Specifikace

Popis a zaměření

DeriNet je lexikální síť modelující slovotvorné vztahy mezi slovy z českého jazyka. Základní jednotkou této sítě je lexém, jenž obsahuje základní informace o jednotlivých slovech. V tuto chvíli se aktuální verzi databáze *DeriNet 2.3*, již jsem sama používala, nachází přes 1 milion lexémů.

Kvůli takové rozsáhlosti sítě není možné vztahy mezi jednotlivými slovy zadávat a kontrolovat ručně, a tak některé vztahy chybí, nebo naopak přebývají. Mým úkolem bylo sepsat sadu testů, které tyto chyby odhalí a, pokud to bude možné, najdou i možná řešení.

Použité technologie

- Samotná databáze DeriNet 2.3
    - Získána z této webové stránky: https://lindat.mff.cuni.cz/repository/items/62540779-b206-4cf7-ac33-399ce68e35e6
- API pro přehlednější komunikaci a přístup s databází
    - Získáno z této webové adresy: https://github.com/vidraj/derinet/tree/master/tools/data-api/derinet2#usage
    - Před používáním doporučeno se seznámit s popisem na uvedené webové adrese
- Python (jazyk využitý při psaní a vykonávaní testů)

Databáze DeriNet 2.3

Jak jsem již zmínila, základní jednotkou v databázi je lexém, kde se schraňují základní informace o jednotlivých slovech, které jsou pro mé testy signifikantní.

Tyto informace můžeme rozdělit na ty týkající se pouze daného lexému (které se dále dělí na editovatelné a needitovatelné) a na ty zabývající se vztahy mezi lexémy.

Do první kategorie patří tyto informace

- informace již neměnitelné
    - *lemma* – základní podoba slova
    - lemid – kombinbace lemma + morfologický tag podle Prague Dependency Treebank (PDT)
    - *pos* – slovní druh
- editovatelná data
    - *feats* – základní informace (mluvnické kategorie anebo zda je slovo cizího původu)
    - *segmentation* – segmentace slova na morfémy (nejmenší vydělitelná část slova, která je nositelem věcného nebo gramatického významu)
    - *misc* – statistika o slově v korpusu (absolutní počet výskytů, jak časté dané slovo je oproti ostatním anebo relativní frekvence)

Do druhé kategorie patří tyto data

- *all_parents* – seznam obsahující všechny lexémy, z nichž vzniklo zkoumané slovo
- *parent* – hlavní předek daného slova (nejvýznamnější, pokud jde o vztah skládání, a jediné, pokud jde o vztah odvozovací)
- *all_children* – seznam všech lexémů, jimž je dané slovo předkem

Ve svých testech využívám hlavně *parent, feats, misc, pos* a *lemma*.

Testy

Název testů

Názvy jsou vymyšleny tak, aby bylo na první pohled poznat, co se v daném testu kontroluje.

Struktura samotných testů

Načtení databáze

Jelikož jsem celou dobu pracovala s databází přes API, bylo potřeba pro každý test nejdříve danou databázi načíst.

Pro ukázku přikládám, jak vypadalo načítání databáze na mém zařízení, kde se soubor s databází nacházel ve stejném adresáři jako všechny testy.

```
import derinet.lexicon as dlex
import os
lexicon = dlex.Lexicon()
adresar = os.getcwd()  # aktuální adresář
cesta_k_souboru = os.path.join(adresar, "./derinet-2-3.tsv")  #sestavení cesty
lexicon.load(cesta_k_souboru)
```

Hlavní část – logika testu

Každý program se v této části liší. Někdy se zde kontroluje, zda lexém, který by měl mít podle mé teorie předka, ho opravdu má, a pokud ne, tak se zkusí najít možný předek. Nebo naopak, zda daný lexém na sebe nemá špatně napojeného potomka, a pokud ano, tak co by měl být pro potomka správný předek.

Výstupem této části je jeden až několik seznamů, které obsahují slova nalezena daným testem.

Výpis

Pro přehlednost každý test vypisuje svůj výstup do souboru se stejným názvem. Nejprve se do souboru napíše, co bylo cílem daného testu a jak bude výstup strukturován. Poté se přejde k vypsání nalezených slov.

Výpis se nejčastěji dělí na slova, která skoro jistě spadají do této chyby a/nebo k nim bylo nalezeno řešení, a poté na slova, jež je potřeba více zkontrolovat a/nebo k nim nebylo nalezeno řešení. Obě kategorie jsou od sebe odděleny.

