문화체육관광부
국립국어원

국립국어원 비출판물 말뭉치

(버전 1.2)

· **자료명:** 국립국어원 비출판물 말뭉치

· **공개일**
 · (버전 1.0) 2021. 8. 30.
 · (버전 1.1) 2021. 12. 1.
  - 일부 태그 오류 수정
 · (버전 1.2) 2022. 11. 4.
  - 일부 태그 오류 수정

· **자료 유형:** 텍스트

· **관련 사업:** 개인적 글쓰기 자료 수집 및 말뭉치 지식 강연회 개최(2019)

· **자료 설명**
 · 내용
  - 교정을 거치지 않았으며 어디에도 공개(출판, 본인 및 타인 블로그 게시 등 포함)되지 않은 글
  - 글 종류: 시(동시), 일기, 편지 글, 소설(동화), 감상문, 기타
  - 초등학생부터 80대 성인까지 총 5,937명의 자료
  - 초등학생은 손 글씨 원자료(PDF, 이미지 파일)를 받아 전문 요원이 글을 입력함.

 · 분량
  - 총 210만 어절

 · 파일 형식: xml(UTF-8 인코딩)

 · 파일 수 및 크기: 파일 10,753개, 총 28MB

· **인용:** 국립국어원(2021). 국립국어원 비출판물 말뭉치(버전 1.0).
 URL: https://corpus.korean.go.kr

· **예시**

```
<?xml version="1.0" encoding="UTF-8"?>
<SJML>
    <header>
        <fileInfo>
            <fileId>WDRW1900520500</fileId>
            <annoLevel>원시</annoLevel>
            <category>비출판물 > 편지글</category>
        </fileInfo>
        <sourceInfo>
            <title>제목없음</title>
            <author id="P20500" age="9" occupation="초등학생" sex="M" submission="오프라인" handwriting="Yes">신재혁</author>
        </sourceInfo>
    </header>
    <text date="20190000" subclass="영화">
        <p>이모에게 어제 나랑</p>
        <p>영화 봤잖아 그 영화 이모가 추천해서 봤는데 난 그 영화가 내 취향에는 맞지않았어</p>
        <p>난 왜냐 하냐면 난 액션 영화가 추리하는 영화 보다 액션 영화가</p>
        <p>더 좋아 그래도 재미 있었어^^.</p>
        <p>그리고 이모가 말했더 갈릭 팝콘이 방구냄세 난 다고 했잖아 ㅋㅋㅋ 다음에는 ~~~~ 갈릭팝콘방구 냄새나니까 주문하지 말자!</p>
        <p>그리고 내가 말한액션 영화!</p>
        <p>봐줄 거지 이모 그럼 언제 보로 갈가</p>
        <p>그럼 재미 있는 영화 나랑 보러가자</p>
        <p>이모 그럼 인기 ,액션영화</p>
        <p>보자고 그것도 다음에 정해보자</p>
        <p>난 미션 임파 서블</p>
        <p>갓은 것도 괜찮고 아니면</p>
        <p>이모가 생각 하는 영화 있어?</p>
        <p>아니면 내가 생각 해 볼 게</p>
        <p>이모 안녕</p>
    </text>
</SJML>
```

* fildId(WDRW190XXYYYYY): XX(글 식별 번호), YYYYY(저자 식별 번호)
author: 저자명 비공개를 원하는 경우 [개인글작성자]
text date: 글 작성일, 정확한 날짜를 모를 경우 0000
subclass: 글감, 저자가 글감을 기입하지 않은 경우 [null_글감]으로 검수자가 기입
본문 개인 정보 비식별화: 이름 - &name1&, &name2&, ...
... : 줄 바꿈(category가 '시'인 경우: ... - 연, <s>...</s> - 행)

· **자료 내용 문의:** 02-2669-9607