자연어 처리를 활용한 전세계 전염병 알림 사이트

곽찬우[1] · 김예찬[2] · 최진황[3]
[1·2·3]명지대학교 정보통신공학과
[1]yhntfx777@naver.com / [2]dpcksdl3@naver.com
[3]chjh2425@naver.com

A global epidemic notification site using natural language processing

Chan-Woo Gwak[1] · Ye-Chan Kim[2] · Jin-Hwang Choi[3]
[1·2·3]Dept, of Information And Communication Engineering, Myongji University

요 약

본 논문에서는 글로벌화가 진행됨에 따라 전 세계의 재난 경보시스템의 중요성을 인지하고, 현재 유행하고 있는 코로나 바이러스를 중점으로 알림 사이트를 개발하였다. 기존의 정보 제공 사이트들과 차별성을 두고자, 기존의 정보들을 분석하고 재분류하여 새로운 형태의 사이트의 형태를 가진다. 이를 위해 인공지능의 한 분야인 자연어처리를 활용하여 기존의 정보를 수집하고 가공하여, 보다 투명하고, 효율적이고, 가치 있는 정보를 게시한다. 정보의 정확성과 데이터 절감을 위하여 여러 조건을 통해 기존의 정보들을 재분류 작업 이후 WATSON NLU(Natural Language Understanding)를 통해 분석하여, 필요한 정보들을 각 대시보드에 게시한다. 각 대시보드는 NLU분석에서 얻을 수 있는 정보들을 기반으로 구성되어 있으며, 간결성과 가시성을 기반으로 정보를 확인할 수 있는 알림 사이트이다.

1. 서론

최근 중국 우한에서 시작하여 전 세계적으로 코로나 바이러스가 유행하고 있다. 이러한 전염병에 관한 정보의 관심도는 점차 증가하고, 글로벌화가 진행됨에 따라, 전 세계에 재난 경보 시스템과 신속한 전염병 정보 전달의 중요성이 부각되고 있다. SNS, 검색엔진에서 다양한 사용자들의 정보를 활용하여 바이러스 경보 알림을 줄 수 있는 사이트의 개발을 목적으로 한다.

<그림 1> 자연어 처리의 기능

정보를 수집(크롤링)하고, 이를 WATSON NLU(Natural Language Understanding)를 이용하여 새로운 형태의 정보로 가공하고 정형화한다. 정보들을 분류하여 게시하는 기존의 사이트들과는 달리 자연어처리[1] 과정을 통하여 분석한다. 분석 이후 키워드, 감정, 구문해석 등의 가치 있는 정보로 가공되며, 조금 더 효율적이고 투명한 정보를 활용한 바이러스 경보사이트를 개발하였다.

또한 전 세계 유사언론 및 출처를 알 수 없는 곳에서 생성된 다양한 페이크 뉴스로 인해 전염병의 잘못된 정보 전달의 심각성 또한 커지고 있다. 전염병의 중요한 부분인 신속하고 정확한 정보 전달을 위하여 본 프로젝트를 설계 및 구현하였다.

2. 시스템 기획 및 구성

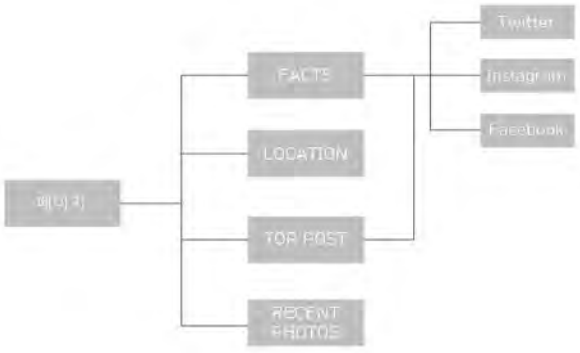
시스템 구성에 관하여 설명하기에 앞서 자연어처리 즉, NLP(Natural Language Processing)에 관하여 주목할 필요성이 있다.