Pro ukázku připadám výpis jednoho z testů.

```
GEOGRAFICKÉ POJMY, KTERÉ MAJÍ IDENTICKÉ SLOVO ZAČÍNAJÍCÍ MALÝM PÍSMENEM A NEJSOU K SOBĚ PŘIPOJENI
GEOGRAFICKÝ POJEM   ODVOZENÉ SLOVO

Aalen               aalen
Aberdeen            aberdeen
Abies               abies
Adamantina          adamantina
```

Kategorie testů

Rozdělení podle vážnosti chyby

Ne všechny testy odhalují chyby na stejné úrovni vážnosti, a proto jsem rozdělila testy do tří skupin

- *Info* – testy obsahující malé množství chyb, anebo pozorování o zvláštních jevech, jichž jsem si za tu dobu všimla (nemusí to být vyloženě chyba, ale stojí to za zaznamenání)
- *Warning* – testy naznačující chybějící vztahy (něco pravděpodobně chybí)
- *Error* – testy ukazující na vztahy, které by se v databázi neměly nacházet (něco pravděpodobně přebývá)

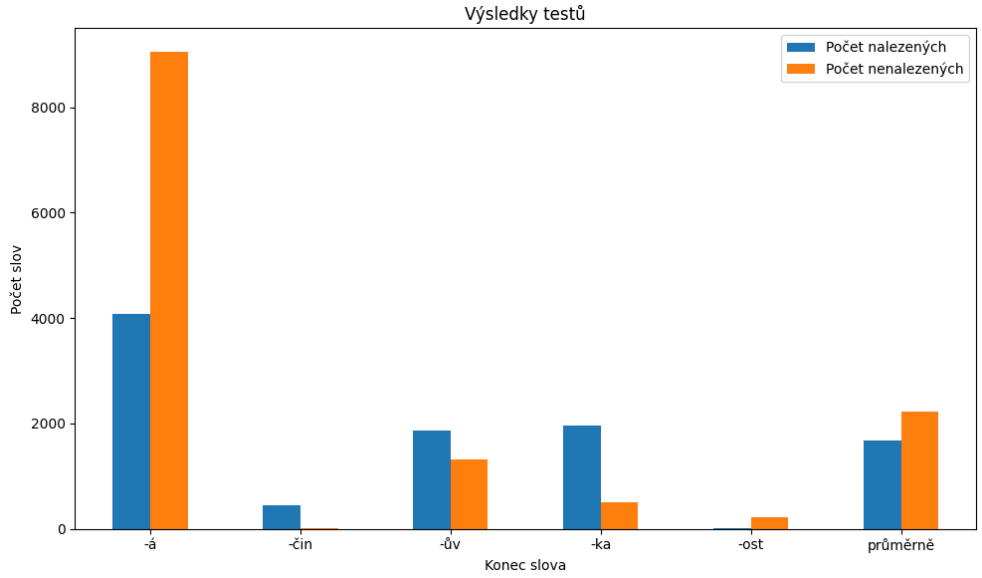
Určená skupina pro daný test se nachází na prvním místě názvu daného testu.

Statistiky

1. *Úspěšnost nalezení možného řešení pro jeden typ testů*

   Nejčastěji jsem testovala slova podle koncovek. Například přídavná jména končící na ‚čin' by měla být připojena k podstatnému jménu. Ne pro všechna slova, která tyto testy našly, se mi podařilo najít možného předka.

   Přikládám tedy graf podle (ne)nalezení možného předka u pěti testů zabývající se podobnou myšlenkou, a k tomu průměrné hodnoty.



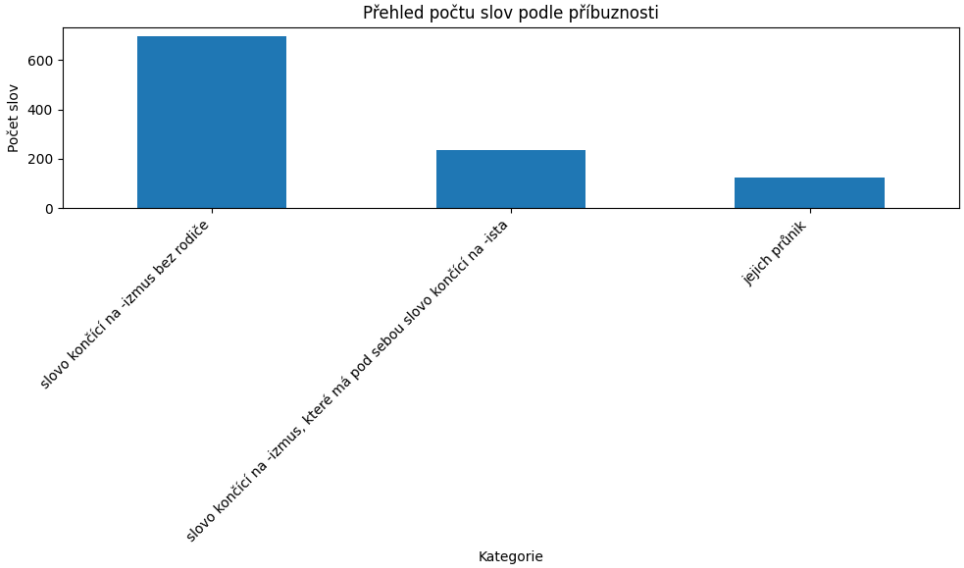
   Jde vidět, že ačkoliv všechny testy jsou na stavěné na podobný typ chyby, neznamená to, že najdou podobný počet hodnot.

