클라우드 서비스를 이용한 영어 말뭉치 구축 도구*

김성동, 김민우
한성대학교 컴퓨터공학부
e-mail:sdkim@hansung.ac.kr, tairian20002@gmail.com

English Corpus Construction Tool Based Using Cloud Services

Sung-Dong Kim, Minwoo Kim
School of Computer Engineering, Hansung University

요 약

본 논문에서는 영어 신문 사이트를 크롤링하여 뉴스 기사를 수집하여 영어 말뭉치를 구축하는 도구를 제안한다. 클라우드 서비스를 이용함으로써 장소와 시간에 구애받지 않고 말뭉치를 지속적으로 확장시킬 수 있을 뿐만 아니라 쉽게 구축된 말뭉치를 활용할 수 있다. 제안한 도구는 수집된 영어 신문 기사에 대한 통계 정보 즉, 문장 수, 단어 수 등을 제공한다. 웹 플랫폼에서 동작하므로 여러 명이 동시에 많은 데이터를 수집할 수 있다. 수집된 데이터는 자연어 처리 및 기계학습 연구에 활용될 수 있다.

1. 서 론

말뭉치는 자연언어처리에 있어 매우 중요한 정보의 원천이 된다. 언어는 지속적으로 변하고 새로운 단어, 용어 들이 계속 생성되고 있어 정보의 원천으로 신문기사는 매우 중요한 역할을 할 수 있다. 또한 신문 사이트에는 하루에도 수십 개씩 다양한 분야의 기사가 업로드 되고 있다. 이러한 신문 사이트를 크롤링하면 쉽고 빠르게 수많은 텍스트 데이터를 얻을 수 있다.

본 논문에서는 지속적으로 영어 말뭉치를 구축하기 위해 영어 신문 사이트로부터 신문 기사를 크롤링하여 텍스트 데이터를 추출하는 도구를 제안한다. 신문 기사는 계속 생성되기 때문에 시간, 장소의 제약 없이 언제든지 신문 기사를 추출하여 말뭉치를 확장할 수 있도록 클라우드 서비스를 이용한 웹 기반의 말뭉치 구축 도구를 개발하였다. 클라우드 서비스를 이용함으로써 서버 관리 및 필요한 소프트웨어의 유지 보수를 위한 노력없이 지속적으로 말뭉치를 수집하고 관리, 활용할 수 있는 기반을 마련할 수 있다.

2 장에서는 말뭉치 구축 도구의 구조를 설명한다. 3장에서는 신문 기사 크롤링 과정과 말뭉치를 활용하는 과정을 기술한다. 4장에서 현재 구축된 말뭉치의 통계와 이를 구축하기 위해 어느 정도의 노력이 필요했는지를 설명하면서 논문을 마무리 한다.

2. 클라우드 서비스를 이용한 말뭉치 구축 도구의 구조

* 이 논문은 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2017R1D1A1B03030878).

시간과 장소의 제한없이 말뭉치 구축 도구를 이용할 수 있도록 웹에서 도구를 사용할 수 있도록 설계하였으며 이를 위해 서버와 필요한 소프트웨어를 클라우드 서비스를 활용한다. 클라우드 서비스를 이용함으로써 운영 환경에 대한 고려를 할 필요없이 말뭉치 구축, 관리에 필요한 솔루션 개발에만 집중할 수 있다. 그림 1은 본 논문에서의 말뭉치 구축 도구의 구조를 보여준다.

(그림 1) 말뭉치 구축 도구의 구조

AWS(Amazon Web Service)의 EC2(Elastic Compute Cloud)[1] 서버에 Node.js 서버를 설치하고 이것을 이용하여 영어 신문 사이트로부터 크롤링을 수행하는 도구를 개발하였다. 크롤링 프로그램은 신문 기사에 대한 정보(신문 소스, 분야, 날짜, url 등의 메타 데이터)를 MySQL DB에 저장하고 신문 기사 본문을 텍스트 파일 형식으로 저장한다. 이 때 MySQL DB는 AWS의 RDS(Relational Database Service)[2]를 이용하고, 본문 텍스트 파일은 AWS의 S3(Simple Storage Service)를 이용한다.

