

Handleiding LeTs Preprocess demo

LeTS Preprocess

Met deze demo kun je gebruikmaken van LeTs Preprocess, een tool voor **automatische zinsontleding** die gebruikmaakt van machinelearning. LeTs Preprocess is getraind voor vier talen: Nederlands, Frans, Engels en Duits. Je kunt voor elk van die vier talen een tekst (maximaal 1000 tekens, inclusief spaties) invoeren om die automatisch te laten verwerken.

Aan de slag

Klik op 'Start the demo' om toegang te krijgen tot de interface van LeTs. Volg dan de onderstaande stappen:

1. **Taalselectie:** Kies de taal van de ingevoerde tekst in het dropdownmenu 'Language'.
2. **Tekst invoeren:** Kopieer en plak je tekst in het veld 'Input text', of upload een tekstbestand door te klikken op 'Upload file'. Let op: de demo accepteert alleen platte tekstbestanden (met de extensie .txt). Teksten in andere formaten, zoals Word- of Excel-documenten, moeten eerst worden opgeslagen als een .txt-bestand. Dat kan je doen door te kiezen voor: 'Opslaan als...' → 'Tekst zonder opmaak (.txt)' Let erop dat je 'Unicode (utf-8)'. Zorg ervoor dat je 'Unicode (utf-8)' kiest als tekencodering.
3. **Verwerken:** Nadat je de taal hebt geselecteerd en de tekst hebt ingevoerd of geüpload, klik je op 'Submit' om de analyse te starten.

Verwerkingsstappen

LeTs Preprocess voert de volgende stappen uit:

1. **Zinssplitsing en tokenisering:** De tekst wordt automatisch verdeeld in zinnen en tokens. Een token kan een woord, een leesteken, een cijfer of een speciaal teken zijn (bv. hashtags of emoji's).
Voorbeeld:
 - o Originele tekst: "De prijzen? Die zijn gedaald met 22% 📉"
 - o Getokeniseerde versie: "de" / "prijzen" / "?" / "die" / "zijn" / "gedaald" / "met" / "22" / "%" / 📉
2. **Woordsoorttoekenning (POS-tagging):** Elk token krijgt een label dat de woordsoort (zoals zelfstandig naamwoord, bijvoeglijk naamwoord, werkwoord, enz.) aangeeft.
3. **Eigenaamherkenning (NER - Named Entity Recognition):** Het systeem herkent of een token een eigenaam is en bepaalt over welke categorie het gaat. Er worden zes types eigennamen herkend, die onderaan deze handleiding verder worden uitgelegd.
4. **Lemmatisering:** Elk woord wordt herleid tot de basisvorm (lemma). Bijvoorbeeld, "vriendjes" wordt herleid tot "vriend" en "zwom" wordt herleid tot "zwemmen".
5. **Chunking:** Op basis van de woordsoorten worden zinsdelen (chunks) aangeduid, zoals naamwoordelijke en werkwoordelijke gezegdes. Zo kunnen syntactische eenheden in de zin worden onderscheiden.

Uitvoer bekijken

Na verwerking wordt de uitvoer binnen enkele seconden op het scherm getoond. Rechts op het scherm verschijnt de tekst, waarbij verschillende visuele elementen worden gebruikt:

- **Zwarte kaders:** Die markeren zinsdelen (chunks), zoals een naamwoordelijk gezegde. Als woorden geen zwart kader hebben, dan werden ze door het systeem niet als 'chunks' beschouwd, zoals voegwoorden ('en', 'maar', 'of', ...).
- **Kleuren:** De achtergrondkleur van een woord of zinsdeel geeft de woordsoort aan. Bijvoorbeeld:
 - o **Paars:** werkwoord
 - o **Appelblauwzeegroen:** voegwoord
 - o **roze:** voornaamwoord
- **Rode tekst:** Eigennamen worden in rood aangeduid, zoals 'Hoogduits' in Figuur 1.

LeTs Demo result

Nev

After the processing you will find your text here with the option to hide or show specific information.

By checking out the checkboxes you can hide specific information.

When you hover over a word, you will get all the information LeTs returns about a word.

☒ Named entities

☒ Chunks

☒ Part of speech

adjective adverb determiner noun symbol conjunction pronoun

preposition verb punctuation mark

[Download result](#)

Rodelen (in het Hoogduits / Zwitsers : rotteln , rütteln , schütteln) is
het op een slede van de berg afglijden .
In tegenstelling tot bij skeleton ligt de rodelaar op zijn rug , met de voeten
naar voren .
De wedstrijdrodelaar die wordt gebruikt , wijkt af van de normale slee , doordat
de constructie niet star is , maar beweegbare delen bezit die het mogelijk maken
te sturen .

Figuur 1: uitvoer LeTs Preprocess.