인간의 언어는 놀랄 만큼 복잡하고 다양하다. 인간은 말과 글을 사용하여 무한한 방식으로 자신을 표

현한다. 언어에는 수백 가지의 종류와 방언이 존재하는 데다, 문법과 구문 규칙, 용어, 속어도 저마다 다르다. 우리는 글을 쓸 때 단어의 철자를 틀리거나, 약어를 사용하거나, 혹은 구두점을 생략하기도 한다. 말을 할 때는 지방마다 특유의 억양이 드러나며, 웅얼거리거나, 말을 더듬거나 다른 언어의 용어를 차용하기도 한다. 언어의 모호성을 완화하고, 음성 인식이나 텍스트 분석 같은 다수의 데이터에 유용한 숫자 구조를 추가가능 하기에 자연어처리 과정의 중요성이 증가하고 있다[2].

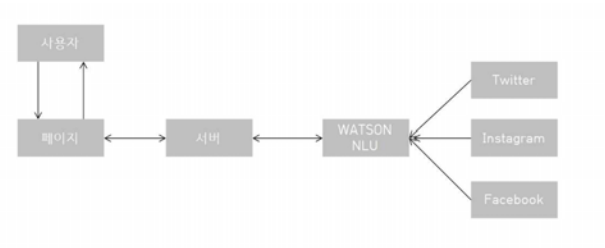
본 논문에서는 기존의 사이트들과는 차별성을 둔 NLU를 활용하여 수집한 데이터를 가공하여 조금 더 효율적이고 투명한 정보를 제공하기 위한 시스템을 기획하였다.

<그림2>와 같이 페이지의 접속을 통하여 정보를 대시보드별로 한눈에 파악할 수 있도록 구성하였다. UI의 구성은 Django 프레임워크를 이용하고, 대시보드별 항목의 가시성을 최우선으로 하여 간결하게 구성하였다. 대시보드는 항목별 게시글 수를 확인할 수 있는 FACTS, 게시글별 위치분포를 비율로 표시한 LOCATION, 일자, 페이지마다 가장 인기 있는 게시글과 키워드를 표시하는 TOP POST와 최근 게시글들의 사진을 확인할 수 있는 RECENT PHOTOS까지 4개의 대시보드를 통하여 정보를 확인할 수 있다.



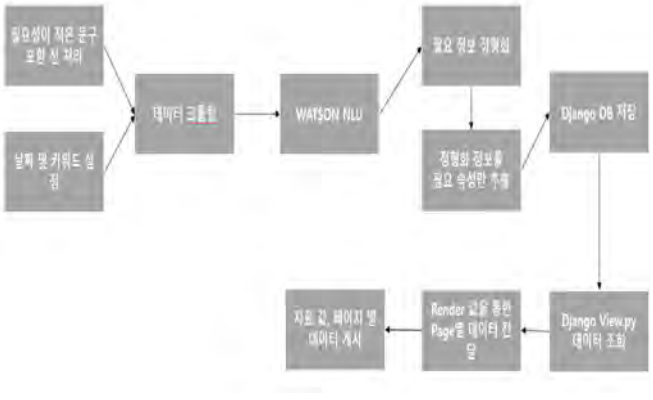
<그림 2> 페이지 대시보드

사이트는 <그림3>과 같이 각 페이지의 크롤러를 통하여 전염병에 관련한 정보들을 크롤링한다. 이후 정보들을 페이지별로 분류하여 WATSON NLU를 통하여 자연어 분석과정을 거치고, 정형화된 데이터로 변형하게 된다. 데이터들을 포함한 서버에서 페이지를 통하여 사용자에게 정보를 제공한다.



<그림 3> 시스템 구성도

사이트는 <그림 4>와 같은 구조로 동작한다. 원하는 정보의 키워드와 기간을 조건으로 하여, 데이터를 수집하고 WATSON NLU 분석을 통한 정형화 정보를 확인한다. 분석 후에는 관계성, 이름, 언어, 분석텍스트 본문, 국가, 개체, 키워드, 카테고리, 긍정 또는 부정의 점수, 관계 등을 포함한 다양한 정형화된 정보를 얻을 수 있다. 이 중 페이지 게시에 필요한 개체, 국가, 점수 등만을 추출하여 이를 저장하고 Django Framework를 이용하여 화면에 정보를 게시한다.

