Czech HS Contracts Dataset (CHSC) 1.0

Czech Contracts dataset was created as a part of the thesis Low-resource Text Classification (2021), A. Szabó, MFF UK.

Contracts are obtained from the Hlídač Státu web portal. Labels in the development and training set are automatically classified on the basis of the keyword method according to the thesis Automatická klasifikace smluv pro portál HlidacSmluv.cz, J. Maroušek (2020), MFF UK. For this reason, the goal in the classification is not to achieve 100% on the development set, as the classification contains a certain amount of noise. The test set is manually annotated. The dataset contains a total of 97 493 contracts.

- Train: 86823
- Development: 9646
- Test: 1024

Note that we also add the development set (`dev10.jsonl`) that we use, which is 10% of the original develepment set.

Categories

The categories are two-level, which allows the classification into 22 main categories or into all 105 categories. The list of individual categories is in a separate `.json` file `categories.json`. The main categories are recognized by the fact that their labels are divisible by one hundred.

Format

Train, Development and Test sets are saved in `.jsonl` format.

An example of a record:


```
{
"identifikator": {"idSmlouvy": "8016459", "idVerze": "8655331"},
"CalculatedPriceWithVATinCZK": 157300.0,
"CalcutatedPriceQuality": 2,
"casZverejneni": "2019-04-02T13:34:10+02:00",
"Classification": {"LastUpdate": "2020-10-04T05:38:10.3604997+02:00", "Version": 1, "Types":
  [{"TypeValue": 10901, "ClassifProbability": 5.0, "RootClassification": false,
    "TypeName": "Paliva a oleje"}]},
"ConfidenceValue": 0.0,
"datumUzavreni": "2019-03-28T00:00:00",
"Enhancements": [{"Created": "2019-04-02T13:37:26.1756949+02:00",
  "Title": "Text přílohy extrahován z obsahu dokumentu ", "Description": "",
  "Changed": {"ParameterName": "item.Prilohy[0].PlainTextContent", "PreviousValue": "", "NewValue": ""},
  "Public": true, "EnhancerType": "HlidacStatu.Plugin.Enhancers.TextMiner"}],
"hodnotaBezDph": 130000.0,
"Id": "8655331",
"Issues": [],
"LastUpdate": "2020-07-28T05:02:55.8289184+02:00",
"odkaz": "https://smlouvy.gov.cz/smlouva/8655331",
"Platce": {"adresa": "Souběžná 2349/7, 46601 Jablonec nad Nisou, CZ", "datovaSchranka": "r986i9w",
  "ico": "25475509", "nazev": "Technické služby Jablonec nad Nisou, s.r.o."},
"platnyZaznam": true,
"PravniRamec": 2,
"predmet": "Kupní smlouva - malotraktor John Deere 2520",
"Prijemce": [{"adresa": "Karlovice - Svatoňovice 24, 51101, Turnov", "ico": "61199567",
  "nazev": "Lubomír Janků"}],
"Prilohy": [{"FileMetadata": [], "nazevSouboru": "Kupní smlouva John Deere 2520.pdf",
  "hash": {"algoritmus": "sha256",
    "Value": "e16a9896a96026058913af5aa2c819ee745e574f0d15c8a6a5afffd8d1f36267"},
  "odkaz": "https://smlouvy.gov.cz/smlouva/soubor/11444911/Kupn%C3%AD%20smlouva%20John%20Deere%202520.pdf",
  "PlainTextContentQuality": 3, "LastUpdate": "2019-04-02T13:37:26.1756949+02:00", "Lenght": 2543,
  "WordCount": 310, "UniqueWordsCount": 247, "WordsVariance": 0.005515470108369287, "Pages": 0,
  "EnoughExtractedText": true}],
"spadaPodRS": true,
```

```
"SVazbouNaPolitiky": false,
"SVazbouNaPolitikyAktualni": false,
"SVazbouNaPolitikyNedavne": false,
"Hint": {"SmlouvaULimitu": 0, "DenUzavreni": 0, "SmlouvaSPolitickyAngazovanymSubjektem": 0,