Resultaten downloaden

Voor een meer gedetailleerde weergave van de resultaten kun je de uitvoer downloaden door te klikken op 'Download results'. Dat levert een .txt-bestand op dat je kunt openen in programma's zoals SublimeText, NotePad, Kladblok, Word of Excel. Als je het bestand opent in Word of Excel, zorg er dan voor dat je de tekencodering instelt op utf-8 voor een correcte weergave.

Tip voor Excel: In Excel kun je de resultaten gemakkelijk filteren op bijvoorbeeld woordsoort of eigennamen om een grondige analyse te maken.

```
Of52fe6a-0b84-11ef-88c8-4f30bb0e00c5.txt
Rodelen rodeel N(soort,mv,basis) B-NP 0
( LET() 0
in in VZ(init) B-PP 0
het het LID(bep) B-NP 0
Hoogduits Hoogduits N(eigen,ev,basis,onz,stan) I-NP B-PRO
/ / LET() I-NP 0
Zwitsers Zwitser N(soort,mv,basis) I-NP 0
: : LET() 0
, rotteln rotteln N(soort,ev,basis,zijd,stan) B-NP 0
, , LET() 0
, rütteln rütteln SPEC(vreemd) B-NP 0
, , LET() 0
, schütteln schütteln N(soort,ev,basis,zijd,stan) B-NP 0
) ) LET() 0
is zijn VV(pv,tow,ev) B-VP 0
het het VNW(pers,pron,3,ev,onz) B-NP 0
op op VZ(init) B-PP 0
een een LID(onbep) B-NP 0
slede slede N(soort,ev,basis,zijd,stan) I-NP 0
van van VZ(init) B-PP 0
de de LID(bep) B-NP 0
berg berg N(soort,ev,basis,zijd,stan) I-NP 0
afglijden afglijden VV(vd,vrij,zonder) B-VP 0
. . LET() 0

In in VZ(init) B-PP 0
tegenstelling tegenstelling N(soort,ev,basis,zijd,stan) B-NP 0
tot tot VZ(init) 0 0
bij bij VZ(init) B-PP 0
skeleton skeleton N(soort,ev,basis,zijd,stan) B-NP 0
ligt liggen VV(pv,tow,met-t) B-VP 0
de de LID(bep) B-NP 0
rodelaar rodelaar N(soort,ev,basis,zijd,stan) I-NP 0
op op VZ(init) B-PP 0
zijn zijn VNW(bez,det,3,ev) B-NP 0
rug rug N(soort,ev,basis,zijd,stan) I-NP 0
, , LET() 0
, met met VZ(init) B-PP 0
de de LID(bep) B-NP 0
voeten voet N(soort,mv,basis) I-NP 0
naar naar VZ(init) 0 0
voren voren BW() B-ADVP 0
. . LET() 0 0
```

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

Analyse van de uitvoer

Het gedownloade .txt-bestand bevat vijf kolommen met gedetailleerde informatie:

1. **Kolom 1:** De originele invoertekst, opgedeeld per token (woord, cijfer, leesteken). Lege regels scheiden zinnen van elkaar.

2. **Kolom 2:** Het lemma (de woordenboekvorm) van elk token.
3. **Kolom 3:** De woordsoort van elk token, gebaseerd op de categorisatie in de bijbehorende manual (zie link naar het overzicht op de hoofdpagina van de demo).
4. **Kolom 4:** De zinsdelen (chunks) worden aangegeven met de codes 'B' (begin) en 'O' (outside). Bijvoorbeeld, 'B-PP' bij het derde token in Figuur 2 betekent 'begin van een prepositional phrase'.
5. **Kolom 5:** Eigennamen worden aangeduid met 'B-' gevolgd door het type. Er worden zes types onderscheiden:
 - **PRO:** Product (bv. commerciële producten, voertuigen, diensten).
 - **PER:** Persoon (bv. familienamen, voornamen, artiestennamen).
 - **ORG:** Organisatie (bv. overheidsorganisaties, bedrijven, politieke partijen).
 - **LOC:** Locatie (bv. landen, steden, continenten, waterwegen).
 - **EVE:** Evenement (bv. wedstrijden, conferenties, oorlogen, festivals).
 - **MISC:** Diversen (bv. juridische termen, onjuiste hoofdletters).

Aanvullende tips:

Ervaar je problemen bij het opladen van een bestand? Controleer dan zeker het volgende:

- **Unicode-codering:** Zorg ervoor dat je altijd Unicode (utf-8) selecteert als tekencodering bij het opslaan van bestanden.
- **Tekstformaten:** De tool ondersteunt alleen tekstbestanden zonder opmaak (.txt). Andere bestandstypen moet je dus eerst converteren naar een .txt-bestand.

Contact

Technische problemen bij deze demo of vragen over de tool zelf?

Contacteer dan Orphee.DeClercq@UGent.be en Cynthia.VanHee@UGent.be.