 때문에 어간 추출에 비해 정확한 추출이 가능하지만, 새로운 단어 즉 표제어 리스트에 포함되어 있지 않은 단어의 경우에는 추출할 수 없고, 표제어 리스트를 기반으로 하기 때문에 리스트를 미리 만들어야 하는 제약이 있다. 위 방법들을 이용하여 사전처리된 데이터를 다양한 키워드 추출방법을 통해 키워드를 추출할 수 있다. 이 프로젝트에서는 다양한 키워드 추출방법 중 빈도기반 방법과 Term Frequency Inverse Document Frequency (TF-IDF) 그리고 Rapid Automatic Keyword Extraction (RAKE) 방법을 테스트하였다. 먼저 빈도기반 키워드 추출의 경우 통계를 이용한 가장 간단한 키워드 추출방법이다. 문서의 전체적인 맥락을 파악하는 방법으로는 효과적일 수 있으나, 동의어와 같은 단어의 의미나 순서 등을 무시하고 단순히 빈도만을 이용한다는 단점이 있어 전문적인 키워드 추출방법에 사용되기에는 제약이 있다. TF-IDF는 각 말뭉치(corpus)의 TF와 IDF 값을 곱하여 값이 큰 순서로 키워드를 추출하는 방법이다. 여기서 TF 값은 각각의 단어가 출연한 빈도수를 의미한다. 중요한 단어일수록 더 자주 출현한다는 가정을 반영하였으며, 문서의

	크기에 따른 편중현상을 방지하기 위해 실제 TF 값은 각 단어의 출현빈도 수를 모든 단어의 총 출현 회수로 나누어 정규화한 값이다. 이때, 높은 TF 값을 가지는 단어가 실제 키워드가 아닌 불용어일 경우를 방지하기 위해 IDF 값을 사용하였다. IDF 값은 총 문서 집합에 포함된 문서의 수를 특정 단어가 나타난 문서의 수로 나눈 값이다. 특정 단어가 나타난 문서의 수가 많을수록 그 단어가 보편적일 가능성이 크다. 즉, 보편적인 단어일수록 작은 IDF 값을 갖게 되고 반대로 큰 IDF 값을 가지고 있는 단어는 그 단어가 포함된 문서에서만 자주 나오는 단어로써 그 문서의 키워드가 될 가능성이 클 것이라는 의도를 반영한 것이다. 최종적으로 TF 값과 IDF 값을 곱한 결과를 이용하여 키워드를 추출할 수 있다. RAKE 방법은 일반적으로 키워드가 다수의 단어와 소수의 불용어로 이루어져 있다는 관찰을 기반으로 한 방법이다. RAKE 방법은 먼저 기능어와 불용어를 이용하여 문장 내의 키워드를 분할한 후, 문자열의 개수와 어구의 포함되는 단어들의 최대개수, 그리고 전체 문서에서 등장하는 빈도를 이용하여 점수를 환산하고 이를 기반으로 키워드 후보군을 뽑는다. 또한, 후보군 중에서 동시에 등장하는 단어들이 더 중요한 의미가 있는 것으로 간주한다. 그러므로 RAKE 방법은 다수의 단어로 이루어진 문장형태의 키워드들이 우선하여 추출되는 경우가 많다. TF-IDF 방법과 RAKE 방법을 이용하여 테스트한 결과, RAKE 방식을 통해 추출된 키워드는 여러 개의 단어로 이루어진 문장의 형태로 된 키워드가 주로 추출되었으며,

