Big Data techniques with Python

Project and overview of the relevant technologies

Supervisor: Dr. hab. Inż. Dariusz Król
Presented by: Oleksii Kyrylchuk 223224
17.11.2017

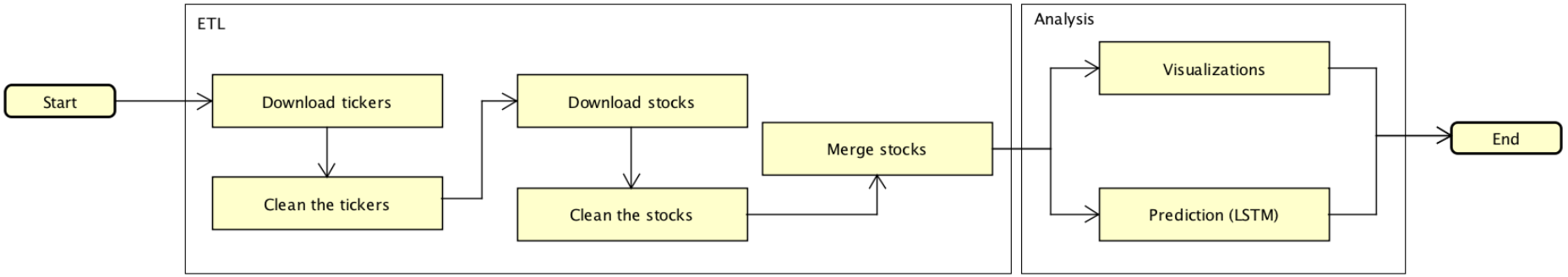
HR EXCELLENCE IN RESEARCH

Wrocław University of Science and Technology

Project Overview

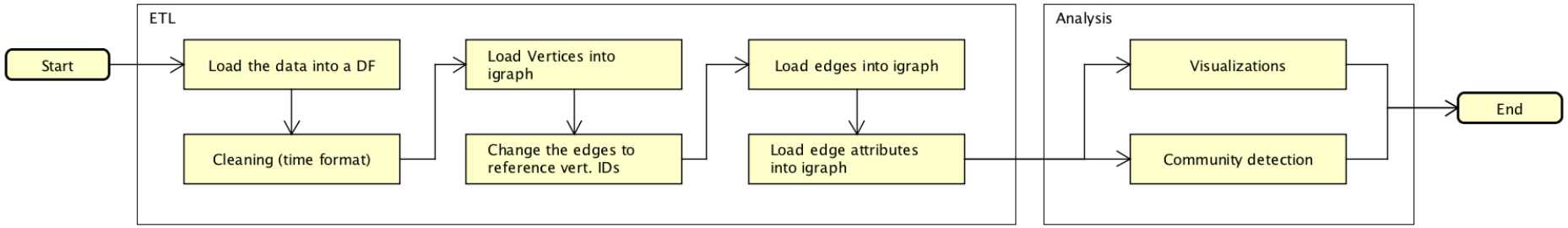
Stock data analysis (NYSE)

Goal: Demonstrate how to handle time series data in Python



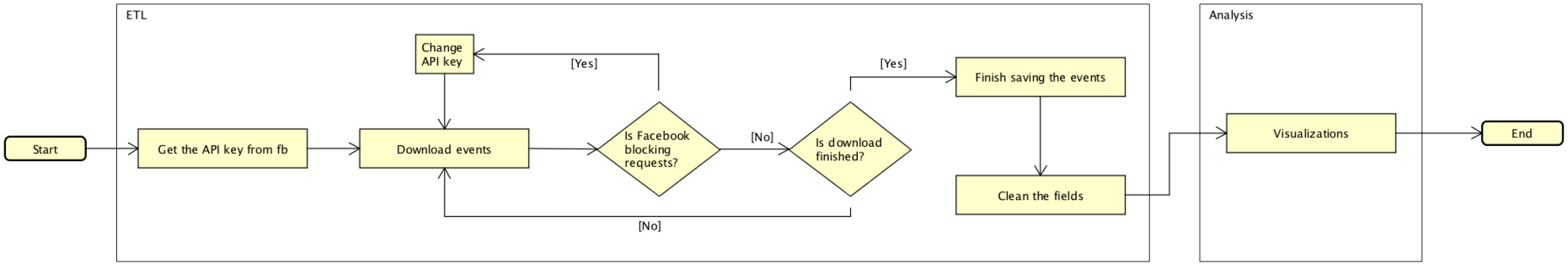
YouTube network analysis

Goal: Demonstrate how to handle network data in Python



Facebook geodata analysis

Goal: Demonstrate how to handle geodata in Python



Requirements and more

FURPS

- Functional
  - Code should be able to produce results similar to those found in other papers
- Usability
  - The experiments' code must be well-explained in the thesis
- Reliability
  - Code shouldn't crash on its own
  - (I am not responsible for the website uptime)
- Performance
  - Experiments should be able to finish in less than a day on a suitable server
- Supportability
  - The code should be written with the future (python3) in mind

Verification

- NYSE Stock market (time series data)
  - Charts look good and are understandable
  - Prediction is better than random
- YouTube network (network data)
  - Graphs look similar to those in other papers
- Facebook locations (geolocation data)
  - Charts look good and are understandable

