com6513-cwi-tom-dakin

Complex Word Identification in English and Spanish

A classifier to identify complex words in the **CWI Shared Task 2018** https://sites.google.com/view/cwisharedtask2018/home . The main datasets are included in ./datasets
A frequency index for Spanish words is included here. This is derived from the work of Matthias Buchmeier - https://en.wiktionary.org/wiki/User:Matthias_Buchmeier .

Dependencies

- sci-kit learn
- spacy - en_core_web_lg and es_core_news_md models
- nltk - wordnet corpus
- random, numpy, sys, csv

Usage

- Run `python demo.py` to train and test English and Spanish CWI models on the datasets included.
- Use command-line argument `cv` to run cross-validation