Určování obtížnosti textu v českém jazyce

Dokumentace

Michal Vlasák - FI MUNI

23. května 2014

Obsah

1 Úvod

Program slouží pro analýzu textu v českém jazyce a určení jeho objektivní obtížnosti na základě několika různých indexů. Určení obtížnosti textu je důležitý prvek pro zkvalitnění výuky a zatímco ve světě již pro angličtinu existují nástroje volné (zde a zde) i komerční, pro český jazyk zatím žádný takový automatizovaný systém není. Projekt je dostupný na adrese `http://nlp.fi.muni.cz/projekty/cstexteval/py.cgi`.

1.1 Použité indexy

V práci je použito několik indexů pro určení obtížnosti textu. Zde následuje jejich popis. Jelikož některé vzorce mívají společné prvky výpočtu, zde je zavedeno jejich společné značení:

T Obtížnost textu

R Doporučený ročník pro daný text

$\overline{U}$ Průměrná délka věty v poštu slov

H Procento obtížných slov (vysvětleno dále)

N Počet slov v textu

1.1.1 Gunning-Fog index

Jeden z prvních indexů obtížnosti textu z roku 1944. Do dneška jeden z nejjednodušších a nejspolehlivějších indexů pro angloamerické texty. Dosáhnul korelace r=.91 s naměřenou dovedností čtení s porozuměním [2]. Obtížnost se dle tohoto indexu spočítá jako

$$T = 0.4 \cdot \left(\overline{U} + 100 \cdot H\right)$$

Obtížná jsou v původním vzorci definována jako slova skládající se z více než dvou slabik, ale vzhledem k tomu, že původně byl index zamýšlen pro texty v angličtině, zde byl jako hraniční limit zvoleno slabik 5.

1.1.2 Flesch-Kincaid index

Původně určen pro americké námořnictvo, nyní využíván v americkém školském systému a v Microsoft Word. Dosáhnul korelace r=.91 s naměřenou dovedností čtení s porozuměním [2]. Obtížnost dle tohoto indexu je

$$T = 0.39 \cdot \overline{U} + 11.8 \cdot \left(\frac{S}{N}\right) - 15.59$$

kde S je počet slabik. Výsledkem je číslo od 0 do 100, kde vyšší číslo reprezentuje jednodušší text. Převodní tabulka na věk žáků je pak zhruba následující

Skóre	Interpretace
90 – 100	pro žáky do 11 let
60 – 70	pro žáky mezi 13 a 15 lety
0 – 30	pro vysokoškolské studenty

1.1.3 Dale-Chall index

Využívá seznamu „lehkých“ slov. Dosáhl korelace r=.93 s porozuměním textu. Jedná se o jeden z nejspolehlivějších vzorců a proto se využívá i ve výzkumné oblasti [3]. Obtížnost textu je

$$T = 0.1579 \cdot H + 0.0496 \cdot \overline{U}$$

Pokud navíc procento obtížných slov překročí hranici 5%, pak se k obtížnosti přičte 3.6365. V původní práci jsou obtížná slova ta, která nejsou na seznamu lehkých slov. V mém řešení jsou to ta, sestávající z více než 5 slabik. Převodní tabulka je pak následující

Skóre	Interpretace
4.9 nebo nižší	pro žáky 4. či nižšího ročníku
5.0 – 5.9	pro žáky 5. či 6. ročníku
6.0 – 6.9	pro žáky 7. či 8. ročníku
7.0 – 7.9	pro žáky 9. či 10. ročníku
8.0 – 8.9	pro žáky 11. či 12. ročníku
9.0 – 9.9	pro studenty bakalářského studia
10.0 nebo vyšší	pro studenty s bakalářským nebo vyšším vzděláním

1.1.4 Coleman-Liau index

Založen čistě na kvantitativních vlastnostech textu, nevyžaduje tedy hlubší analýzu textu. Obtížnost je určena jako

$$T = 0.0588 \cdot L - 0.296 \cdot S \ - 15.8$$

kde L je průměrný počet písmen na 100 slov a S je průměrný počet čísel na 100 slov.

