

[붙임]

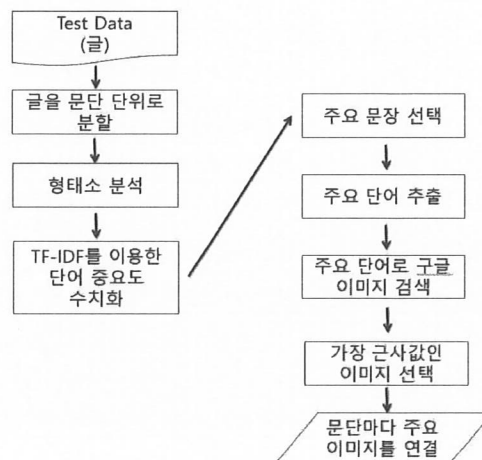
DU-도전학기 결과보고서

성명		학번	
단과대학		학과(전공)	
도전학기 과제명	(한글) 글에서 중요 단어들을 추출하여 해당하는 이미지 연결하기 (영문) Extract important words from the article and connect with similar image		
지도교수 의견	해당 과제는 학부생의 수준으로 높은 정확률을 얻기에는 한계가 있는 과제였으나, 다양한 아이디어를 알고리즘으로 구현하여 적정한 결과를 얻어서, 원래의 계획대로 인문사회과학기술융합학회의 하계학술대회 논문으로 게재 확정되었음.		

1. 도전 과제의 목표

소설, 문학과 같은 글에서 그림 또는 이미지는 극히 일부에 불과하거나 아예 없는 경우가 많다. 이는 독자의 상상력을 증진시키며 오히려 긍정적인 효과를 낼 수도 있지만 자신만의 상상이기에 달린 그림 밖에 떠오르지 않는다. 또한 글에 관련된 지식이 미리 없다면 상상조차도 할 수 없는 애매한 표현들이 많기에 이에 글을 알맞은 단위로 끊어 프로그램으로 주요 단어들을 조합해 이와 관련된 이미지를 보여 주면서 상상을 도와주는 프로그램을 만든다.

2. 도전 과제 내용



[그림 1] 이미지 검색 시스템 절차

해당하는 글의 주요 단어들을 분석해 아주 밀접한 이미지 또는 그림을 연결해주어 상상하기 힘들거나 윤곽조차 잡히지 않는 내용들에 대해 이해를 돕는 것으로 아래의 과정을 통한다.

과정 1. 입력으로 들어온 글을 문단 단위로 분할하여 나눈다.

과정 2. 문단단위로 나뉜 글들을 각각 KoNLPy라는 파이썬 패키지로 형태소 분석을 하고 원하는 명사만을 추출해 내며 불용어를 삭제한다.

과정 3. 포털 이미지검색 에 쓰일 주요 단어를 가려내기 위해 TF-IDF 알고리즘을 통해 문단별 각 명사 단어들의 중요도를 판별한다.

과정 4. 과정 3.에서의 중요도 수치를 기반으로 주요 문장을 선택한다.

과정 5. 과정 4.에서의 선택된 주요 문장에서 검색에 쓰일 주요 단어들을 골라낸다.

과정 6. 과정 5.에서의 추려내진 단어들로 포털사이트에서 검색한다.

과정 7. 6번 과정에서 검색 되어진 이미지들 중 몇 개의 이미지를 어떤 방식으로 반환할지 처리한다.

과정 8. 마지막으로 1번 과정에서 나뉜 문단단위 글들과 각 문단에 대한 이미지를 매칭시켜 하나의 프로그램으로써 시각적으로 보여준다.

3. 도전 과제의 성과

나 개인으로서의 성과로는 부분 부분의 작은 프로그램이 아니라 하나의 큰 프로젝트로써 아이디어 구상, 계획, 구현, 수정, 보완하는 과정을 겪었으며 이를 통해 프로그래밍에 있어서 동료와의 협업이 왜 필요한지를 아이디어 부분에서 스스로의 달린 사고만 계속 하게 되면서 지독하게 느꼈으며 도전학기제 기간 동안 프로그램을 만들 때 계속 허점이 드러난 프로그램 코드의 최적화(간결성, 가독성), 업데이트 시 버전관리, DB관리, 웹 앱 또는 모바일 앱의 구현, 전혀 다른 종류의 프로그래밍 언어 및 플랫폼간의 호환, 프로그램 보안성(SQL injection, DDos), 정해진 프로젝트 기간 내에서 프로젝트의 각 부분을 알맞게 시간분배 하여 일정을 조절하며 최적의 결과물을 도출하는 과정. 그리고 담당 교수님과 의 짧은 면담을 통해 프로젝트가 잘못된 길로 나아가고 있다는 것을 짚어주시고 고쳐나가며 방법론을 찾는 과정 및 보완하는 과정을 중간 중간 핵심을 조언해주셔서 완전히 엉뚱한 방향으로 빠지는 것을 면했다. 이로 인해 프로그래머로서 많은 부족함을 느끼며 알게 되었다.