2. *Geografické názvy*

Některé geografické názvy, které začínají velkým písmenem, mají v databázi lexémy lišící se pouze tím, že začínají malým písmenem (např. *Absurdistán* a *absurdistán*). Zajímalo ně, jak častý je to fenomén obecně mezi geografickými pojmy začínající velkým písmenem a zda jsou k sobě navzájem připojena anebo ne.

3. *Nejasný typ vztahu*

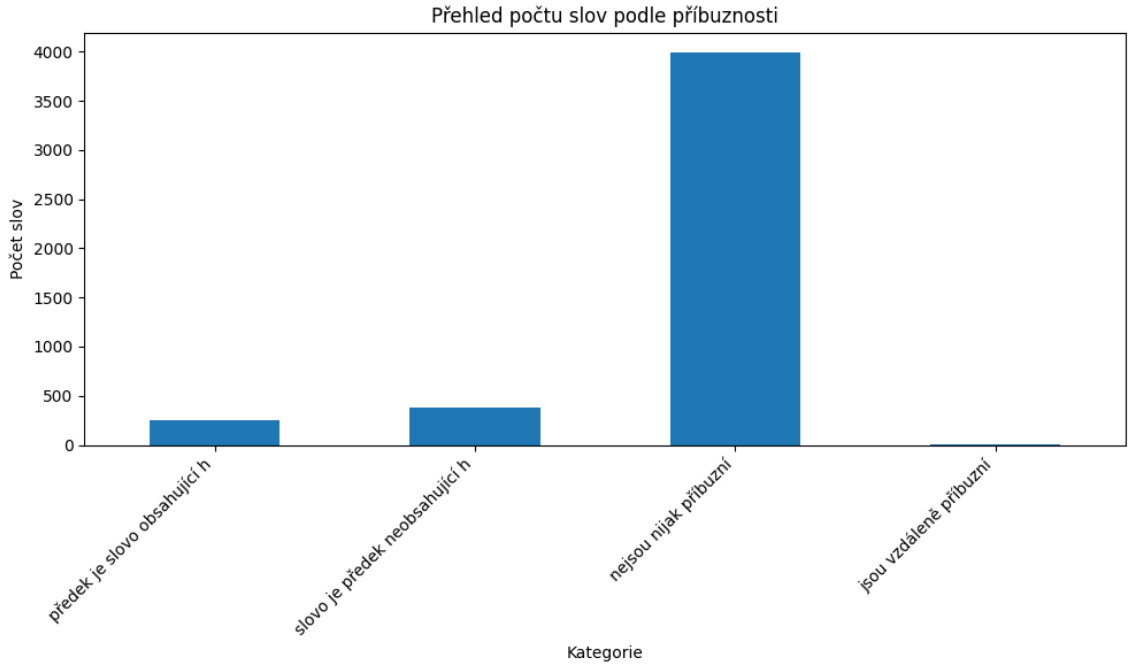
V databázi se nachází spoustu slov, které se liší pouze vynechaným h, tedy například *gothaj* a *gotaj*. Jde ale vidět, že neexistuje jednotné pravidlo, jak a zda vůbec by tato slova měla být spojena. Napsala jsem tedy test, který tento úkaz v databázi projde a poté seřadí jednotlivé kategorie podle toho počtu slov.



Na graficky znázorněném testu jde vidět, že všechny kategorie jsou zahrnuty, ačkoliv ta poslední obsahuje pouze 2 případy.

4. *Izmus a ismus*

Velkou kapitolou samy o sobě jsou slova, končící koncovkou ‚-ismus' a jejich „dětmi". Ani zde neexistuje jednotné pravidlo, a tak mě zajímalo, jaký je vztah mezi tím, že slovo končící na ‚-izmus' je bez rodiče (sloupec 1), a zároveň k sobě váže slovo končící na ‚-ista' (sloupec 2). Jde vidět, že ten první úkaz je o dost častější než ten druhý, ale pokud se dané slovo nachází v druhé kategorii, tak má vysokou pravděpodobnost, že se bude nacházet i v té první.



Čeho jsem si ale všimla, neexistuje případ, kde by slovo ‚-ista' mělo za rodiče jak ‚-izmus', tak ‚-ismus', vždy je to maximálně jeden z nich.

5. *Jmenný rod podstatného jména*

Často jsem testovala podstatná jména a několikrát se mi stalo, že lexému daného slova neobsahoval informaci, kterého rodu je. Zajímalo mě tedy, jak častý je to úkaz.