1.1.5 SMOG

Dosahuje korelace r=.88 s porozuměním textu. Doporučován pro použití v lékařství [2]. Ročník, pro který by měl být text vyhovující se pak přímo určí jako

$$R = 3 + \sqrt{F}$$

kde F je počet slov, skládajících se ze dvou a více slabik (v mé implementaci z více než pěti) v přepočtu na 30 vět.

1.1.6 FORCAST

Původně určeno pro americkou armádu. Dosahuje korelace r=.66 s porozuměním textu [2]. Ročník je přímo určen jako

$$R = 20 - \left(\frac{E}{10}\right)$$

kde E je počet jednoslabičných slov na každých 150 slov.

1.1.7 Průcha-Pluskal-Nestlerová

Zabývá se kvantitativní i kvalitativní stránkou textu. Jediný rozpracovaný index pro texty v českém jazyce. Již mnohokrát použit zejména pro určování obtížnosti učebnic.

Z výše zmíněných indexů je pro český jazyk určující zejména poslední jmenovaný. Ačkoliv jsou v programu implementovány i ostatní, jejich výsledky neodpovídají realitě - a to zejména proto, že jsou původně určeny pro anglický text.

Index obtížnosti dle Průchy (dále upraven Pluskalem a Nestlerovou) uvažuje „syntaktickou“ i „sémantickou“ obtížnost textu, což je další věc, která ho odlišuje od předchozích zmíněných. Ty to totiž onu zmiňovanou „sémantickou“ obtížnost uvažují maximálně na úrovni, jestli slovo je, nebo není „složité“. Průcha rozlišuje hned několik kategorií pojmů, dle kterých obtížnost určuje. Syntaktickou obtížnost určuje jako:

$$T_S = 0.1 \cdot \overline{U} \cdot \overline{V}$$

kde $\overline{U}$ je průměrná délka věty (v počtu slov) a $\overline{V}$ je průměrný počet slov na jedno sloveso. Sémantickou obtížnost pak definuje jako:

$$T_P = 100 \cdot \frac{P}{N} \cdot \frac{P_1 + 3P_2 + 2P_3 + 2P_4 + P_5}{N}$$

kde P je počet pojmů (pojmem Průcha rozumí podstatná jména včetně přídavných jmen zpodstatnělých a číslovky), N celkový počet slov a P_i pak jednotlivé pojmové kategorie

P_1 Běžné pojmy

P_2 Odborné pojmy

P_3 Faktografické pojmy

P_4 Číselné údaje, číslovky

P_5 Opakované pojmy

Celková obtížnost je pak

$$T = T_S + T_P$$

2 Popis implementace

2.1 Požadované moduly

Zejména kvůli práci s textem jsou vyžadovány některé další moduly, jejichž popis a možnost instalace následuje.

2.1.1 Enum

Slouží k internímu ukládání kategorií jednotlivých slov

Instalace

```
wget https://pypi.python.org/packages/source/e/enum/enum-0.4.4.tar.gz
tar -zxvf enum-0.4.4.tar.gz
cd enum-0.4.4
python setup.py install
```

2.1.2 Manatee

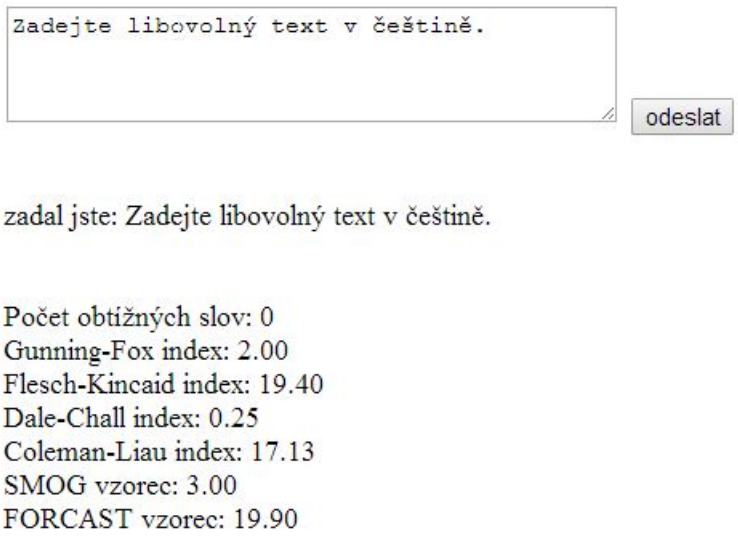
Soubor využívá nástrojů Centra zpracování přirozeného jazyka na Fakultě informatiky Masarykovy univerzity. Zejména potřebuje mít pro práci s korpusem přístupny modul `manatee`.

