Handleiding LeTs Preprocess demo

LeTS Preprocess

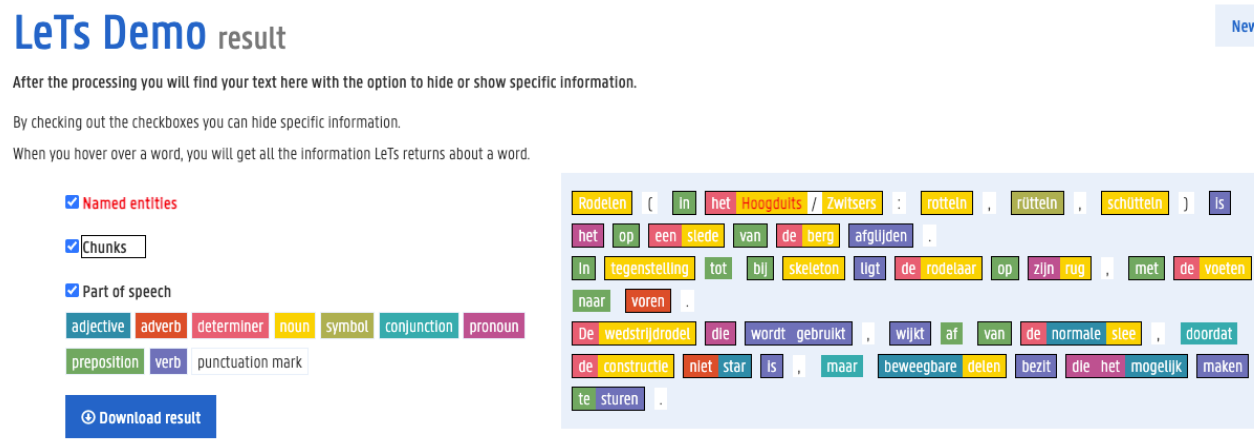
Met deze demo kan je gebruikmaken van LeTs Preprocess, een tool voor **automatische zinsontleding** met behulp van machinelearning. LeTs is momenteel getraind voor vier talen: Nederlands, Frans, Engels en Duits. Dat betekent dat je voor elk van die vier talen een tekst (van maximaal 1000 tekens) kan opladen en laten verwerken.

Aan de slag

Klik op 'Start the demo' om toegang te krijgen tot de LeTs interface. Kies daarna de taal van je tekst die je invoert in het dropdownmenu 'Language'. Kopiëer en plak je tekst in het vak 'Input text' of laad een bestand op via 'Upload file'. Let wel: de demo kan enkel .txt-bestanden verwerken. Als je tekst in een Word- of Excel-document wil verwerken, dan moet je het eerst opslaan als een .txt-bestand. Dat kan via 'opslaan als…' → 'tekst zonder opmaak (.txt)'. Let erop dat je 'Unicode (utf-8)' kiest als tekencodering.
Heb je de taal van je brontekst aangeduid en je tekst opgeladen? Klik dan op 'Submit' om de demo te laten starten.

Uitvoer

Na enkele seconden krijg je de uitvoer van de demo te zien. Rechts op het scherm zie je de ingevoerde tekst met verschillende kleuraanduidingen.



Figuur 1: uitvoer LeTs Preprocess.

