문화체육관광부
국립국어원

국립국어원 개체명 분석 말뭉치 2022

(버전 1.1)

· **자료명:** 국립국어원 개체명 분석 말뭉치 2022

· **공개일**
  · (버전 1.0) 2023. 6. 30.
  · (버전 1.1) 2023. 9. 27.

· **자료 유형:** 텍스트

· **관련 사업:** 2022년 말뭉치 개체명 분석 및 개체 연결(2022)

· **자료 설명**

※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > '2022년 말뭉치 개체명 분석 및 개체 연결' 사업 보고서 참고

· 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 500만 어절(신문 300만 어절, 일상 대화 100만 어절, 온라인 대화 100만 어절)

· 파일 형식: JSON(UTF-8 인코딩)

· 파일 수 및 크기: 파일 3개, 총 96.7MB(ZIP 파일 기준)

· 인용:

(국문) 국립국어원(2023). 국립국어원 개체명 분석 말뭉치 2022(버전 1.1). URL: https://kli.korean.go.kr/corpus

(영문) National Institute of Korean Language (2023). NIKL Named Entity Corpus 2022

(v.1.1). 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	유형	구성 방법	주석 단계		구축 연도		일련번호(8자리)							
정의값	N: 신문 S: 일상 대화 M: 온라인 대화	X: 추출	NE: 개체명		22: 2022년		00000001 ~ 99999999 (여덟 자리 일련번호)							
※ 예시: NXNE2202211217.json 2022년에 구축한 신문 자료 가공 개체명 말뭉치 파일 SXNE2202211217.json 2022년에 구축한 일상 대화 자료 가공 개체명 말뭉치 파일 MXNE2202211217.json 2022년에 구축한 온라인 대화 자료 가공 개체명 말뭉치 파일														

· **예시**


```
                    {
                "id": "MDRW2100010933.1.10",
                "form": "캐나다 동부쪽 몬트리올 퀘백 이쪽은 프랑스어 쓰더러구요 프랑스에서 이민을 많이 왔나봐요",
                "word": [
                    {
                        "id": 1,
                        "form": "캐나다",
                        "begin": 0,
                        "end": 3
                    },
                    {
                        "id": 2,
                        "form": "동부쪽",
                        "begin": 4,
                        "end": 7
                    },
                    {
                        "id": 3,
                        "form": "몬트리올",
                        "begin": 8,
                        "end": 12
                    },
                    {
                        "id": 4,
                        "form": "퀘백",
                        "begin": 13,
                        "end": 15
                    },
```

```
            {
                "id": 5,
                "form": "이쪽은",
                "begin": 16,
                "end": 19
            },
            {
                "id": 6,
                "form": "프랑스어",
                "begin": 20,
                "end": 24
            },
            {
                "id": 7,
                "form": "쓰더러구요",
                "begin": 25,
                "end": 30
            },
            {
                "id": 8,
                "form": "프랑스에서",
                "begin": 31,
                "end": 36
            },
            {
                "id": 9,
                "form": "이민을",
                "begin": 37,
                "end": 40
            },
            {
                "id": 10,
                "form": "많이",
                "begin": 41,
                "end": 43
            },
            {
                "id": 11,
                "form": "왔나봐요",
                "begin": 44,
                "end": 48
            }
        ],
        "NE": [
            {
                "id": 1,
                "form": "캐나다",
                "label": "LCP_COUNTRY",
                "begin": 0,
                "end": 3
```

```
                    },
                    {
                        "id": 2,
                        "form": "동부",
                        "label": "TM_DIRECTION",
                        "begin": 4,
                        "end": 6
                    },
                    {
                        "id": 3,
                        "form": "몬트리올",
                        "label": "LCP_CITY",
                        "begin": 8,
                        "end": 12
                    },
                    {
                        "id": 4,
                        "form": "퀘백",
                        "label": "LCP_CITY",
                        "begin": 13,
                        "end": 15
                    },
                    {
                        "id": 5,
                        "form": "프랑스어",
                        "label": "CV_LANGUAGE",
                        "begin": 20,
                        "end": 24
                    },
                    {
                        "id": 6,
                        "form": "프랑스",
                        "label": "LCP_COUNTRY",
                        "begin": 31,
                        "end": 34
                    }
                ]
            },
```


· **자료 내용 문의:** 02-2669-9638