2.2 Přístup z internetu

Soubor je CGI skript, tudíž musí být umístěn ve složce přístupné z webu a být spustitelný

```
chmod +x cstext.cgi
```

Poté můžete přistoupit na danou webovou adresu a měl byste vidět oblast pro zadání textu a tlačítko pro odeslání textu.



2.3 Popis běhu programu

Po odeslání textu se spočítají některé základní údaje. Nejprve je to počet slabik textu. Text je dle mezer rozdělen na slova a v každém slově jsou samohlásky s diakritikou převedny na verzi bez diakritiky. V rámci každého slova jsou pak slabiky určeny pomocí regulárního výrazu jako libovolně dlouhá skupina souhlásek následována libovolně dlouhou skupinou samohlásek. V jedné z verzí byl počet slabik určován pomocí modulu `PyHyphen`, který pro určení slabik využívá slovníku pro LibreOffice, ale výsledky byly horší, než s výše popsaným postupem.

Dle počtu slabik je pak určen počet „obtížných“ (viz.kapitolu 1.1) a jednoslabičných slov. K určení počtu vět je použit modul `TextBlob`. Dále je za pomoci regulárních výrazů určen počet znaků a počet čísel v celém textu. Na základě takto získaných údajů je již možné spočítat všechny indexy, kromě posledního popsaného - Průcha-Pluskal-Nestlerová - který je pro nás však nejdůležitější.

Pro určení indexu Průcha-Pluskal-Nestlerová musí být text zpracován důkladněji. Nejdříve je nutné určit počet podstatných jmen, číslovek a sloves. První dvě skupiny slouží jako základ pro práci s pojmy, počet sloves pak spoluurčuje syntaktickou část obtížnosti, která už v tuto chvíli může být spočítána.

V následujícím kroku je nutné rozdělit jednotlivá slova do daných kategorií, což je nezbytné pro správné určení sémantické obtížnosti. Před samotným určením kategorií je ovšem nutné o slovech získat dodatečné informace. To je učiněno pomocí nástrojů Centra zpracování přirozeného jazyka, především pomocí morfologického analyzátoru `majka`. Pomocí nich jsou ke slovům určena jejich lemmata, a morfologické značky. Dále je za pomocí modulu manatee a korpusu cztenten12 určena frekvence každého z lemmat.

Jednotlivé kategorie pojmů mají danou prioritu (tzn. pokud slovo patří do více kategorií, je mu přiřazena kategorie s vyšší prioritou), které jsou v kódu reprezentovány pořadím určování jednotlivých kategorií. Priority jsou následující (od nejvyšší po nejnižší)