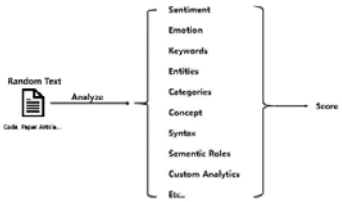


<그림 4> 사이트 동작 구성도

3. 사이트 동작 원리 및 구성

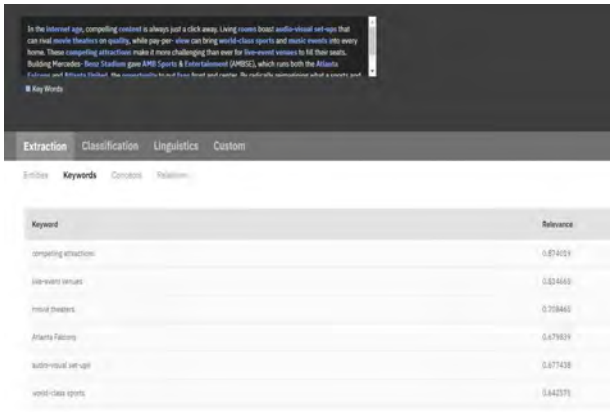
정보를 수집하기에 앞서 정보의 정확성과 데이터 용량 처리의 절감을 위하여 욕설과 불필요한 문구가 포함된 데이터를 예외처리하여 크롤링한다. '코로나바이러스'라는 키워드를 제시하고 이에 관련한 최근 1주일 분량의 정보를 수집한다. 데이터베이스에 최초로 1회로 저장한 이후, 매일 업데이트 될 때마다, 이전 날의 하루 동안의 게시글을 추가로 데이터베이스에 저장하게 된다. 이후 자연어 처리 과정을 거치

게 된다. WATSON NLU 처리 과정은 <그림5>와 같다.

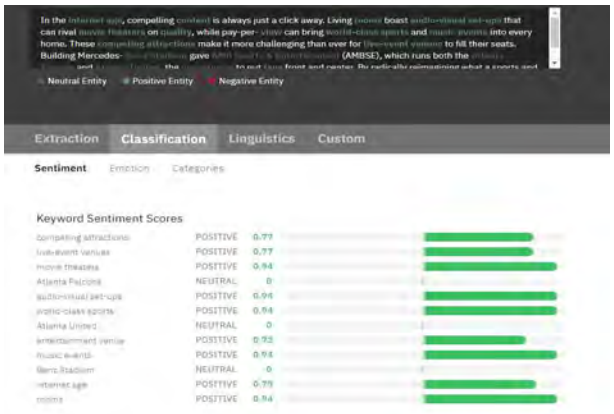


<그림 5> NLU 처리 과정

크롤링한 텍스트 데이터들을 WATSON NLU를 이용하여 분석하게 되면 <그림 6, 7>과 같이 Sentiment, Entities, Keywords와 같은 10개 이상의 항목들로 분류되어 분석결과가 발생한다. 문맥의 긍정, 부정을 평가한 점수(Score), 중심이 되는 키워드(Keyword), 문맥의 개체(Entities), 카테고리 분류(Categories) 등의 정보를 얻을 수 있다. 이 과정에서 전염병 관련 정보들에 필요한 감정 관련 점수의 조건을 설정하여, 추가로 데이터를 선별하고, 필요한 정보인 지역, 카테고리, 점수 등을 Django 데이터베이스에 저장하게 된다.

