문화체육관광부
국립국어원

국립국어원 온라인 게시 자료 말뭉치 2022

(버전 1.0)

· **자료명:** 국립국어원 온라인 게시 자료 말뭉치 2022

· **공개일**

· (버전 1.0) 2023. 6. 30.

· **자료 유형:** 텍스트

· **관련 사업:** 온라인 게시 자료 수집 및 정제(2022)

· **자료 설명**

※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > '2022년 온라인 게시 자료 수집 및 정제' 사업 보고서 참고

· 내용

- 게시판, 누리 소통망 등에서 수집한 언어 자료로 구성한 말뭉치

종류	매체 이름
게시판	네이버 카페, 네이트판, 다음 카페, 디시인사이드, 루리웹, 보배드림, 뽐뿌, 에스엘알클럽, 에펨코리아, 엠엘비파크, 오늘의 유머, 웃긴대학, 인벤, 클리앙
누리 소통망	인스타그램, 페이스북

· 분량

- 게시판: 8,282건
- 누리 소통망: 296,892건

· 파일 형식: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 파일 763개, 총 123MB(ZIP 파일 기준)

· **인용:**

- **(국문)** 국립국어원(2023). 국립국어원 온라인 게시 자료 말뭉치 2022(버전 1.0). URL: https://kli.korean.go.kr/corpus
- **(영문)** National Institute of Korean Language (2023). NIKL Online Posting Materials Corpus 2022(v.1.0). URL: https://kli.korean.go.kr/corpus

· 파일 명명 규칙

자리	1	2	3	4	5	6	7	8	9	10	11	12	13	14
속성	매체	장르	주석 단계		구축 연도		일련번호(8자리)							
정의값	E: 웹	P: 게시판 S: 누리 소통망	RW: 원시 말뭉치		22: 2022년		00000001 ~ 99999999 (여덟 자리 일련번호)							
※ 예시: EPRW2200000010.json 2022년에 구축한 게시판 자료 원시 말뭉치 파일 ESRW2200000710.json 2022년에 구축한 누리 소통망 자료 원시 말뭉치 파일														

· 예시


```
{
    "id": "EPRW2200000752.860",
    "metadata": {
        "title": "중독성이 상당하네요",
        "author": "writer752",
        "publisher": "디시인사이드",
        "date": "20220612",
        "topic": "문화/예술_영화/드라마/방송",
        "crawl_date": "20221005 00:10:00",
        "url": "https://gall.dcinside.com/board/view/?id=haebang&no=29937&page=8"
    },
    "paragraph": [
        {
            "id": "EPRW2200000752.860.1",
            "form": "시작부터 느낌이 좋았습니다.",
            "original_form": "시작부터 느낌이 좋았습니다."
        },
        {
            "id": "EPRW2200000752.860.2",
            "form": "따뜻한 느낌, 별거없고 적막한 시골인데 계속 보고 싶게 만드는 드라마네요. 분위기가 제일 마음에 들었습니다.",
            "original_form": "따뜻한 느낌, 별거없고 적막한 시골인데 계속 보고 싶게 만드는 드라마네요. 분위기가 제일 마음에 들었습니다."
        },
        {
            "id": "EPRW2200000752.860.3",
            "form": "각자의 이야기들로 이루어진 점도 좋았습니다.",
            "original_form": "각자의 이야기들로 이루어진 점도 좋았습니다."
        },
        {
            "id": "EPRW2200000752.860.4",
            "form": "복수를 위해 가족전체가 움직여서 복수하는둥, 남의 애인한테 가서 이래라저
```

래라하는 그런 쓸데없는 전개도 없다는 점.",
"original_form": "복수를 위해 가족전체가 움직여서 복수하는둥, 남의 애인한테 가서 이래라저래라하는 그런 쓸데없는 전개도 없다는 점."
},
{
"id": "EPRW2200000752.860.5",
"form": "오히려 각자의 이야기에 더 집중할 수 있어서 좋았습니다.",
"original_form": "오히려 각자의 이야기에 더 집중할 수 있어서 좋았습니다."
},
{
"id": "EPRW2200000752.860.6",
"form": "결말부분은 마음이 멍해지는 느낌이였습니다.",
"original_form": "결말부분은 마음이 멍해지는 느낌이였습니다."
},
{
"id": "EPRW2200000752.860.7",
"form": "선택지를 주고 알아서 상상하는 결말이 아닌",
"original_form": "선택지를 주고 알아서 상상하는 결말이 아닌"
},
{
"id": "EPRW2200000752.860.8",
"form": "대회처럼 '해방'을 주제로 결말을 알아서 생각하라는 붕 뜬 결말 같았습니다. 그런데 희한하게 그게 여운이 길게 가네요.",
"original_form": "대회처럼 '해방'을 주제로 결말을 알아서 생각하라는 붕 뜬 결말 같았습니다. 그런데 희한하게 그게 여운이 길게 가네요."
},
{
"id": "EPRW2200000752.860.9",
"form": "다른 결말이였어도 이렇게 마음과 생각이 차분해지고 여러 생각이 들었을지 잘 모르겠네요.",
"original_form": "다른 결말이였어도 이렇게 마음과 생각이 차분해지고 여러 생각이 들었을지 잘 모르겠네요."
},
{
"id": "EPRW2200000752.860.10",
"form": "이런 분위기의 드라마나 영화 추천해주시면 감사하겠습니다.",
"original_form": "이런 분위기의 드라마나 영화 추천해주시면 감사하겠습니다."
}

※ "original_form": 수집한 언어 자료의 원문을 그대로 유지한 형태(개인 정보 등은 비식별화)
"form": 원문에서 연속된 여러 개의 공백(스페이스, 탭 등), 비식별화 기호 등을 제거하여 전처리한 형태

· **자료 내용 문의:** 02-2669-9638