KWJP$\frac{1}{2}$M

Paczka `samples.tar.gz` zawiera podkorpus KWJP$\frac{1}{2}$M, czyli losowe próbki tekstów z Korpusu Współczesnego Języka Polskiego. KWJP$\frac{1}{2}$M składa się ze 11 108 próbek obejmujących łącznie 513 009 słów tekstowych „od spacji do spacji". Pojedyncza próbka to jedno lub więcej zdań występujących kolejno w tekście korpusu.

Liczba próbek pochodzących z każdego tytułu (książki, dziennika, tygodnika etc.) wchodzącego w skład KWJP jest proporcjonalna do udziału tego tytułu w stumilionowym korpusie zrównoważonym.

Próbki zapisane są w formacie `.json`. W pojedynczym pliku `xyz_sample.json` zebrane są wszystkie próbki wylosowane z dokumentu korpusu o identyfikatorze `xyz`. Przykładowa zawartość takiego pliku wygląda następująco:


```
{
    "meta": {
        "id": "146648",
        "meta_author": "",
        "meta_title": "Kącik Gier Historycznych: Ostrołęka 26 maja 1831",
        "meta_editor": "",
        "meta_main_title": "Rebel Times",
        "meta_published": "2015-11-24",
        "meta_first_published": "",
        "meta_publisher": "",
        "meta_pubplace": "Gdańsk",
        "meta_isbn": "",
        "meta_issn": "",
        "meta_pubnumber": "98",
        "meta_topic": "",
        "meta_kwjp_type": "typ_fakt",
        "meta_kwjp_channel": "kanal_prasa_miesiecznik",
        "meta_label": "rebel-times_98"
    },
    "samples": [
        {
            "sentence_ids": [
                "text_structure.xml/para_5/sent_4",
                "text_structure.xml/para_5/sent_5"
            ],
            "text": "Oddziały kawalerii są w skali dywizjonu (jeden dywizjon to
dwa szwadrony, dwa dywizjony składają się zazwyczaj na pułk kawalerii), czyli
jeden oddział to zwykle 200-400 ludzi. W przypadku artylerii oddziały są w skali
baterii - jedna bateria to na ogół od 6 do 12 dział i 150-300 ludzi obsługi."
        },
        {
            "sentence_ids": [
                "text_structure.xml/para_18/sent_3",
                "text_structure.xml/para_18/sent_4",
```

```
                "text_structure.xml/para_18/sent_5"
            ],
            "text": "Trudno zrobić szybko dobrą, trochę bardziej zaawansowaną grę
historyczną. Trzeba do tego i wiedzy, i umiejętności, i praktyki (trzeba przez
dłuższy czas grać w takie gry). Trzeba mieć jakąś własną wizję, pomysł, jak taka
gra ma wyglądać i czym będzie się różnić od innych jej podobnych."
        }
    ]
}
```


Słownik reprezentujący próbki z pojedynczego dokumentu zawiera dwa klucze: `meta` (metadane próbki) oraz `samples` (właściwe próbki).

Elementy (być może puste, jeśli dana informacja była nieznana lub niedostępna) metadanych to:

— `id` – identyfikator dokumentu,
— `meta_author` – autor,
— `meta_title` – tytuł tekstu,
— `meta_editor` - redaktor,
— `meta_main_title` – tytuł główny (np. tytuł czasopisma),
— `meta_published` – data wydania,
— `meta_first_published` – data pierwszego wydania,
— `meta_publisher` – wydawca,
— `meta_pubplace` – miejsce wydania,
— `meta_isbn` – ISBN,
— `meta_issn` – ISSN,
— `meta_pubnumber` – numer periodyku (dotyczy tylko prasy),
— `meta_topic` – tematyka (wartość opisowa; dotyczy tylko książek),
— `meta_kwjp_type` – typ; możliwe wartości:
  — `typ_fikcja`,
  — `typ_fakt`,
  — `typ_publicystyka`,
— `meta_kwjp_channel` – kanał; możliwe wartości:
  — `kanal_ksiazka`,
  — `kanal_prasa_dziennik`,
  — `kanal_prasa_tygodnik`,
  — `kanal_prasa_miesiecznik`,
  — `kanal_prasa_inne`,
  — `kanal_internet`,
— `meta_label` – etykieta.

W przypadku książek pole `meta_title` zawiera tytuł książki, zaś pole `meta_main_title` pozostaje puste. W przypadku wydawnictw cyklicznych pole `meta_title` zawiera tytuł dokumentu (np. artykułu), zaś pole `meta_main_title` – tytuł dziennika, tygodnika etc. W szczególności może być kilka plików o tym samym tytule głównym, zawierających próbki pochodzące z różnych dokumentów w obrębie jednego tytułu np. prasowego.

Próbki tekstów zapisane są w postaci listy słowników, które z kolei zawierają dwa klucze:
— `sentence_ids` – lista identyfikatorów zdań składających się na próbkę w pełnym korpusie,
— `text` – tekst próbki.