TSLL TechXplorations 2017

TextFeaturesExtractor: Python tool for text mining and corpus analysis

Contributors: Sowmya Vajjala and Sagnik Banerjee
Iowa State University, USA
(tool under development)

Background: We are developing a small tool (command line interface, not graphical) to extract different kinds of textual features such as word, part-of-speech sequences and their frequencies and frequencies of various syntactic relations in a text, and we support different forms of pre-processing (stemming, spelling correction etc.).
The goal of the tool is two-fold:

1. support more linguistic corpus analysis
2. support text mining researchers by providing a suite of text features they can use to benchmark text classification. The tool is still under development, and is currently seen as a single large python file with some documentation. Our goal is to release the code for public use by December.

3 Takeaways from this session

1. learning about how to extract different kinds of textual features beyond words and word sequences
2. walking through the process of how to extract such features
3. sharing the python code so that enthusiastic people can try to use it and give feedback for future improvement of the tool

Current version of the code is available for download at: `https://goo.gl/gRF4X9`

Currently supported features (for any text file)

- Extraction of word, character and POS n-grams for any n- and their frequencies
- Skip grams (i.e., n-grams with gaps)
- Word-POS mixed n-gram representations
- Perform phrase level chunking and syntactic (dependency) parsing and collect frequencies of most common syntactic structures
- Pre-processing: lowercasing, stemming, spelling correction, stop word removal

Planned Extensions:

- Working on folders containing many files
- Storing output in human readable formats
- Storing output to be used as input for text mining and machine learning algorithms
- Add support for vectorized representations of words, beyond n-grams

contact: sowmya@iastate.edu