문화체육관광부
국립국어원

국립국어원 개체명 분석 말뭉치 2020

(버전 2.1)

· **자료명:** 국립국어원 개체명 분석 말뭉치 2020

· **공개일**

· (버전 1.0) 2021. 5. 31.

· (버전 2.0) 2022. 3. 31.

- 2021년 개체명 분석 지침 ver. 2.1.에 따라 일부 태그 분석 수정
- 부적절한 내용 포함 문서 삭제

· (버전 2.1) 2022. 9. 16.

- 부적절한 내용 포함 문서 삭제

· **자료 유형:** 텍스트

· **관련 사업:** 2020년 개체명 말뭉치 연구 분석(2020)

· **자료 설명**

※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > '2020년 개체명 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 '2020년 개체명 분석 말뭉치 구축 지침 ver. 1.6.'('2020년 개체명 말뭉치 연구 분석' 보고서, 2020)에 준하여 분석하였음.
- '2021년 개체명 분석 말뭉치 구축 지침 ver. 2.1.'('2021년 개체명 분석 및 개체 연결 말뭉치 연구 분석' 보고서, 2021)에 준하여 버전 1.0 분석 결과를 수정하였음.

· 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

- 약 500만 어절(웹 문서 대상)

· 파일 형식: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 파일 50개, 총 118MB(ZIP 파일 기준)

· 파일 명명 규칙

자리	1	2	3	4	5	6	7	8	9	10	11	12	13	14
속성	유형	구성 방법	주석 단계		최초 배포 연도		일련번호(8자리)							
정의 값	E: 웹	X: 추출	NE: 개체명		21: 2021년		00000001 ~ 99999999 (여덟 자리 일련번호)							
※ 예시: EXNE2105170001.json 2021년에 최초 배포한 웹 자료 가공 개체명 말뭉치 파일														

· 인용:

- **(국문)** 국립국어원(2022). 국립국어원 개체명 분석 말뭉치 2020(버전 2.1). URL: https://kli.korean.go.kr/corpus
- **(영문)** National Institute of Korean Language (2022). NIKL Named Entity Corpus 2020 (v.2.1). URL: https://kli.korean.go.kr/corpus

· 예시


```
{
    "id": "EBRW1908000115.3.2.1",
    "form": "스노우(SNOW)라는 카메라 스티커 어플이 있는데 거기에 고양이 인식이 가능한 스티커가 있다는 것이다.",
    "word": [
        {
            "id": 1,
            "form": "스노우(SNOW)라는",
            "begin": 0,
            "end": 11
        },
        {
            "id": 2,
            "form": "카메라",
            "begin": 12,
            "end": 15
        },
        {
            "id": 3,
            "form": "스티커",
            "begin": 16,
            "end": 19
        },
```

```
            {
                "id": 4,
                "form": "어플이",
                "begin": 20,
                "end": 23
            },
            {
                "id": 5,
                "form": "있는데",
                "begin": 24,
                "end": 27
            },
            {
                "id": 6,
                "form": "거기에",
                "begin": 28,
                "end": 31
            },
            {
                "id": 7,
                "form": "고양이",
                "begin": 32,
                "end": 35
            },
            {
                "id": 8,
                "form": "인식이",
                "begin": 36,
                "end": 39
            },
            {
                "id": 9,
                "form": "가능한",
                "begin": 40,
                "end": 43
            },
            {
                "id": 10,
                "form": "스티커가",
                "begin": 44,
                "end": 48
            },
            {
                "id": 11,
                "form": "있다는",
                "begin": 49,
                "end": 52
            },
            {
                "id": 12,
                "form": "것이다.",
                "begin": 53,
                "end": 57
            }
        ],
        "NE": [
```

```
                {
                    "id": 1,
                    "form": "스노우",
                    "label": "TMI_SERVICE",
                    "begin": 0,
                    "end": 3
                },
                {
                    "id": 2,
                    "form": "SNOW",
                    "label": "TMI_SERVICE",
                    "begin": 4,
                    "end": 8
                },
                {
                    "id": 3,
                    "form": "카메라",
                    "label": "TMI_HW",
                    "begin": 12,
                    "end": 15
                },
                {
                    "id": 4,
                    "form": "고양이",
                    "label": "AM_MAMMALIA",
                    "begin": 32,
                    "end": 35
                }
            ]
        },
```


· **자료 내용 문의:** 02-2669-9638