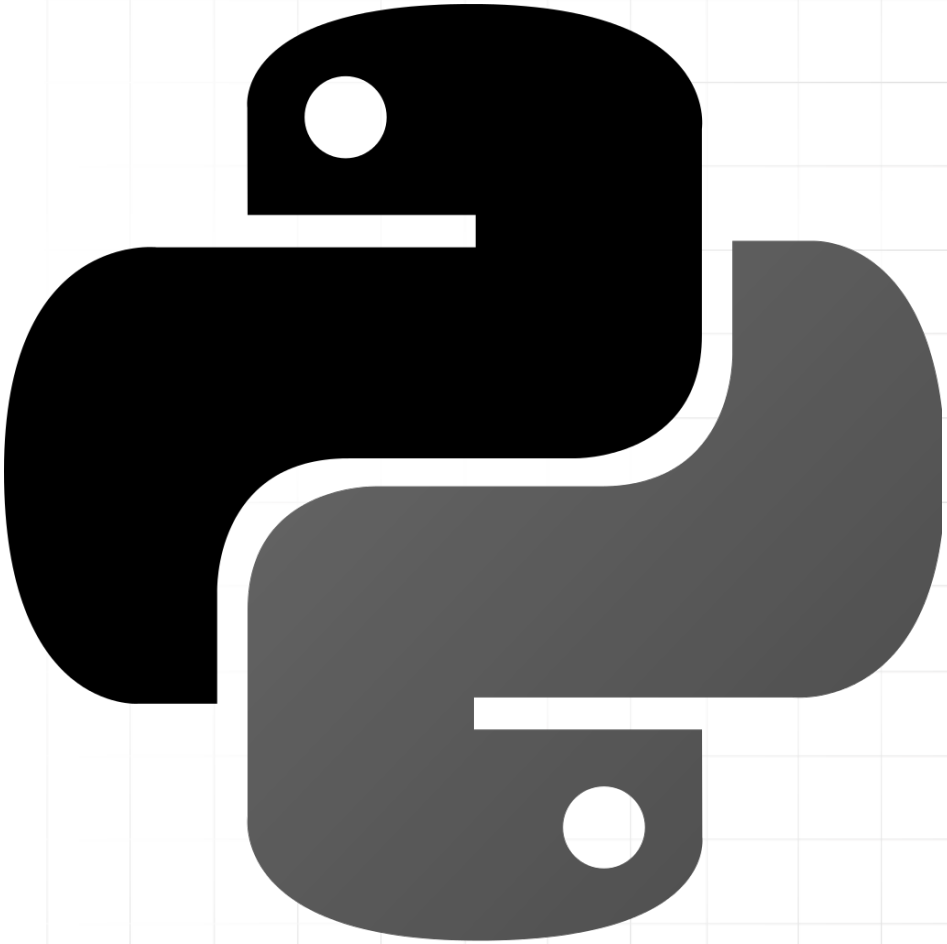
Methodology

- Following the PEP 8 code standard
- Trying to make code as simple as possible so the reader can easily understand it

Technology overview

Language

- Python 3
- Async processing is only available in Python3, as well as some other libraries
- Python2 will not be maintained past 2020



Processing server

- FloydHub[1]
  - Made for data science programming in Python
  - Provides useful CLI tools to upload data to their servers
  - Can run Jupyter notebooks on GPU clusters
  - Easier to get into than AWS[2] or Google cloud[3]

[1] https://www.floydhub.com/
[2] https://aws.amazon.com/
[3] https://cloud.google.com/

Data retrieval

- Stocks
  - requests[1] – for getting stock tickers
  - BeautifulSoup[2] – for parsing web pages
  - Pandas DataReader[3] – for getting the stocks themselves
- YouTube
  - Pre-generated dataset from KONECT[4]
- Facebook
  - aiohttp[5] (asyncio) – to get locations asynchronously

[1] http://docs.python-requests.org/
[2] https://www.crummy.com/software/BeautifulSoup/
[3] https://pypi.python.org/pypi/pandas-datareader
[4] http://konect.uni-koblenz.de/networks/youtube-u-growth
[5] https://aiohttp.readthedocs.io/en/stable/

Data transformation & cleaning

- Numpy[1] – performant arrays, math operations
- Pandas[2] – access to statistic functions and easy transformations
- Built-ins – loop comprehensions, itertools[3], …

[1] http://www.numpy.org/
[2] http://pandas.pydata.org/
[3] https://docs.python.org/3/library/itertools.html

Data storage

- Stocks
  - Pandas (CSV files) – DataReader provides you with CSV files
- YouTube
  - igraph[1] files
- Facebook
  - SQLAlchemy ORM[2] – Load geodata from facebook into a DB

[1] http://igraph.org/python/
[2] https://www.sqlalchemy.org/

Data visualization (current)

- Matplotlib[1] – The base for many python plotting packages
  - Seaborn[2] – High-level interface for matplotlib
  - Pandas – Basic plots that don't require importing an extra library
- Datashader[3] – Geodata visualization
- Bokeh[4] – interactive plots that run in browser
- More?

[1] https://matplotlib.org/
[2] http://seaborn.pydata.org/
[3] http://datashader.readthedocs.io/
[4] https://bokeh.pydata.org/

Data Analysis

- igraph – network analysis

- Keras[2] – predictive analysis, machine learning

[1] https://keras.io/

Thank you for your attention!