	TF-IDF를 이용한 키워드 추출의 경우 한 개의 단어나 두 개 또는 세 개의 단어로만 이루어진 키워드들이 추출되었다. 이 프로젝트에서 요구되는 키워드는 긴 문장형태의 키워드가 아닌 일반적으로 세 개 이하의 단어로 이루어진 키워드가 요구되기 때문에 TF-IDF 방법을 제안한 시스템에 적용하였다. 하지만, TF-IDF의 방법을 사용하는 경우 주요 키워드로 추출된 키워드 대부분이 하나의 단어로 이루어진 경우가 많았다. 하나의 단어로만 이루어진 키워드도 중요하지만 두 개나 세 개의 단어로 이루어진 키워드가 더 명확한 의미를 보여주는 경우도 많다. 예를 들어 'process'라는 키워드가 추출되었다고 했을 때, 이 단어가 시장에서 사용되는 process를 의미하는지 컴퓨터 공학이나 그 외 분야에서 사용되는 process를 의미하는지 명확하지 않지만, 'software process'나 'software process improvement'인 경우 키워드의 의미가 좀 더 명확해진다. 그러므로, 이러한 문제점을 해결하기 위해 threshold 값을 이용하였다. TF-IDF를 이용하여 하나의 단어로 이루어진 키워드들을 먼저 추출하고 추출된 단어를 포함하고 있는 2개나 3개의 단어로 이루어진 키워드 중에 threshold 값을 넘는 경우에만 키워드로 설정하였다. 추출된 각각의 키워드들은 Open Education Resource (OER) 서비스와 연결된다. 대표적인 OER 서비스로는 Google, Wikipedia, MOOC, Google Scholar 등이 있다. 관리자에 의해 새로운 OER 서비스를 전체 목록에 추가하거나 불필요한 서비스를 목록에서 삭제할 수도 있다. 사용자는 설정된 OER 서비스 목록에서 원하는 서비스만을 선택하여 개인화된 OER 서비스 목록을 유지할 수 있다.

	또한, 개인화된 튜터링을 제공하기 위해 사용자 정보를 수집하였다. 사용자 정보는 암시적인 방법이나 명시적인 방법을 통해 수집하고 분석할 수 있다. 많은 양의 사용자 정보가 이미 수집된 경우에는 수집된 사용자 정보를 분석하여 현재 사용자와 유사한 패턴을 가지는 집단의 정보를 이용하여 암시적으로 현재 사용자의 행동을 예측할 수 있다. 하지만, 이 프로젝트와 같이 분석할 데이터가 충분하지 않은 초기 시스템에는 적용이 어려우므로 간단한 질문을 통한 명시적인 방법으로 사용자 정보를 수집하였다. 질문에 대한 사용자의 답변을 기반으로 k-means 클러스터링 방법을 이용하여 사용자를 그룹화하였다. 이때, 몇 개(k)로 그룹화할지에 대한 판단을 위해서 elbow 방법을 사용하였다. Elbow 방법은 초기 k값을 1로 설정한 후 각각의 포인트와 중앙점 사이의 Sum of Square Error (SSE) 값을 계산한다. k 값을 1씩 증가시키면서 SSE 값을 계산하여 elbow 형태의 곡선이 나오는 지점을 최적의 k 값으로 선택하는 방법이다. K 값이 선택된 후, k-means 알고리즘을 이용하여 사용자 데이터를 k 개의 그룹으로 그룹화시킨다. 이와 함께, 사용자에게 적합한 키워드 그룹을 매칭시키기 위해 추출된 키워드들을 그룹화시켰다. 키워드 간의 거리를 계산하기 위해 텍스트 형태의 데이터를 컴퓨터가 쉽게 이해할 수 있는 숫자의 벡터 형태로 변환하였다. 이렇게 변환된 키워드들에 대해 elbow 방법을 사용하여 키워드 그룹화를 위한 최적의 k 값을 구한 후 k-means 알고리즘으로 그룹화하였다.

	이렇게 그룹화된 사용자 데이터와 키워드 데이터를 이용하여 사용자 특성에 맞게 튜터링 할 수 있는 시스템을 개발하였다. 각각의 사용자는 하나의 키워드 그룹과 매칭되며, 매칭된 키워드를 이용하여 사용자의 학습에 도움을 준다. 사용자는 매칭된 키워드를 이용하여 불필요하다고 생각되는 키워드를 삭제할 수 있으며, 더 중요하다고 생각되는 키워드들에 대해서는 강조(highlight)하여 표시할 수도 있다. 아울러, 키워드 추출 알고리즘에 의해 추출되지 않았지만, 사용자가 중요하다고 생각되는 키워드나 제공된 문서에 존재하지 않지만 다른 키워드들과 연관되어 중요하다고 생각되는 키워드를 추가할 수도 있다. 테스트를 위해 실제 강의에 사용되었던 소프트웨어 공학 과목의 강의 노트를 이용하였다. 사용자 정보를 수집하기 위해서는 실제로 강의를 들었던 학생들의 정보를 수집하여 테스트하여야 하지만, COVID-19로 인한 시간적, 공간적 제약으로 가상의 사용자 데이터를 만들었다. 가상의 사용자 데이터는 명확한 가정을 기반으로 만들었다. 테스트에 사용된 가상의 사용자 데이터는 3개의 그룹으로 그룹화되었으며, 키워드는 4개의 그룹으로 그룹화되었다. 사용자가 처음 시스템을 사용할 때 명시적으로 3가지 질문이 나오며, 질문에 대한 답변을 이용하여 현재 사용자와 가장 적합한 사용자 그룹에 매칭시켰다. 사용자 그룹에는 다수의 이전 사용자들에 대한 정보들이 있으며, 각각의 사용자는 하나의 키워드 그룹을 가지고 있다. 그러므로, 각각의 사용자 그룹에는 일반적으로 한 개 이상의 키워드 그룹

