SIR: Style & Identity Recognition

Jan Rygl

rygl@fi.muni.cz

NLP Centre, Masaryk University, Brno, Czech Republic

4th December 2015

Goals

- Solve authorship recognition and other stylometric problems
- Data analysis functions
- Use external data as they are
- Open-source (GitHub)
- Stand-alone python (pip) package

Solve authorship recognition and other stylometric problems

Status: 30 %

- Problem is defined by a document label and a stylometric feature selection
- Implemented a python API and a console script interface
- Implemented a label-driven attribution
- TODO: a label-driven verification
- TODO: a label-driven clustering
- TODO: a web interface

Data analysis functions

Status: 20 %

- Implemented data filtering (remove rare or predominant labels)
- Implemented data splitting (divide data into tune/train/test data sets)
- TODO: visualize data as a graph
- TODO: GUI with a parameter slider
- TODO: cluster data by common features

Use external data

Status: 70 %

Data acquisition:

- XML format: unpack data from https://nlp.fi.muni.cz/projekty/acb/ and use them
- TODO: other formats

Data storing:

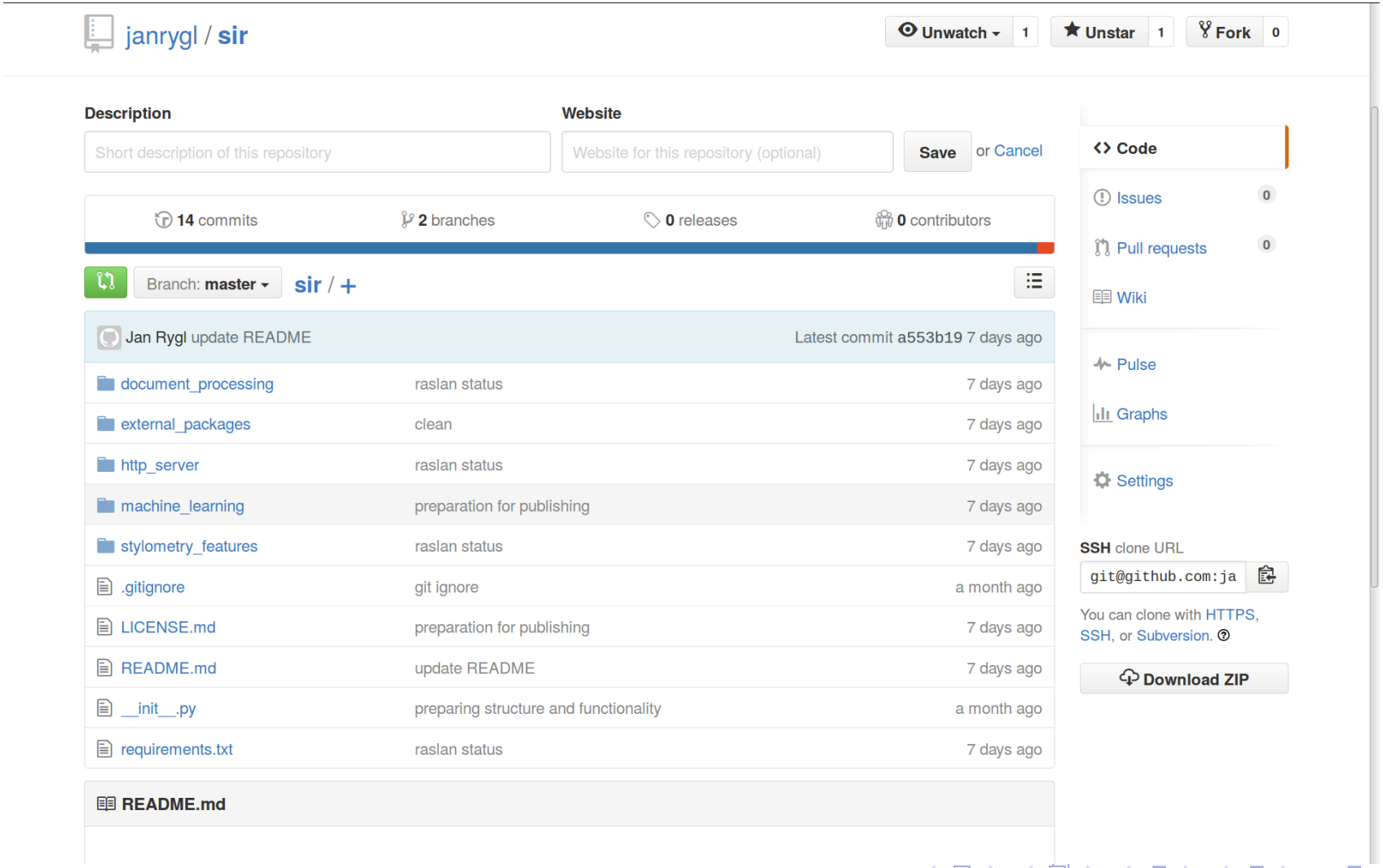
- Switched to a `sqlite3` database
- Database is located in the data folder (copying raw data will also copy preprocessed documents)
- Documents are cleaned, preprocessed and stored as dictionaries into the database (`sqlitedict` python package)