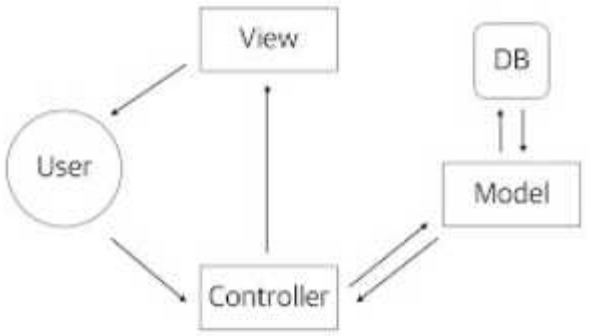


<그림 6> NLU 분석 결과(1)



<그림 7> NLU 분석 결과(2)

데이터베이스에 데이터 저장 시 아래 <그림 8>의 구조의 MVC(Model-View-Controller) 패턴을 활용한다. Model에서 각각의 데이터들의 타입을 지정해주어 장고 데이터베이스에 저장한다. 데이터 저장 확인 후 View 페이지에서 object_all 함수를 통하여 저장된 모든 데이터를 조회하여, 각각의 항목에 맞도록 설정해주고, 이를 Index 페이지로 값을 전달하면 사용자 화면에 정보들이 표시된다.



<그림 8> MVC(Model-View-Controller) 패턴 구조

이러한 구조의 흐름을 통하여 원하는 정보를 수집하고, 자연어 처리 과정을 통한 재가공 후 사이트에 게시할 수 있다. 사이트는 정보습득의 효율성을 최대화하기 위하여 가시성과 간결성을 우선적으로 <그림 9, 10, 11>과 같이 설계하였다.

<그림 9> 구현된 웹 UI (FACTS 보드)

<그림 10> 구현된 웹 UI (LOCATION 보드)

<그림 11> 구현된 웹 UI (TOP POST 보드)

4. 결론

본 논문에서는 전 세계에서 업로드되는 바이러스 관련된 SNS 게시글을 매일매일 수집하여 비정형 데이터를 자연어 처리(NLU) 과정을 통한 정형화 데이터, 수치화된 데이터로 변환하여 웹 사이트에 정보를 보여주는 형태로 설계하고 이를 구현하였다.

구현된 웹 사이트에서는 전 세계 SNS(트위터, 페이스북, 인스타그램)에서 매일 업로드 되는 게시글들을 수집, 여과, 가공의 과정을 거쳐 긍정/부정, 업로드된 위치, 가장 인기 있는 게시글 등등 다양하고 투명한 정보들을 한눈에 파악할 수 있게 구현하였다.

설계 배경으로는 세계적 감염병 유행(pandemic)에 대한 많은 정보들을 SNS 및 여러 플랫폼에서 접할 수 있었지만 검증되지 않은 정보와 허위 사실들이 실시간 검색어에 오르내리며 오히려 더욱 불안감을 조성하는 분위기를 확인할 수 있었다. 유례없는 펜데믹 사태를 맞이한 지금, 페이크 뉴스로 인한 혼란을 줄이고 정확하고 신속한 정보 전달로 더 이상의 불필요한 피해를 최소화해야 한다.

위 프로젝트는 이러한 문제점에서 확실한 대안을 만들기 위해 진행되었다. 또한, 이번 프로젝트는 대한민국에서 한동안 화두였던 국민들의 안전 불감증에 대한 성장의 기회라 생각하며 진행되었다. 전 연령층에게 가시적으로 쉽게 정보 전달을 할 수 있도록 제작하였고 언제든 새로운 전염병에 대한 사실 보도를 즉각적으로 업데이트할 수 있도록 제작하였다.

자연어 처리를 활용한 전 세계 전염병 알림 사이트를 통해 제대로 된 정보 전달을 더욱 빠르게 진행할 수 있을 것으로 기대된다. 또한 언론사를 거치지 않고 전 세계 개개인이 업로드한 게시글을 수집하여 처리한 데이터이기 때문에 보다 투명하고 진실 된 정보를 얻을 수 있을 것이라 기대한다. 더욱 나아가 각 공공기관에서 전염병 관련 데이터 수집에 도움이 될 수 있고 전염병 외에 다른 키워드를 활용하여 더욱 방대한 세계적 이슈(사회, 경제, 정치 등)에 대한 데이터를 수집하여 처리할 수 있을 것으로 기대된다.

- 본 논문은 과학기술정보통신부 정보통신 창의인재 양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] Hong Bae Kim, “Natural Language Processing Technology of Machine Learning”, pp. 15 - 21 Jul 22. 2016

[2] SAS Insights Site, “Definition and Importance of Natural Language Processing ”