**2021년도 물결21 말뭉치 정비 가이드라인**

**1. 작업 정의**

* 목적:
  + 물결21 말뭉치 구축 과정은 자료의 ‘수집’, ‘정비’, ‘형태분석’, ‘통계적 분석’, ‘배포’로 진행된다.
  + 본 작업은 ‘정비’ 단계에 포함되는 작업으로 신문에 포함된 불필요한 텍스트를 제거하고, 기사의 장르를 구분하는 것을 목표로 한다.
* 작업 개요:
  + 비기사 분류:

연작 소설, 인사 소식, 운동 경기 스코어 등 기사가 아니거나 기사로서의 가치가 떨어지는 텍스트를 분류한다.

* + 중복기사 분류:

수집된 신문 데이터 중에는 내용이 중복되는 기사가 있다. 판(version) 갱신에 의해 이전 판의 기사가 남아 있는 것으로, 최종 판을 제외한 나머지 판들을 분류한다.

* + 기사 내 주석 및 마크업 제거:

기사 본문에 작성자/편집자의 주석이나 서식 마크업이 남아 있는 경우가 있다. 실제 기사로 발행되지 않는 내용들로, 기사 본문에서 이러한 문자열을 찾아 거한다.

- 사설 분류:

기사의 장르는 사설과 사설이 아닌 스트레이트기사로 구분한다. 사설인 기사를 찾아 별도로 분류한다.

**2. 비기사 분류**

* 비기사의 유형

단신, 광고성 기사, 사진기사, 부고, 인사, 동정, 경기일정, 신춘문예, 연작 소설, 일기예보, 외국어 학습 등의 일일 정보, 책의 출판사와 가격 등이 연속하는 기사(문화면), 논술문제 및 답안 등

* 비기사 텍스트 제목의 신문사별 특징
  + 동아일보

|  |  |
| --- | --- |
| **분류** | **제목에 포함된 키워드** |
| 스포츠 | [브리핑룸] [200자 브리핑] [200자 뉴스] |
| 문화 | [연재소설] [이러쿵 저러쿵] [신춘문예] [건강브리핑] [건강 단신] [의료계 단신] [학술 단신] [미술 단신] |
| 경제 | [경제계 인사] [생활게시판] [고용] [분양정보] [주총소식] [유통가소식] [비즈파일] [비즈브리핑] [브리핑룸] [금융정보] [증권정보] [창업/단신] [업계소식] [이주의 추천상품] [알뜰장바구니] [생활게시판] |
| 지역 | [수도권/메트로파일] [광주/내고장 인사]광주시 [인천/경기]게시판 [충청/강원]문화가 [인천/경기]단신 [뉴스브리핑] |
| 투데이, 종합 | [부고] [부음] [새의자] [인사] [동정] [인물동정] [알림] [모임] [본사] [본사손님] [본사내방] [사고] [社告] |
| 사회, 종합 | [기상/기후] [오늘의 날씨] [날씨와 과학] [날씨 小史] [날씨소사] |
|  | [알려드립니다] [바로잡습니다] [단신] |

* + 조선일보

|  |
| --- |
| **제목에 포함된 키워드** |
| ․ [사진]  ․ [정정]  ․ [부음(부고)] [사고(社告)] [인사] [알림]  ․ [단신(경제, 학술, 스포츠 등)]  ․ [생활한자] [생활영어(일어/중어 등)]  ․ [오늘의 운세] [오늘의 날씨]  ․ [연재만화]  ․ [주식시세표] [스포츠 대진표] [방송편성표]  ․ [신춘문예] 등 |

* + 한겨레

|  |
| --- |
| **제목에 포함된 키워드** |
| ․ 바로잡습니다, 반론싣습니다, 인사, 동정, 궂긴소식, 알림, 단신, 게시판, 기록실, 뉴스 퍼즐, 토막 뉴스  ․ 주요 기사안내(다른 면에 있는 기사내용 요약 소개, 페이지 안내)  ․ 볼만한 프로, 오늘의 프로  ․ 오늘의 경기, 내일의 경기, 그린 풍향계  ․ 새 DVD, 새 책, 새 상품, 새로 나왔어요, 화제의 전시, 무대 아래  ․ 여행 정보, 함께 떠나요, 함께 가봐요, 가보고 싶어요  ․ 재테크 플라자, 금융 백화점, 금융 플라자, 금융 수첩, 부동산 파일, 부동산 캘린더, 오늘의 증시  ․ NGO 게시판, 이번주의 NGO  ․ 쪽지, 여성계 쪽지, 여행 쪽지, 여가 쪽지  ․ 정보 꾸러미, 대학가 소식, 나눔 소식, 짧은 소식, 토막 소식, 호텔 소식, 디지털 장터, 건강 쌈지, 교육 마당 |

**2. 사설 분류**