1. Opakované pojmy
2. Číselné pojmy
3. Odborné pojmy
4. Faktografické pojmy
5. Běžné pojmy

U každého slova je určeno, zda je podstatné jméno, nebo číslovka a následně jsou u něj od nejvyšší priority určovány kategorie pojmů. Jakmile do nějaké kategorie slovo spadá, je tato jeho výslednou kategorií, se kterou se pak pracuje při výpočtu. V paměti je udržován seznam lemmat již zpracovaných slov, jenž slouží pro určení opakovaných pojmů. Výjimku tvoří číslovky a číselné pojmy, které mají ve výstupu z `majky` stejné lemma, proto je číslovka vždy určena jako číslovka a nikdy jako opakovaný pojem. Vzhledem k povaze číselných pojmů to nepovažuji za škodlivé. Odborné pojmy jsou definovány jako ty, které mají frekvenci lemmatu nižší než 0.2. Tato hodnota se v průběhu vývoje ukázala jako rozumná volba. Faktografické pojmy jsou určena na základě velkého prvního písmena ve slově. `majka` totiž všechna velká písmena slov, jenž nejsou vlastními jmény příp. zkratkami převede na písmena malá. Tím odpadá možnost určit jako faktografický pojem slovo čistě na základě toho, že je na začátku věty. V průběhu se ukázalo, že určování faktografických pojmů tímto způsobem je poměrně spolehlivé - jediný problém nastává s víceslovnými pojmy, které jsou zpracovávány jako dva oddělené pojmy. Jako běžný pojem jsou pak určeny ty pojmy, které se nedostanou do žádné z předchozích kategorií.

Nyní už jsou k dispozici všechny údaje, jenž jsou pro spočtení indexu Průcha-Pluskal-Nestlerová potřeba, tudíž program tak učiní a výsledek vypíše uživateli, včetně počtů slov v jednotlivých pojmových kategorií a jejich poměrového zastoupení v textu.

3 Použití

Do textové oblasti zadejte požadovaný text a klikněte na "Odeslat". Po chvíli by se Vám měl zobrazit původní text a informace o jeho obtížnosti dle několika indexů.

Syntaktická: 2.50
Sémantická: 16.00
Celkem: 18.50

Pojmů: 2
Běžných pojmů: 2
Odborných pojmů: 0
Faktografických pojmů: 0
Číselných pojmů: 0
Opakovaných pojmů: 0

4 Vyhodnocení

Vyhodnocení probíhalo zejména porovnáním naměřených hodnot s hodnotami, naměřenými Petrou Tannenbergovou v její disertační práci, zabývající se obtížností učebnic [1]. Ze 4 učebnic z původní práce byly použité části textu zadány do programu a hodnoty jím naměřené byly porovnány s původní práci (celkem se jednalo o cca 30 stran textu).

Dílčí ukazatele obtížnosti se od původní práce značně lišily. Odborných pojmů byla datekována pouze čtvrtina, faktorgrafických pak zhruba polovina. Naopak bylo detekováno více číselných pojmů. Ačkoliv naměřená sémantická i syntaktická obtížnost byly menší, než v původní práci, korelace s původním hodnocením byla velmi vysoká ($r = .992$). Tento výsledek považuji za velice uspokojivý a program byl následně otestován na dalších sadách textů a to jmenovitě na pohádkách([a] [b] [c]) a na vysokoškolských textech ([a] [b] [c]) . Základní deskriptivní hodnoty naměřené na těchto textech jsou v tabulce 1

	Průměr	Rozptyl	Odchylka
Pohádky	13.12	0.18	0.42
Učebnice pro 6. a 7. třídu	24.20	5.68	2.38
VŠ skripta	36.88	4.62	2.15

Tabulka 1: Výsledky vyhodnocení

Ze zajímavých zjištění pak stojí za povšimnutí několik věcí:

- U pohádek platilo, že čím jsou věty složitější, tím jsou kratší
- To samé platilo u vysokoškolských materiálů
- Učebnice pro středoškoláky i vysokoškolská skripta měla (narozdíl do pohádek) vyvážený poměr mezi syntaktickou a sémantickou obtížností

Reference

[1] TANNENBERGOVÁ, Petra. *Analýza didaktické vybavenosti učebnic dějepisu pro 6. a 7. ročník základní školy.* [online]. 2012 [cit. 2014-05-22]. Disertační práce. Masarykova univerzita, Pedagogická fakulta. Vedoucí práce Jaroslav Vaculík. Dostupné z: <http://is.muni.cz/th/15216/pedf_d/>

[2] DUBAY, W. H. 2006. *Smart language: Readers, Readability, and the Grading of Text.* Costa Mesa:Impact Information.

[3] CHALL, J. S. and E. Dale. 1995. *Readability revisited: The new Dale–Chall readability formula.* Cambridge, MA: Brookline Books.