TA Introductory Session

Arun Kumar Rajasekaran

Python

https://www.python.org/downloads/

Jupyter

https://jupyter.org/

Google colab

https://colab.research.google.com/?utm_source=scs-index#scrollTo=C4HZx7Gndbrh

Syllabus

1 Overview of Natural Language Processing

2 Word: Morphology and Part-of-Speech

3 Supervised machine learning and Perceptron

4 Phrase structure and dependency

5 Transition-based dependency parsing

Project: Multilingual dependency parsing

Coding standards

• Coding standards are guidelines for code style and documentation.

• They may be formal (IEEE) standards, or company specific standards.

• The aim is that everyone in the organization will be able to read and work on the code.

• Coding standards cover a wide variety of areas:

– Program design

– Naming conventions

– Formatting conventions

– Documentation

– Use (or not) of language specific features

• **Why bother with a coding standard?**

– Consistency between developers

– Ease of maintenance and development

– Readability, usability

• Example should make this obvious!

• No standard is perfect for every application.

– If you deviate from the standard for any reason, document it!

Coding style

• There are several examples of coding styles. Often they differ from company to company

• They typically have the following in common:

– Names

• Use full English descriptors

• Use mixed case to make names readable

• Use abbreviations sparingly and consistently

• Avoid long names

• Avid leading/trailing underscores

– Documentation

• Document the purpose of every variable

• Document why something is done, not just what

Coding style

– Accessors

• Use getX(), setX() functions on all class variables.

– Member function documentation

• What & why member function does what it does

• Parameters/return value

• How function modifies object

• Preconditions/postconditions

• Concurrency issues

• Restrictions

– Document why the code does things as well as what it does.

Standards

• Standards rare documented agreements containing technical specifications or other precise criteria to be used consistently as guidelines, rules, or definitions of characteristics, to ensure that materials, products, processes and services are for for their purpose.

• International standards are supposed to contribute to making life simpler, and to increasing reliability and effectiveness of the goods and services we use.

• Standards represent best, or most appropriate, practice:
– They encapsulate historical knowledge often gained through trail and error.
– They preserve and codify organizational knowledge and memory
– They provide a framework for quality assurance.
– Ensure continuity over a project’s lifecycle.

Standards

• There are many industry standards governing all aspects of software development:

– Terminology

– Notation

– Requirements gathering

– Design

– Coding

– Documentation

– Human computer interaction

– Verification and validation

– Quality assurance

– Even ethics!

Who writes standards?

– ISO

International Organization for

Standardization

– SAA

Standards Australia

– BSI

British Standards Institute

– ANSI

American National Standards Institute

– IEEE

Institute for Electronic and Electrical

Engineers

– And about 80 or so others!

Relevant standards

• ISO 646 – 7-bit ASCII with national variants

• ISO 8859 – several 8-bit ASCII extensions:

– ISO 8859-1: West European languages (Latin-1)

– ISO 8859-2: East European languages (Latin-2)

– ISO 8859-5: Latin/Cyrillic

• ISO 6429 – ASCII control codes

• ISO 2382 – Information technology vocabulary

• ISO 8652 – the Ada programming language

• ISO 9899 – the C programming language

• ISO 9660 – CD-ROM volume and fie structure

• ISO 3166 – codes for the representation of names of counties:

– Defines a 2-letter, 3-letter and numeric code for every

country.

– US/USA/840 = United States

– GB/GBR/826 = United Kingdom

• The 2-letter codes are well known as the internet top level domain names.

NLTK

NLTK is a toolkit build for working with NLP in Python.

- Tokenization
- Lower case conversion
- Stop Words removal
- Stemming
- Lemmatization
- Parse tree or Syntax Tree generation
- POS Tagging

Alternatives

Polyglot

Polyglot is a python library for NLP which is especially useful as it supports extensive multilingual applications. According to the documentation of polyglot, it supports tokenization for 165 languages, language detection for 196 languages, part-of-speech tagging for 16 languages and sentiment analysis in more than 130 languages.

So it might be useful if someone is working with non-mainstream language. Also, it works very fast as it uses NumPy.

Alternatives

SpaCy

SpaCy is a Python NLP library that can be useful specifically for industry-level real-world projects containing huge amounts of text data.

The main advantage of using this library is its speed. SpaCy is much faster than other libraries as it is written in Cython which also makes it capable of handling larger amounts of data efficiently.

Support for more than 64 languages, 60 + train pipelines for 19 languages, multi-task learning with pre-trained transformers like BERT and support for modern ML/DL frameworks like Pytorch and Tensorflow makes SpaCy — a good choice for professional projects.

Alternatives

GenSim

GenSim is an NLP library written in Python that is popular due to its amazing speed and memory optimization. All the libraries used in GenSim are memory independent and can run easily with data sets of large size also. It comes with mini useful NLP algorithms like random projections (RP), latent semantic analysis (LSA), hierarchical Dirichlet process (HDP) etc.

GenSim uses SciPy and NumPy for computing and is used in applications like chatbots and semantic search applications etc.

Alternatives

Textblob

Textblob is a Python library that is powered by NLTK. It provides almost all the functionalities of NLTK but in a much simpler and beginner-friendly manner and its API can be used for some common tasks like classification, translation, word inflexion etc.

Many data scientists also use textblob for prototyping as it is much more lightweight to work with.

Alternatives

PyNLPI

PyNLPI also pronounced as pineapple is a Python NLP library that is mainly used for building basic language processing models. It is divided into different models and packages which can be used for different varieties of NLP tasks. One of the most prominent features of PyNLPI is that it comes with an entire library for working with FoLiA XML(format for linguistic annotation)

Alternatives

Pattern

Pattern is a multipurpose Python library that can be used for different tasks like natural language processing (tokenization sentiment analysis POS tagging etc.), Data mining from websites and machine learning using built-in models such as K-nearest Neighbors, Support Vector Machine etc.

This library is easy to understand and implement for beginners due to its simple and straightforward syntax and it is also helpful for web developers who need to work with text data.

Early starters

https://colab.research.google.com/github/gal-a/blog/blob/master/docs/notebooks/nlp/nltk_preprocess.ipynb