  "PocetDniOdZalozeniFirmy": 5663, "VztahSeSoukromymSubjektem": 1},
"VkladatelDoRejstriku": {"adresa": "Souběžná 2349/7, 46601 Jablonec nad Nisou, CZ",
  "datovaSchranka": "r986i9w", "ico": "25475509", "nazev": "Technické služby Jablonec nad Nisou, s.r.o."},
"Label": 10901,
"PlainTextContent": "Kupní smlouva Lubomír Janků Svatoňovice 24 Karlovice 511 01 Turnov Zastoupený
  panem Lubomírem Janků IČ: 61199567, DIČ: CZ7605263446 Tel.: 736415500 dále jen kupující a Technické
  služby Jablonec nad Nisou, s.r.o. IČ: 254 75 509, DIČ: CZ25475509, se sídlem Souběžná 7, PSČ 466 01,
  Jablonec nad Nisou č. účtu: 27-633560227/0100 zastoupená Mgr. Milanem Nožičkou, ředitelem
  a jednatelem společnosti Kontaktní osoba: Miloš Šikola, vedoucí střediska autodopravy dále jen
  prodávající I. Předmětem koupě je malotraktor John Deere 2520, RZ L00 0335, typ 268, varianta FD22
  Tovární značka: John Deere Výrobce: Deere Company, Moline, USA Rok výroby: 2009, stav 3208 Mth
  Motor: Yanmar Diesel, Engine CO, LTD Japonsko Max. výkon: (kW/ot.min.): 19.1/2600 Palivo: motorová
  nafta Barva: zelená, kombinovaná Identifikační číslo (VIN): LV2520E486086 Technická způsobilost
  platná do 10.3.2021, technické osvědčení, ev. č.: ZA 156791, malé OTP č.: ZAA 011231 Kupující byl
  seznámen se skutečností, že je zadní vývodový hřídel nefunkční. II. Prodávající prohlašuje, že
  předmět koupě je jeho vlastnictvím a není zatížen zástavním právem ani jinými závazky. I. Sjednaná
  cena činí 130 000,- Kč bez DPH. Celková cena činí 157 300,- Kč vč. DPH. Sjednaná cena bude zaplacena
  v hotovosti dle převzaté faktury dne 28.3.2019. Kupující byl seznámen se stavem předmětu koupě
  a souhlasí s tím, že na předmět prodeje není poskytnuta záruka. IV. Obě strany prohlašují, že tato
  smlouva nebyla uzavřena za nápadně nevýhodných podmínek a jejímu obsahu rozumějí. Smlouva byla
  vyhotovena ve dvou výtiscích, z nichž každá strana obdrží po jednom. V. „Smluvní strany jsou
  srozuměny s tím, že tato smlouva bude bez jakéhokoli omezení, včetně identifikace smluvních stran,
  zveřejněna v souladu se zákonem č. 340/2015 Sb. zákon o registru smluv, na Portálu veřejné správy
  (http://portal.gov.cz/), včetně případných dodatků a změn. Smluvní strany nepovažují žádnou ze
  skutečností ve smlouvě uvedených za obchodní tajemství ve smyslu § 504 zák. č. 89/2012 Sb., občanský
  zákoník." V Jablonci nad Nisou, dne 28.3.2019 .....
  ..... Kupující Prodávající",
"WindowsRange": [(118, 1018), (756, 1916)]
}
```


Classification

Important objects for classification are the last three ones:

- **Label** – category of the contract
- **PlainTextContent** – text of the contract
- **WindowsRange** – ranges of the text sorted in descending order according to the number of obtained keywords. The original method, with which the training and development set was obtained, used keywords for classification, and therefore we found windows where are most of those keywords in the text. The size of each window is limited to 300 tokens, and keywords are centered in the center of the window. They are specified as character indexes in the PlainTextContent contract. You can find more details in section 2.3.4 of the thesis Low-resource Text Classification (2021), A. Szabó, MFF UK.

License – CC BY-NC-SA 4.0

Contact

- Milan Straka: straka@ufal.mff.cuni.cz
- Adam Szabó: adam.szabo707@gmail.com