도전학기제 자체로서의 성과로는 프로젝트 프로그램 자체의 코드와 6월 1일의 (사)인문사회과학기술융합학회의 하계학술대회 논문 제출이며 이후 7월20일에 기존 논문의 확장을 거쳐 저널 논문용 논문으로 제출을 하게 되었다.

4. 자기평가

도전학기제 중반까지 목표했던 계획대로 따라가며 프로그램을 구현해 나갔으나 학기 후반에 시간 분배에 실패하여 원하던 목표 완성도에 도달하지 못했다.

5. 최종 결과물

결과물 1.

Convergence Research Letter
Vol.1, No.3, June (2018), pp.

단어 및 문장 중요도에 기반한 연관 이미지 검색

Related Image Retrieval based on Word and Sentence Importance

요약

텍스트로만 구성된 문서의 이해도를 높이고, 흥미를 더하기 위해서 텍스트의 내용을 분석하여 중요한 내용을 포함하고 있는 문장을 선택하고, 이 문장을 가장 잘 표현할 수 있는 이미지들 웹에서 자동 검색하여 제시하는 시스템을 구현하고자 한다. 이를 구현을 위해서 TF-IDF에 기반하여 문단 내에서 중요한 문장을 선택하는 방법과 문장 내에서 중요한 단어를 선택하는 방법을 제시한다.

핵심어 : 단어 중요도, 문장 중요도, TF-IDF, 이미지 검색, 텍스트랭크

Abstract

In order to increase understanding and be of interest to readers in reading text-only documents, we propose a system that analyze the contents of the text to select important sentences, and automatically search the web for images that best represent these sentences. To implement this system, we also present methods of selecting important sentences based on TF-IDF within paragraphs, and how to choose important words in a sentence.

Keywords : Word Importance, Sentence Importance, TF-IDF, Image Retrieval, TextRank

1. 서론

어린 아이들이 독서를 할 때, 텍스트만으로 구성된 글을 읽게 되면 쉽게 지루해지고, 또 어려운 단어가 등장할 때에는 글의 의미를 잘 이해하지 못하는 어려움이 발생한다. 본 연구에서는 이러한 문제를 해결하기 위하여 텍스트의 내용에 적합한 그림 이미지들 웹에서 자동으로 검색하여 보여주

¹38403, Dept. Computer Engineering, Daegu Univ., Daegu-daero, Gyeongsan-si, Gyeongsangbuk-do, Republic of Korea
email:

²(Corresponding author) 38403, Dept. Computer Engineering, Daegu Univ., Daegu-daero, Gyeongsan-si, Gyeongsangbuk-do, Republic of Korea
email:

ISSN: 2384-0170 KJLI
Copyright © 2018 KJLI

1

[그림 2] (사)인문사회과학기술융합학회 하계학술대회 논문

결과물 2.

단어 선택 문장 선택	TF-IDF 상위 3개 단어	TF-IDF 상위 4개 단어	TF-IDF 상위 5개 단어	최상위 TF-IDF 주변단어 1개	최상위 TF-IDF 주변단어 2개
문장 내 최상위 TF-IDF 단어가 1순위	50%	40%	40%	40%	40%
문장 내 TF-IDF 상위 3개 단어를 동일한 가중치	45%	30%	35%	45%	45%
TextRank 알고리즘	30%	25%	25%	40%	40%

[표 1] 단어 및 문장 선택 알고리즘에 따른 이미지 정확률

결과물 3.

글 정리.py , 이미지 크롤러.py , Django\myweb*

(.py 는 파이썬 파일이며 Django\myweb 은 Django 기반 웹 프레임워크 파일이다)

* 단 프로그램의 동작 과정을 웹 앱을 통해 눈으로 쉽게 보여주기 위한 Django 파일은 아직 미완성이다.

결과물(예정) 4.

7월 20일 저널 발간용 최종 논문 제출 및 심사
(등재지 게재 or 탈락)

참고문헌

- [1] Yoon-Sung cho and Ki-Hong Ahan and Su-Kyoung Kim, Editors. A Study on Key Sentence Extraction by using Enriched TextRank. Proceedings of KIIT Summer Conference,(2013) 5, pp. 39-44
- [2] Han-mook Yoo and Han-joon Kim and Jae-young Chang, Automatic Construction of Reduced Dimensional Cluster-based Keyword Association Networks using LSI. Journal of KIIE. (2017), Vol.44, No.11, pp.1236-1243.
- [3] Jeong-Hur and Myung-Gill Jang, Editors. Homonym Disambiguation using Sense-Tagged Compound Noun Dictionary. Proceedings of KIIT fall Conference,(2013) Vol32, No.2 pp. 538~540pp
- [4] Hyun-woong Jang and Soo-sun Cho, Automatic Tagging for Social Images using Convolution Neural Networks. Journal of KIIE. (2016), Vol.43, NO.1. pp. 47~53
- [5] 국립국어원 언어정보나눔터, <https://ithub.korean.go.kr>