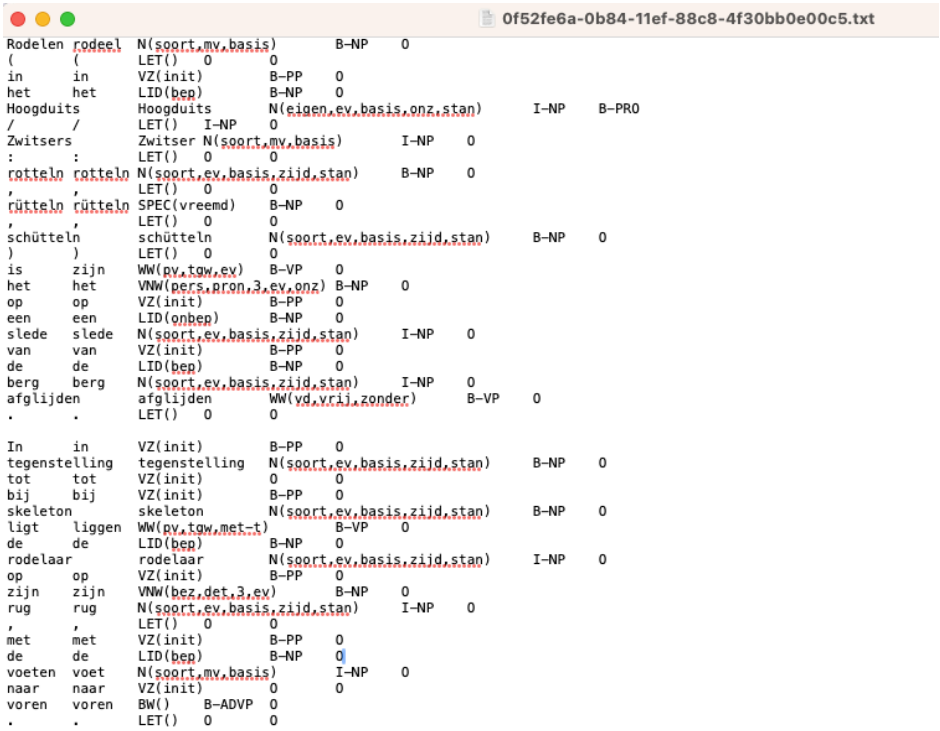
De zwarte kaders rond de woorden duiden zinsdelen (ook *chunks* genoemd) aan, zoals een naamwoordelijk gezegde (bijvoorbeeld: 'de rodelaar'). De individuele woorden (ook *tokens* genoemd) binnen een zinsdeel zijn gescheiden door spaties.

De achtergrondkleuren duiden aan tot welke woordsoort een specifiek zinsdeel of woord behoort. Zo worden zelfstandige naamwoorden aangeduid in het geel, voegwoorden in het appelblauwzeegroen en bijvoeglijke naamwoorden in het donkergroen.

Woorden die in het rood gedrukt staan (zoals 'Hoogduits' in dit voorbeeld), zijn aangeduid als eigennamen (ook *named entities* genoemd).

Voor een meer gedetailleerde uitvoer, klik je op 'Download results'. Met die knop ontvang je automatisch een .txt-bestand in de map 'Downloads'. Dat bestand kan je openen met een tekstverwerker als SublimeText, NotePad of Kladblok, maar ook met Word of Excel (klik rechts met de muis en selecteer 'open met…'). In het pop-upscherm dat je dan te zien krijgt, dien je wel 'utf-8' als tekencodering aan te duiden.

In Excel krijg je zo een mooi overzicht van de uitvoer dat je kan filteren zoals gewenst (bijvoorbeeld op woordsoort, eigennaam, …).



Figuur 2: LeTs uitvoer gedownload als .txt-bestand en geopend met Kladblok.

Analyse uitvoer

Het .txt-bestand dat je hebt gedownload, bevat vijf kolommen met informatie:

- In de eerste kolom staat je invoertekst, gesegmenteerd per *token* (i.e. een woord of een interpunctieteken). Tussen verschillende zinnen staat telkens een lege regel.
- In de tweede kolom vind je het lemma of de woordenboekvorm van het *token* in de eerste kolom.
- In de derde kolom staat de woordsoort van het *token* in de eerste kolom. Die woordsoorten zijn gebaseerd op de categorisatie in de bijbehorende manual (zie link naar PDF-bestand "Taxonomie_PoS-tags").
- In de vierde kolom zijn de *chunks* aangegeven. 'B' en 'O' staan voor 'begin' en 'outside'. 'B-PP' bij het derde token betekent dus: 'begin van een prepositional phrase' (en meteen ook het einde).
- In de vijfde en laatste kolom wordt aangegeven of het token in kolom 1 een eigennaam is, opnieuw met 'B' of 'O'. Als een token een eigennaam is, wordt het voorvoegsel 'B' gebruikt, gevolgd door een liggend streepje en het type eigennaam. Er worden in totaal zes types onderscheiden:
    1. **PRO =** product (commerciële producten, voertuigen, drukwerk, prijzen, diensten, …)
    2. **PER =** persoon (familienamen, voornamen, fictieve namen, artiestennamen, …)
    3. **ORG =** organisatie (overheidsoragniasties, commerciële organisaties, politieke partijen, …)
    4. **LOC =** locatie (landen, continenten, waterwegen, marktpleinen, luchthavens, pretparken, ziekenhuizen, fictieve locaties, …)
    5. **EVE =** evenement (wedstrijden, conferenties, prijzen, feestdagen, oorlogen, festivals, …)
    6. **MISC =** miscellaneous (bv. aansprekingen, wetten, woorden die ten onrechte met een hoofdletter zijn geschreven, …)

Contact

Vragen of problemen bij deze demo? Contacteer dan Michael.Lumingu@UGent.be.