Text and Sequential Pattern Mining

Text Mining Practical Work

Solen Quiniou

solen.quiniou@univ-nantes.fr

IUT de Nantes

Master 2 DS – 2020-2021

(last update: January 5, 2021)

Outline

Outline

Main Frameworks for Text Mining

- Python is the programming language most commonly used for Text Mining
  - → The most known and used frameworks are: Gensim, NLTK, and spaCy [1]

gensim
topic modelling for humans

- 2009
- Tokenization,
- Topic Modeling (LDA, HDP, LSI),
- Stemming,
- Word Embedding.

Natural Language Analyses with NLTK

- 2001
- Tokenization de mots et de phrases,
- POS-Tagging,
- NER,
- Analyse de sentiments,
- Stemming,
- Algorithmes de classifications,
- Corpora de données.

spaCy

- 2015
- Tokenisation,
- POS-Tagging,
- NER,
- Analyse de sentiments (toujours en développement),
- Lemmatisation,
- Vecteurs de mots pré-entraînés.

[1] https://www.ekino.com/articles/introduction-au-nlp-partie-ii

Comparison between NLTK and spaCy[2]

	⊕ PROS	⊖ CONS
Natural Language ToolKit	+ The most well-known and full NLP library + Many third-party extensions + Plenty of approaches to each NLP task + Fast sentence tokenization + Supports the largest number of languages compared to other libraries	– Complicated to learn and use – Quite slow – In sentence tokenization, NLTK only splits text by sentences, without analyzing the semantic structure – Processes strings which is not very typical for object-oriented language Python – Doesn't provide neural network models – No integrated word vectors
spaCy	+ The fastest NLP framework + Easy to learn and use because it has one single highly optimized tool for each task + Processes objects; more object-oriented, comparing to other libs + Uses neural networks for training some models + Provides built-in word vectors + Active support and development	– Lacks flexibility, comparing to NLTK – Sentence tokenization is slower than in NLTK – Doesn't support many languages. There are models only for 7 languages and "multi-language" models

[2] https://activewizards.com/blog/comparison-of-python-nlp-libraries/

Outline

Setup of the working environnment

- Installing spaCy and the language models
  - The website `https://spacy.io/usage` gives all the needed information to install spaCy as well as the language models in English
  - The given cheat sheet gives you additional information on using spaCy

- Data
  - A sample of positive movie reviews (10 files)
  - A sample of negative movie reviews (10 files)

Practical work using spaCy

1. Work on one sentence
   - Choose one sentence from one of the movie review files
     - Use spaCy to display the lemmas of the words of the sentence
     - Use spaCy to display the part-of-speech of the words of the sentence
   - Write a function that takes a sentence as the input and that outputs the sequence of itemsets corresponding to the sentence where each itemset contains a word, its lemma (if different from the word), and its part-of-speech

2. Work on one movie review file
   - Write a function that takes a movie review file as the input and that outputs a file where each line corresponds to the sequence of itemsets of the associated sentence of the input file, using the previous function

3. Work on several movie review files
   - Write a function that takes a directory of movie review files as the input and that outputs a directory of movie review files containing sequences of itemsets, using the previous function