문화체육관광부
국립국어원

국립국어원 개체명 분석 말뭉치 2021

(버전 1.0)

· **자료명:** 국립국어원 개체명 분석 말뭉치 2021

· **공개일**

· (버전 1.0) 2022. 4. 1.

· **자료 유형:** 텍스트

· **관련 사업:** 2021년 개체명 분석 및 개체 연결 말뭉치 연구 분석(2021)

· **자료 설명**

※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > '2021년 개체명 분석 및 개체 연결 말뭉치 연구 분석' 사업 보고서 참고

· 내용

- 개체명의 경계를 인식하고 150개 세부 의미 분류 체계에 따른 표지를 부착한 말뭉치임.
- 한국전자통신연구원(ETRI)의 '세부분류 개체명 가이드라인 2018'(2018. 12.)을 기본으로 하여 국립국어원에서 수정한 '2021년 개체명 분석 말뭉치 구축 지침 ver. 2.1.'('2021년 개체명 분석 및 개체 연결 말뭉치 연구 분석' 보고서, 2021)에 준하여 분석하였음.
- 15개 대분류 의미 분류 체계에 따라 분석한 '국립국어원 개체명 분석 말뭉치(버전 1.0)'(국립국어원 2020, 문어 200만 어절, 구어 100만 어절)에 대해 세분류로 다시 분석함.
- 2019년, 2020년에 구축한 국립국어원 신문 말뭉치 중 "IT/과학", "문화", "건강" 분야에서 100만 어절을 새로이 추출하여 세분류로 분석함.
- 2019년에 구축한 구어 말뭉치 중 "IT/과학", "문화", "건강" 관련 분야에서 200만 어절을 새로이 추출하여 세분류로 분석함.

· 분석 표지: 150개

대분류	세분류
1. PERSON(PS)	PS_NAME, PS_CHARACTER, PS_PET
2. STUDY_FIELD(FD)	FD_SCIENCE, FD_SOCIAL_SCIENCE, FD_MEDICINE, FD_ART, FD_HUMANITIES, FD_OTHERS
3. THEORY(TR)	TR_SCIENCE, TR_SOCIAL_SCIENCE, TR_MEDICINE, TR_ART, TR_HUMANITIES, TR_OTHERS
4. ARTIFACTS(AF)	AF_BUILDING, AF_CULTURAL_ASSET, AF_ROAD, AF_TRANSPORT, AF_MUSICAL_INSTRUMENT,

	AF_WEAPON, AFA_DOCUMENT, AFA_PERFORMANCE, AFA_VIDEO, AFA_ART_CRAFT, AFA_MUSIC, AFW_SERVICE_PRODUCTS, AFW_OTHER_PRODUCTS
5. ORGANIZATION(OG)	OGG_ECONOMY, OGG_EDUCATION, OGG_MILITARY, OGG_MEDIA, OGG_SPORTS, OGG_ART, OGG_MEDICINE, OGG_RELIGION, OGG_SCIENCE, OGG_LIBRARY, OGG_LAW, OGG_POLITICS, OGG_FOOD, OGG_HOTEL, OGG_OTHERS
6. LOCATION(LC)	LCP_COUNTRY, LCP_PROVINCE, LCP_COUNTY, LCP_CITY, LCP_CAPITALCITY, LCG_RIVER, LCG_OCEAN, LCG_BAY, LCG_MOUNTAIN, LCG_ISLAND, LCG_CONTINENT, LC_SPACE, LC_OTHERS
7. CIVILIZATION(CV)	CV_CULTURE, CV_TRIBE, CV_LANGUAGE, CV_POLICY, CV_LAW, CV_CURRENCY, CV_TAX, CV_FUNDS, CV_ART, CV_SPORTS, CV_SPORTS_POSITION, CV_SPORTS_INST, CV_PRIZE, CV_RELATION, CV_OCCUPATION, CV_POSITION, CV_FOOD, CV_DRINK, CV_FOOD_STYLE, CV_CLOTHING, CV_BUILDING_TYPE
8. DATE(DT)	DT_DURATION, DT_DAY, DT_WEEK, DT_MONTH, DT_YEAR, DT_SEASON, DT_GEOAGE, DT_DYNASTY, DT_OTHERS
9. TIME(TI)	TI_DURATION, TI_HOUR, TI_MINUTE, TI_SECOND, TI_OTHERS
10. QUANTITY(QT)	QT_AGE, QT_SIZE, QT_LENGTH, QT_COUNT, QT_MAN_COUNT, QT_WEIGHT, QT_PERCENTAGE, QT_SPEED, QT_TEMPERATURE, QT_VOLUME, QT_ORDER, QT_PRICE, QT_PHONE, QT_SPORTS, QT_CHANNEL, QT_ALBUM, QT_ADDRESS, QT_OTHERS
11. EVENT(EV)	EV_ACTIVITY, EV_WAR_REVOLUTION, EV_SPORTS, EV_FESTIVAL, EV_OTHERS
12. ANIMAL(AM)	AM_INSECT, AM_BIRD, AM_FISH, AM_MAMMALIA, AM_AMPHIBIA, AM_REPTILIA, AM_TYPE, AM_PART, AM_OTHERS
13. PLANT(PT)	PT_FRUIT, PT_FLOWER, PT_TREE, PT_GRASS, PT_TYPE, PT_PART, PT_OTHERS
14. MATERIAL(MT)	MT_ELEMENT, MT_METAL, MT_ROCK, MT_CHEMICAL
15. TERM(TM)	TM_COLOR, TM_DIRECTION, TM_CLIMATE, TM_SHAPE, TM_CELL_TISSUE_ORGAN, TMM_DISEASE, TMM_DRUG, TMI_HW, TMI_SW, TMI_SITE, TMI_EMAIL, TMI_MODEL, TMI_SERVICE, TMI_PROJECT, TMIG_GENRE, TM_SPORTS