웹 브라우저를 통해 크롤링을 요청하면 이를 처리할 작업 객

체(Crawler)를 생성하여 처리한다. 만약 Herald 뉴스 사이트를 선택하여 요청하면 Herald 뉴스 사이트를 크롤링할 수 있는 Crawler 작업 객체가 생성된다. 각 뉴스 사이트에 대한 작업 객체는 Crawler라는 추상 클래스를 상속받아 구현하게 되는데, 이 추상 클래스는 게시판형 사이트 크롤링 과정에서 수행되는 '게시판 URL 얻어오기', '게시판 파싱', '뉴스 기사 본문 파싱' 기능의 함수들을 추상화하여 만들어 졌다. 크롤링 할 새로운 뉴스 사이트에 대한 작업 객체를 추가로 구현할 때마다 Crawler 추상 클래스를 상속받아 3가지 추상 함수를 구현한다. 그림 2는 Crawler 추상 클래스의 구조를 보여준다.

Crawler
-newspaper: String -newsCategory: String -newsDivision: String -startDate: String -endDate: String
+startCrawling(callback: func) +crawling(page: int, callback: func) *#getNewsBoardUrl(category: String, division: String, page: int): String* *#parsePage(body, page): String[][]* *#parseNewsText(body): String*

(그림 2) Crawler 추상 클래스

getNewsBoardUrl은 크롤링 할 뉴스 기사의 분류 및 게시판 페이지를 통해 게시판 URL을 생성하여 반환한다. 그러면 Crawler는 게시판 URL에 접속하여 parsePage 함수를 호출하며 HTML Body 객체를 넘겨준다. parsePage에서는 게시판의 각 행에 있는 뉴스 기사 제목, 작성날짜, 본문 URL을 배열에 저장하여 반환한다. Crawler가 다시 본문 URL로 접속하여 parseNewsText 함수를 호출하며 뉴스기사의 HTML Body 객체를 넘겨주면 뉴스 본문을 파싱하여 반환하도록 구현한다.

3. 말뭉치 구축 도구의 활용

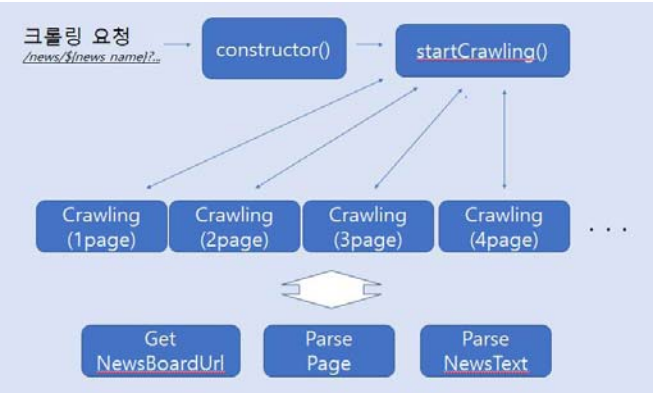
3.1 크롤링 과정

그림 3은 논문에서 제안한 웹기반 말뭉치 구축 도구를 이용하여 "데일리중앙" 신문의 "Sports" 분야의 "Domestic" 세부 분야에서 2019-01-31 ~ 2019-01-31 (1일분)의 기사를 크롤링 하는 모습을 보여준다.

뉴스 사이트 이름, 분류, 날짜를 선택하여 크롤링을 하면 수집한 뉴스 기사의 제목, 분류, 작성날짜, 본문 문장 수 및 글자 수, 원본 URL은 데이터베이스에, 그리고 본문 내용은 클라우드 저장소에 저장한다.

그림 4는 크롤링 과정을 보여준다. 크롤링 요청을 처리할 작업을 생성하고 뉴스 사이트 게시판 1페이지부터 날짜 범위를 비교하면서 뉴스 기사 크롤링을 진행한다. 각 페이지에 대한 크롤링(Crawling n page) 작업이 전부 끝나면 콜백 함수(callback function)를 통해 결과를 저장한다.

(그림 3) 신문 기사를 크롤링 하는 모습

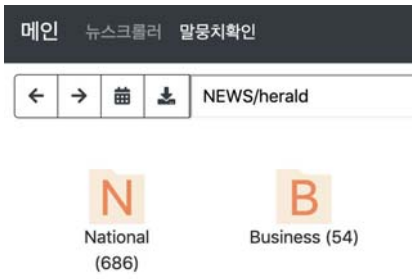


(그림 4) 신문 기사 크롤링 과정

3.2 말뭉치 활용

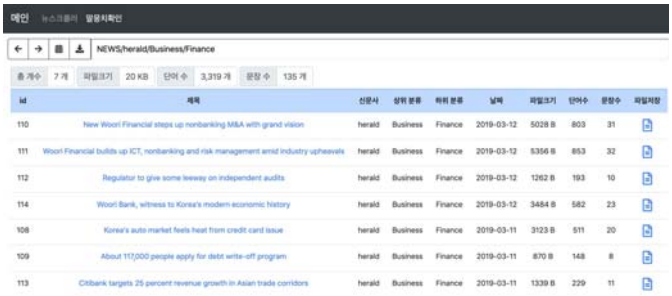
구축된 말뭉치는 그림 5에서와 같이 신문 사이트 이름으로 분류되고 각 신문마다 정의한 세부분야별로 구분되어 저장된다. 그림 6은 "herald" 신문의 세부 분야별 구축된 말뭉치 현황을 보여준다.