	을 가지고 있다. 현재 사용자가 속한 사용자 그룹에 있는 키워드 그룹들의 키워드들을 사용자가 확인한 후 사용자의 선택에 따라 현재 사용자에게 적합한 하나의 키워드 그룹이 매칭된다. 이렇게 매칭된 키워드는 사용자가 학습하기 전과 후 그리고 학습 중에 다양한 방법으로 튜터링에 사용된다. 키워드를 이용하여 학습 전에는 예습에 활용할 수 있고, 학습이 끝난 후에는 복습에 키워드가 이용될 수 있다. 학습 중간에는 키워드 정보를 이용하여 사용자가 수업 내용을 정확하게 이해하고 있는지 확인하거나 추가적인 정보를 제공할 수도 있으며, 과제에 이용될 수도 있다. 아울러, 공정거래위원회의 영문 웹사이트에서 제공하고 있는 데이터를 이용하여 키워드 추출을 테스트하였다. 이렇게 추출된 키워드를 이용하여 사용자가 해당 데이터를 학습할 때 튜터링 서비스를 제공하여 사용자의 역량을 강화하는데 제안한 시스템이 이용될 수 있다. 기존의 튜터링 시스템은 일반적으로 경우 많은 시간과 노력이 요구되지만 여기서 제안한 시스템은 주어진 데이터를 분석하여 추출된 키워드를 활용한 비교적 간단한 튜터링 시스템이다. 이러한 시스템은 사용자의 학습을 도와 역량을 강화하는데 이용될 수 있다. 3. 결론 이 프로젝트에서는 공정거래 감시역량을 강화하기 위한 하나의 방안으로써 공정거래위원회에 축적된 문서를 학습할 때 튜터링을 통해 학습을 도와주는 시스템을 제안하였다. 이 프로젝트의 목적은 주어진 텍스트 형태의 데이터를

	분석하여 추출한 키워드를 이용하여 사용자가 해당 데이터를 학습할 때 튜터링 서비스를 통해 학습에 도움이 되는 시스템으로써의 가능성에 관한 연구이다. 제안한 시스템은 간단한 질문에 대한 사용자의 답변을 기반으로 현재 사용자와 가장 유사한 사용자 그룹에 매칭시키고, 이와 동시에 주어진 데이터로부터 추출된 키워드를 그룹화하여 현재 사용자에게 가장 적합한 키워드 그룹을 매칭시킨다. 이렇게 매칭된 키워드를 활용하여 사용자가 데이터를 학습할 때 다양한 방법으로 도움을 줄 수 있는 튜터링 서비스를 제공한다. 사용자에게 도움이 되는 시스템을 위해서는 사용자들의 적극적인 참여가 무엇보다 중요하다. 사용자가 참여한 사용자 정보를 이용하여 그룹화가 가능하기 때문이다. 또한, 현재 시스템은 영문 분석에 최적화되어있다. 한글을 사용하기 위해서는 자연어 형태인 한글의 사전처리 등 추가적인 개발이 필요하다. 마지막으로 현재는 음성데이터나 동영상 데이터에 대한 분석은 가능하지 않고 오직 텍스트 형태의 데이터만 분석만이 가능하다. 이러한 제약에도 불구하고, 이 프로젝트에서 제안한 튜터링 시스템이 사용자의 데이터 학습을 도와 사용자 역량을 강화하는데 이용될 수 있다는 가능성을 확인하였다.