· 분량

총 약 600만 어절(문어 300만, 구어 300만 어절)

- 2019년 대분류 개체명 분석 말뭉치 세분화: 300만 어절(문어 200만, 구어 100만 어절)
- 2021년 신규 추출: 300만 어절(문어 100만, 구어 200만 어절)

· 파일 형식: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 파일 273개, 총 105MB(ZIP 파일 기준)

· 인용:

(국문) 국립국어원(2022). 국립국어원 개체명 분석 말뭉치 2021(버전 1.0). URL: https://kli.korean.go.kr/corpus

(영문) National Institute of Korean Language (2022). NIKL Named Entity Corpus 2021 (v.1.0). URL: https://kli.korean.go.kr/corpus

· 파일 명명 규칙

<table>
<tr><th>자리</th><th>1</th><th colspan="2">2</th><th>3 4</th><th>5 6</th><th colspan="8">7 8 9 10 11 12 13 14</th></tr>
<tr><td>속성</td><td>유형</td><td colspan="2">구성 방법</td><td>주석 단계</td><td>구축 연도</td><td colspan="8">일련번호(8자리)</td></tr>
<tr><td rowspan="2">정의 값</td><td>N: 신문</td><td rowspan="2">X: 추출</td><td></td><td rowspan="2">NE: 개체명</td><td rowspan="2">21: 2021년</td><td colspan="8" rowspan="2">00000001 ~ 99999999
(여덟 자리 일련번호)</td></tr>
<tr><td>S: 구어</td><td>A: 공적 독백
B: 공적 대화</td></tr>
<tr><td colspan="14">※ 예시: NXNE2102203310.json 2021년에 구축한 신문 자료 가공 개체명 말뭉치 파일
SANE2100000895.json 2021년에 구축한 구어-공적 독백 자료 가공 개체명 말뭉치 파일</td></tr>
</table>

· 예시


```
{
    "id": "SARW1900000895.1.1.24",
    "form": "그래서 지금 제가 동영상을 하나 갖고 왔는데 그 고대 그리스의 천동설이 예측하는 행성의 움직임이 어떤 것인지를 보여주는 동영상이에요.",
    "word": [
        {
            "id": 1,
            "form": "그래서",
            "begin": 0,
            "end": 3
        },
        {
```

```
            "id": 2,
            "form": "지금",
            "begin": 4,
            "end": 6
        },
        {
            "id": 3,
            "form": "제가",
            "begin": 7,
            "end": 9
        },
        {
            "id": 4,
            "form": "동영상을",
            "begin": 10,
            "end": 14
        },
        {
            "id": 5,
            "form": "하나",
            "begin": 15,
            "end": 17
        },
        {
            "id": 6,
            "form": "갖고",
            "begin": 18,
            "end": 20
        },
        {
            "id": 7,
            "form": "왔는데",
            "begin": 21,
            "end": 24
        },
        {
            "id": 8,
            "form": "그",
            "begin": 25,
            "end": 26
        },
        {
            "id": 9,
            "form": "고대",
            "begin": 27,
            "end": 29
        },
        {
            "id": 10,
            "form": "그리스의",
```

```
            "begin": 30,
            "end": 34
        },
        {
            "id": 11,
            "form": "천동설이",
            "begin": 35,
            "end": 39
        },
        {
            "id": 12,
            "form": "예측하는",
            "begin": 40,
            "end": 44
        },
        {
            "id": 13,
            "form": "행성의",
            "begin": 45,
            "end": 48
        },
        {
            "id": 14,
            "form": "움직임이",
            "begin": 49,
            "end": 53
        },
        {
            "id": 15,
            "form": "어떤",
            "begin": 54,
            "end": 56
        },
        {
            "id": 16,
            "form": "것인지를",
            "begin": 57,
            "end": 61
        },
        {
            "id": 17,
            "form": "보여주는",
            "begin": 62,
            "end": 66
        },
        {
            "id": 18,
            "form": "동영상이에요.",
            "begin": 67,
            "end": 74
```

```
                    }
                ],
                "NE": [
                    {
                        "id": 1,
                        "form": "하나",
                        "label": "QT_COUNT",
                        "begin": 15,
                        "end": 17
                    },
                    {
                        "id": 2,
                        "form": "고대",
                        "label": "DT_DYNASTY",
                        "begin": 27,
                        "end": 29
                    },
                    {
                        "id": 3,
                        "form": "그리스",
                        "label": "LCP_COUNTRY",
                        "begin": 30,
                        "end": 33
                    },
                    {
                        "id": 4,
                        "form": "천동설",
                        "label": "TR_SCIENCE",
                        "begin": 35,
                        "end": 38
                    }
                ]
            },
```


· **자료 내용 문의:** 02-2669-9638