(그림 5) 말뭉치 확인 화면



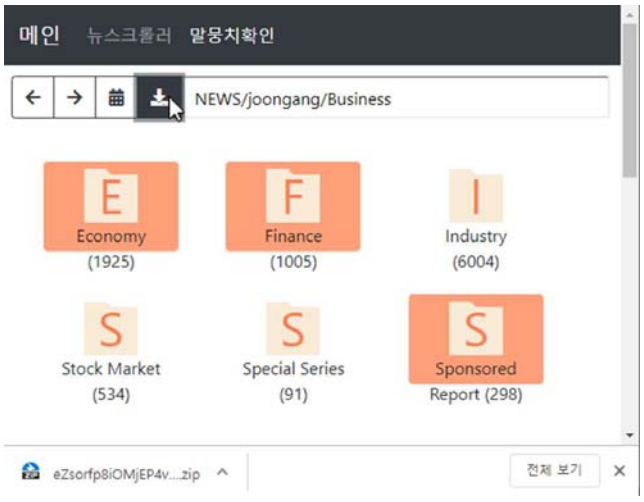
(그림 6) herald 신문의 세부 분야별 말뭉치 현황

그림 6의 특정 폴더를 클릭하면 해당 분야의 신문 기사 목록이 그림 7과 같이 나타난다. 기사의 제목으로 파일 이름이 할당되었으며 기사 아이콘을 누르면 기사를 확인할 수 있으며 개별 기사를 다운받을 수도 있다.



(그림 7) 특정 신문의 특정 분야 신문 기사 목록의 모습

특정 폴더에 있는 모든 신문 기사를 다운로드 받을 수 있는데, 폴더 아이콘을 선택하여 한 번에 여러 개의 파일을 ZIP 파일로 받을 수 있다. 그림 8은 여러 개의 파일을 한 번에 다운로드 받는 모습을 보여준다.



(그림 8) 여러 개의 텍스트 파일을 한 번에 다운로드 받기

4. 결론

본 논문에서는 지속적으로 시간과 장소의 제약없이 영어 말뭉치를 구축하기 위해 웹기반의 말뭉치 구축 도구를 제안하였다. 웹기반 도구를 개발하기 위해 클라우드 서비스를 이용하였으며, 새로운 언어 현상을 지속적으로 수용할 수 있는 말뭉치의 구축을 위해 영어 신문 기사 사이트로부터 크롤링을 통해 신문 기사를 추출하고 이를 말뭉치로 구축하였다. 표 1은 2019년 4월까지 구축한 말뭉치의 현황을 보여준다.

표 1 말뭉치 구축 현황

신문사/분류	뉴스개수	파일크기	단어수	문장수
joongang	10,524 개	26,080.6 KB	4,402,787 개	196,652 개
herald	740 개	1,773.3 KB	301,071 개	13,407 개
reuters	14,128 개	37,540.2 KB	6,426,751 개	266,909 개

현재 약 45만 문장, 1,100만 단어의 영어 말뭉치가 구축되었다. 이를 위해 사람이 들인 노력은 신문 사이트, 분야, 기간 설정을 하는 것뿐이며, 기간 설정을 한 달 단위로 수행하여 여러 기간에 걸친 기사를 수집하였다. 따라서 사람의 노력은 거의 소요되지 않으면서도 많은 말뭉치를 빠른 시간에 구축할 수 있었다. 또한 신문 기사는 계속 작성되고 업로드 되고 있어 지속적인 말뭉치 확장을 쉽게 할 수 있다. 그리고 구축한 말뭉치를 언제 어디서든 확인하고 필요한 말뭉치를 다운받아 기계학습, 자연언어처리 등의 연구에 활용할 수 있어 연구에 매우 중요한 자원의 역할을 할 수 있다고 기대한다.

참 고 문 헌

[1] https://aws.amazon.com/ko/ec2/
[2] https://aws.amazon.com/ko/rds/
[3] https://aws.amazon.com/ko/s3/