TreeTagger Tutoriel

1. Télécharger TreeTagger : https://www.cis.lmu.de/~schmid/tools/TreeTagger/data/tree-tagger-windows-3.2.3.zip

2. Extraire les fichiers compressés et déposer le fichier « TreeTagger » dans C:/Programmes

3. Télécharger les fichiers « .par.gz » pour les langues voulues sur le lien suivant : https://www.cis.lmu.de/~schmid/tools/TreeTagger/#Windows

Parameter files

- Bulgarian parameter file (gzip compressed, UTF-8, tagset documentation, trained on the Bulgarian Treebank)
- Catalan parameter file (gzip compressed, UTF8, tagset documentation)
- A Chinese parameter file and tokenizer created by Serge Sharoff are available here
- A Coptic parameter file created by Amir Zeldes is available here
- Czech parameter file (gzip compressed, UTF-8, trained on the Czech Academic Corpus)
- Danish parameter file trained on the ePAROLE corpus (gzip compressed, UTF-8, tagset documentation)
- Dutch parameter file (gzip compressed, UTF-8, tagset documentation)
- Another Dutch parameter file (gzip compressed, UTF8, trained on the Eindhoven corpus, tagset documentation (starts on page 9))
- English parameter file (PENN tagset) (gzip compressed, UTF8, tagset documentation, trained on the Penn treebank)
- English parameter file (BNC tagset) (gzip compressed, UTF8, tagset documentation, trained on the British National Corpus)
- Estonian parameter file (gzip compressed, UTF-8, tagset documentation)
- Finnish parameter file trained on the Finnish Treebank (gzip compressed, UTF-8, tagset documentation).
- French parameter file (gzip compressed, UTF-8, tagset documentation)
- Spoken French parameter file (gzip compressed, UTF-8, tagset documentation) trained on the Perceo corpus
- A parameter file for spoken French texts can be found here
- Old French parameter file (gzip compressed, UTF-8, tagset documentation) trained on the Base de Français Médiéval
- Galician parameter file (gzip compressed, UTF8, tagset documentation)
- German parameter file (gzip compressed, UTF-8, tagset documentation)
- Spoken German parameter file (gzip compressed, Latin-1, tagset documentation)
  trained on the FOLK corpus provided by the Institut für Deutsche Sprache (IDS) Mannheim
- Middle High German parameter file trained by Sarah Schulz on the Middle High German Conceptual Database (gzip compressed, UTF-8, paper (in German))
- Greek parameter file trained on the INTERA corpus (gzip compressed, UTF8, tagset documentation)
- Ancient Greek parameter file (UTF8 encoding or beta encoding) trained on the PROIEL and Perseus treebanks and kindly provided by Alessandro Vatri and Barbara McGillivray (gzip compressed, no ler
- A Hausa parameter file created by Amir Zeldes is available here
- Hungarian parameter file (gzip compressed, UTF8, trained on data annotated with magyarlanc)
- The Indonesian parameter file (gzip compressed, UTF8, tagset documentation) has been trained by Prihantoro on the UI corpus using lexical information from the Kateglo dictionary.
- Italian parameter file (gzip compressed, UTF8, tagset documentation)
- Marco Baroni's Italian parameter file (gzip compressed, Latin1, tagset documentation)
- Korean parameter file (gzip compressed, UTF8, tagset documentation). This parameter file was created in joint work with Prof. Lee Minhaeng on data kindly provided by the KLPLAB headed by Prof. O
- Latin parameter file (gzip compressed, tagset info in Italian)
  The corpus and lexicon for training the Latin parameter file have been compiled by Gabriele Brandolini from various resources
- Another Latin parameter file (gzip compressed, tagset info) which has been trained on the Index Thomisticus Treebank which was kindly provided by Marco Passarotti.
- Mongolian parameter file (gzip compressed) created from a small Mongolian corpus by Khuder Altangerel.
- Norwegian (Bokmaal) parameter file trained on the Norwegian Dependency Treebank (gzip compressed, UTF-8) with tags mapped to the universal dependency tagset
- Persian (Farsi) parameter file trained on the Persian Dependency Treebank (gzip compressed, UTF8, tagset description).
- Persian (Farsi) parameter file with coarse tagset trained on the Persian Dependency Treebank (gzip compressed, UTF8, tagset description).
- Polish parameter file trained on the Polish National Corpus (gzip compressed, UTF8, tagset description).
- Portuguese parameter file provided by Pablo Gamallo (gzip compressed, UTF8, tagset description).
- Portuguese parameter file with fine-grained tagset provided by Pablo Gamallo (gzip compressed, UTF8, tagset description).
- Another Portuguese parameter file trained on the Floresta Sintá(c)tica corpus and the Unitex lexicon (gzip compressed, UTF8).
- Romanian parameter file (gzip compressed, UTF8, tagset, created with the help of Cristian Chirita using a MULTEXT-East corpus and lexicon)
- Russian parameter file (gzip compressed, UTF8, tagset trained on a corpus created by Serge Sharoff)
- Slovak parameter file (gzip compressed, UTF8)
  The Slovak parameter file was trained on the Slovak National Corpus. The tagset was simplified.
- Slovak parameter file (full tags) (gzip compressed, UTF-8)
  The Slovak parameter file was trained on the Slovak National Corpus. The tagset was not simplified (just a marker for typos was removed). Many thanks to Vladimir Benko for suggesting to train on the
- Slovenian parameter file (gzip compressed, UTF-8)

4. Décompresser les fichier .par.gz pour obtenir des fichiers .par (à l'aide de 7zip ça marche très bien)
5. Déposer les fichiers .par dans TreeTagger/lib

Dans la console anaconda exécuter la commande suivante :

- `pip install treetaggerwrapper`

Le fichier « treetaggerwrapper.py » doit être dans le même dossier que le fichier de code à exécuter.

Pour l'utiliser dans un notebook, l'import suivant est requis :

- `import treetaggerwrapper`

Documentation : https://treetaggerwrapper.readthedocs.io/en/latest/