|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | |  | | **제12회 공개SW 개발자 대회 개발 계획서** | |  | |

**□ 참가팀 개요**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **구 분** | | **세부내용** | | | |
| **팀 명** | | 95Avenue | | | **총 인원 ( 4 명)** |
| **팀 구 성** | | **성명** | **소속** | **부서/학과** | **직위/학년** |
| **팀 장** | | 김태진 | 광운대학교 | 컴퓨터공학과 | 4학년 |
| **참가**  **지원** | **부문** |  | |  | |
| **자유과제** |  |  |  |  |
|  |  |  |  |
| **지정과제** |  | |  |  |
| ※ 지정과제 및 자유과제 중복 지원 가능(단, 중복 수상은 제한) | | | |
|  | |  | | | |
| **□ 출품작 요약** | | | | | |
| 사람들이 필요한 정보를 검색하게 되면, 회사의 협찬을 받아 작성된 광고 글이 포털 사이트에 노출되게 된다. 이러한 광고 글들의 양이 많아지게 되면, 사용자는 필요한 정보를 얻는 것에 어려움을 겪게 된다. 검색 포털 사이트에서 광고를 필터링하여도, 최근에는 문구를 이미지화하여 필터를 피해 가는 등의 여러 방법으로 상위에 노출되도록 광고 글을 작성한다.  이러한 경험에서 아이디어를 얻어, 사용자가 브라우저를 통해 포털 사이트에서 검색했을 때, 글의 내용과 삽입된 이미지에 집중하여 해당 글이 광고 글인지 판단을 하는 프로그램을 계획하게 되었다.  브라우저 검색 API를 이용하여 특정 키워드에 대한 블로그 검색 결과를 수집하고, 블로그 내의 본문 텍스트와 이미지를 수집한다. 블로그 본문 내용은 트레이닝 데이터로 이용한다. 분류 결과의 정답 데이터를 이용하여 데이터 셋을 만들고, 해당 데이터 셋은 gensim 모델의 doc2vec을 레이블링 된 텍스트 데이터들로 새로운 텍스트 데이터가 광고 글인지를 판단하는 데 사용한다.  해당 블로그 글이 광고 글인지 판단하는 방법은 블로그 내 이미지를 OpenCV와 Tesseract-OCR엔진을 통해 이미지에서 텍스트를 추출하여 재가공하는 방법을 사용한다. ‘협찬’, ‘원고료’와 같은 대가성 광고 글이라는 것을 암시하는 텍스트가 존재한다면 광고라고 판단한다. 광고 글을 암시하는 텍스트가 존재하지 않더라도, 여러 블로그에서 사용한 이미지가 같거나 추출한 텍스트 내용을 텍스트 데이터로 읽어, 광고 글인지 아닌지 판단한다. | | | | | |