JAPE - týmový úkol

Anastasiya Kiptsevich, Sára Juranková

4. dubna 2019

Obsah

1 Cíl projektu

Cílem našeho projektu bylo vytvořit gramatiku extrahující z textů název modelu notebooku, procesoru a grafické karty.

2 Trénovací dokumenty

Do trénovací skupiny dokumentů byly zařazeny tři dokumenty:

- Recenze notebooku MSI GE75 Raider na `www.pcmag.com`
- Stránka notebooku Acer Aspire E 15 na `www.amazon.com`
- Stránka notebooku HP Stream na `www.ebay.com`

Tyto tři dokumenty byly vybrány protože reprezentují typické webové stránky jejichž obsah je zaměřen na poskytování informací o noteboocích. Recenze je příklad volnějšího textu, spravujícího o vlastnostech notebooku, další příklady jsou stránky eshopů, které často poskytují informace ve strukturovanější podobě (výčet specifikací modelu).

3 Principy gramatiky

3.1 Preprocess.jape

Cílem gramatiky v Preprocess.jape je předzpracování tokenů pro další extrakci. Názvy notebooků, CPU a GPU často obsahují alfanumerické řetězce (případně ještě doplněné o pomlčky). `ANNIE English Tokenizer` v programu GATE ovšem tyto řetězce rozděluje, například „MSI GE75“ bude rozděleno na posloupnost tří tokenů: „MSI“ „GE“ „75“. Z pohledu extrakčních gramatik je ovšem žádoucí pracovat nad celými alfanumerickými řetězci, tato gramatika tedy řetězec „GE75“ spojí do jednoho tokenu.

Dále obsahuje pravidlo které detekuje SpaceToken typu control a anotuje jej jako „Control“. Důvodem je, že při detekci entit nechceme, aby mohly přesahovat do jiné věty, nebo byly její složky vizuálně odděleny bílými znaky (to je případ výčtů specifikací). Tato anotace tedy slouží abychom mohly jako Input v gramatikách použít `SpaceToken.kind=="control"`.

Využité prvky: regulární výrazy nad řetězci, token.kind + token.string, opakování (?,+,*), disjunkce

3.2 Rozpoznávání modelů notebooků

Gramatika rozpoznávající název modelu notebooku nepracuje s žádným kontextem, v rámci trénovací množiny dokumentů nebyl společný kontext nalezen (s výjimkou možného levého kontextu „Model:“ v dokumentu ze stránky ebay, takovýto kontext ovšem pokládáme za nedostatečně specifický).

Místo toho tedy hledá sekvence tokenů začínající nějakou známou značkou výrobce notebooků. Tyto značky jsou definované v gazetteeru `pc_brands.lst`. Dále tato sekvence musí obsahovat minimálně jeden alfanumerický řetězec, který je případně obklopený omezeným počtem tokenů začínajících na velké písmeno.

Druhé pravidlo v rámci fáze slouží k odstranění anotací u chybně detekovaných řetězců. Například firma Microsoft vyrábí notebooky, ale i operační systém Windows. Pokud tedy detekovaný název notebooku obsahuje „Windows“, je tato anotace odstraněna.

Využité prvky: gazetteer, regulární výrazy nad řetězci, značky POS, opakování ([n,m]), disjunkce, oddělovače vět

3.3 Rozpoznávání grafických karet

Gramatika sloužící k rozpoznávání GPU obsahuje pravidla pro tři hlavní značky grafických karet notebooků: NVIDIA, AMD a Intel. Pravidlo pro každou z těchto značek je odvozené ze seznamu názvů grafických karet, kde se hledaly společné znaky a sekvence řetězců.

Dále je do gramatiky zařazené obecnější pravidlo, využívajícího pravého kontextu „graphics card“. Za GPU bude označený předcházející text o 3-5 tokenech, skládající se z tokenů začínajících na velké písmeno, nebo skládajících se z alfanumerických znaků (s případnou pomlčkou).

Využité prvky: regulární výrazy nad řetězci, opakování ([n,m]), disjunkce, oddělovače vět

Pozn.: Trénovací dokument HP_Stream_ebay.txt obsahuje název grafické karty „AMD Radeon VII“. Jedná se o novinku na trhu a netypický název GPU pro výrobky této společnosti. Tato entita není pokryta pravidly v gramatice, pokud ovšem začne AMD podobně pojmenovávat své nové produkty, je možné to do pravidel doplnit.

3.4 Rozpoznávání procesorů

4 Zhodnocení úspěšnosti

4.1 Striktní hodnocení

4.1.1 Modely Notebooků

V dokumentu `Dell_XPS_cnet.txt` bylo úspěšně nalezeno 24 modelů, neúspěšně 17 , žádné nebyly nalezené navíc. Přesnost je tedy 100 %, úplnost 58,5 %.

V souboru `top10_laptopunderbudget.txt` byla striktní úspěšnost mírně horší. 47 Modelů bylo nalezeno úspěšně, 35 neúspěšně a opět žádné navíc. Přesnost je tedy opět 100 %, úplnost 57,3 %.

4.1.2 Grafické karty

V dokumentu `Dell_XPS_cnet.txt` byla přesnost i úplnost 100 %.

V dokumentu `top10_laptopunderbudget.txt` se úspěšně našlo 20 názvů GPU, neúspěšně 5, žádné navíc. Přesnost je 100 %, úplnost 83 %.

4.2 Fuzzy hodnocení

4.2.1 Modely Notebooků

Ve většině případů v dokumentu `Dell_XPS_cnet.txt` kdy nebyla entita rozpoznána se jednalo o neúplný název (např. „the XPS 13“ místo „Dell XPS 13“. Pouze čtyři nerozeznané výskyty obsahovaly všechny potřebné informace (čtyřikrát „Apple MacBook Air(13-inch, 2018)“). Pokud by podmínkou bylo vyhledávat pouze kompletní názvy, stoupla by úplnost na 85,7 %.

Ve 21 případech se v souboru `top10_laptopunderbudget.txt` byl název modelu chybně označen jako delší (jedno až dvě slova). V dalších 10 případech se opět jednalo o neoznačení neúplného názvu. U obou těchto chyb byl ovšem z textu jinde extrahován i jejich celý název správně. Zcela neoznačené modely byly ve výsledku čtyři. Pokud bychom za chyby počítali pouze ty, stoupne úplnost na 95 %.

4.2.2 Grafické karty

3 ze 4 neodhalených výskytů v dokumentu `top10_laptopunderbudget.txt` vycházely ze zapsání značky Nvidia jako „NVidia“. Jedná se o nestandardní

(až chybný) způsob zápisu. Pokud by se tyto případy vůbec nebraly v potaz, byla by úplnost 95 %.

5 Rozbor přepoužití