Open-source

Status: 90 %

- Published at GitHub: https://github.com/janrygl/sir
- Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License
- TODO: prepare a pip package and find a good unique name
- Focus on parallelism, efficiency and PEP8
- I had to completely rewrite the SIR tool
- New characteristics can be written by students as pull requests

GitHub



janrygl / sir

Unwatch 1 | Unstar 1 | Fork 0

Description Short description of this repository

Website Website for this repository (optional)

Save or Cancel

14 commits | 2 branches | 0 releases | 0 contributors

Branch: master | sir / +

Jan Rygl update README — Latest commit a553b19 7 days ago

document_processing	raslan status	7 days ago
external_packages	clean	7 days ago
http_server	raslan status	7 days ago
machine_learning	preparation for publishing	7 days ago
stylometry_features	raslan status	7 days ago
.gitignore	git ignore	a month ago
LICENSE.md	preparation for publishing	7 days ago
README.md	update README	7 days ago
__init__.py	preparing structure and functionality	a month ago
requirements.txt	raslan status	7 days ago

README.md

Code | Issues 0 | Pull requests 0 | Wiki | Pulse | Graphs | Settings

SSH clone URL git@github.com:ja

You can clone with HTTPS, SSH, or Subversion.

Download ZIP

Stand-alone python (pip) package

Status: 90 %

- ideally, all dependencies should be available as `pip` packages:
  requirements.txt

  Pattern==2.6; gensim==0.12.1; ipython==3.1.0; numpy==1.9.2;
  requests==2.7.0; scikit-learn==0.16.1; scipy==0.15.1; smart-open==1.2.1;
  sqlitedict==1.4.0; xmltodict==0.9.2; langid==1.1.5; Flask==0.10.1;
  Flask-Cors==2.0.1; Flask-Mako==0.3; gunicorn==19.3.0; argparse==1.2.1;
  nltk==3.0.2
- Exception 1: **Chared** (no `pip` package, but one of authors 'Prof.' Suchomel gave me permission to bundle it in the software)
- Exception 2: **RFTagger** (due licence restrictions, use have to download it himself), if not present, morphology based features are omitted and tokenization is done by `pattern`.

Removed tools

- Majka (incomplete data in a free version and a missing disambiguation)
- Desamb (cannot be freely shared)
- cz_accent (cannot be freely shared)
- Set (no `pip` package, probably will be added later)

Inspiration – pip package Pattern

Pattern has tools for data mining (Google, Twitter and Wikipedia API, a web crawler, a HTML DOM parser), natural language processing (part-of-speech taggers, n-gram search, sentiment analysis, WordNet), machine learning, network analysis and canvas visualization.

Free, well-document and bundled with 50+ examples and 350+ unit tests.

Acknowledgements

This presentation has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.