For English, go to page 3

Norske idiom

Dette er eit datasett som består av 3537 norske idiom og fraser som finst fleire enn 100 gonger i Nettbiblioteket.
Det er 3455 idiom på bokmål og 88 på nynorsk. I framtida vil vi prøve å leggje til fleire nynorske idiom.

Idiomfrekvensar

Fila idiom_freqs/all_idioms_freq.json inneheld alle idioma og frekvensen deira i nettbiblioteket som par av (nykel, verdi).
Fila idiom_freqs/nno_idioms_freq.json inneheld dei nynorske idioma og frekvensen deira i nettbiblioteket som par av (nykel, verdi).
Fila idiom_freqs/nob_idioms_freq.json inneheld idioma på bokmål og frekvensen deira i nettbiblioteket som par av (nykel, verdi).

Idiomgrafar

Det er stort språkleg overlapp mellom idioma. Til dømes er det for dei 3537 unike idioma berre 796 unike startord.

Difor har vi ordna idioma som tre, der rotnodane er startorda, og ein finn idioma ved å fylgje ein sti frå ein rotnode til ein løvnode.

Døme:

```
"alt": {
    "er": {
        "bare": {
            "fryd": {
                "og": {
                    "gammen": {}
                }
            }
        },
        "klappet": {
            "og": {
                "klart": {}
            }
        },
        "såre": {
            "vel": {}
        }
    },
    "går": {
        "sin": {
```

```
            "vante": {
                "gang": {}
            }
        }
    },
}
```

Her er eit tre der "alt" er rota, og idioma "alt er bare fryd og gammen", "alt er klappet og klart", "alt er såre vel" og "alt går sin vante gang" kan utleiast ved å traversere treet. Desse trea ligg i idiom_graphs/.

Det finst også flate versjonar av trea (der sekvensar av nodar/ord med berre eitt barn blir limt saman)

Døme:

```
"alt": {
    "er": {
        "bare fryd og gammen": "",
        "klappet og klart": "",
        "såre vel": ""
    },
    "går sin vante gang": "",
}
```

Idiomfullføring som ei språkteknologioppgåve

Idioma er splitta i idiomstartar (dei fyrste N-1 orda) og godkjende idiomssluttar (ei liste med mogelege ord som kan fullføre idiomet).
Av dei 3245 radene er det 156 der det finst meir enn eitt rett svar.

Dette datasettet kan bli brukt til å måle ein generativ språkmodell si evne til å fullføre kjente idiom, eller som ei 'masked language modelling' oppgåve.

Datasettet kan bli lasta inn med datasets-biblioteket og ligg òg på huggingface.

Norwegian idioms

This is a dataset that consists of 3537 Norwegian idioms and phrases that appear more than 100 times in the online library of the National Library of Norway.
There are 3455 Bokmål idioms and 88 Nynorsk idioms. In the future, we want to add more Nynorsk idioms.

Idiom frequencies

idiom_freqs/all_idioms_freq.json contain all the idioms and their frequency in the online library as key,value pairs
idiom_freqs/nno_idioms_freq.json contain the Nynorsk idioms and their frequency in the online library as key,value pairs
idiom_freqs/nob_idioms_freq.json contain the Bokmål idioms and their frequency in the online library as key,value pairs

Idiom graphs

There is considerable linguistic overlap between the idioms. For example, though there are 3537 unique idioms, there are only 796 unique starter words.

We have arranged the idioms as trees, where the roots are the start words, and the idioms are found by following a path from a root node to a leaf node.
Example:

```
"alt": {
    "er": {
        "bare": {
            "fryd": {
                "og": {
                    "gammen": {}
                }
            }
        },
        "klappet": {
            "og": {
                "klart": {}
            }
        },
        "såre": {
            "vel": {}
        }
    },
    "går": {
        "sin": {
            "vante": {
                "gang": {}
            }
        }
    },
}
```

Here is a tree where "alt" is the root, the idioms "alt er bare fryd og gammen", "alt er klappet og klart", "alt er såre vel" and "alt går sin vante gang" can be found by traversing the tree.
These trees can be found in idiom_graphs/.

There are also flat versions of the trees (where consecutive nodes/words with only one child are merged into word sequences)

Example:

```
"alt": {
    "er": {
        "bare fryd og gammen": "",
        "klappet og klart": "",
        "såre vel": ""
    },
    "går sin vante gang": "",
}
```

Idiom completion as an NLP task

The idioms are split into idiom starts (the first N-1 words) and accepted completions (a list of possible last words to complete the idiom). Of the 3245 rows, there are 156 where there are more than one accepted completion.

This dataset can be used to measure a generative language models' ability to complete well known idioms, or as a masked language modelling task.

The dataset can be loaded with the datasets library and is also on huggingface.