Assignment 3: Scientific Area And Similarity Classifier (SAASC)

Advanced Information Retrieval 22/23

Group 23:
Alma Emkic (Developer/Analyser)
Julian Jautz (Architect/AI-Model)
Christina Mandlez (Developer/Visualization)
Paul Scheibelmasser (Project Manager/AI-Dataset)

09.01.2022

https://github.com/pscheibel/AIR_ASSIGNMENT3

Introduction

Research Questions:

(Goals) Motivation:

- Get training data via Arxiv.
  - Automated retrieval of recent papers for various categories
- Implement Network classifying scientific belonging
  - Extract nouns via text processing (textblob)
  - Creation and training of Classification Network based on nouns
- Classify and Compare any Papers (URLs).
  - Input:
    - URL's of papers where the Network is applied on
  - Output:
    - Categories per URL
    - Plot of training/network statistic
    - Plots of Similarity

Data + Methods TODO(optional theoretical background)

- papers from https://arxiv.org/
  - Retrieves automatically x newest papers of configured categories y categories.
  - Downloads content of URL and processes its nouns into labelled files which are stored via caching.
  - Creates labelled dataset for machine learning (Categorization)

  - cs
    - https://arxiv.org/list/cs/pastweek?show=1000
  - q-bio
    - https://arxiv.org/list/q-bio/pastweek?show=1000
  - physics
    - https://arxiv.org/list/physics/pastweek?show=1000
  - eess
    - https://arxiv.org/list/eess/pastweek?show=1000
  - econ
    - https://arxiv.org/list/econ/pastweek?show=1000

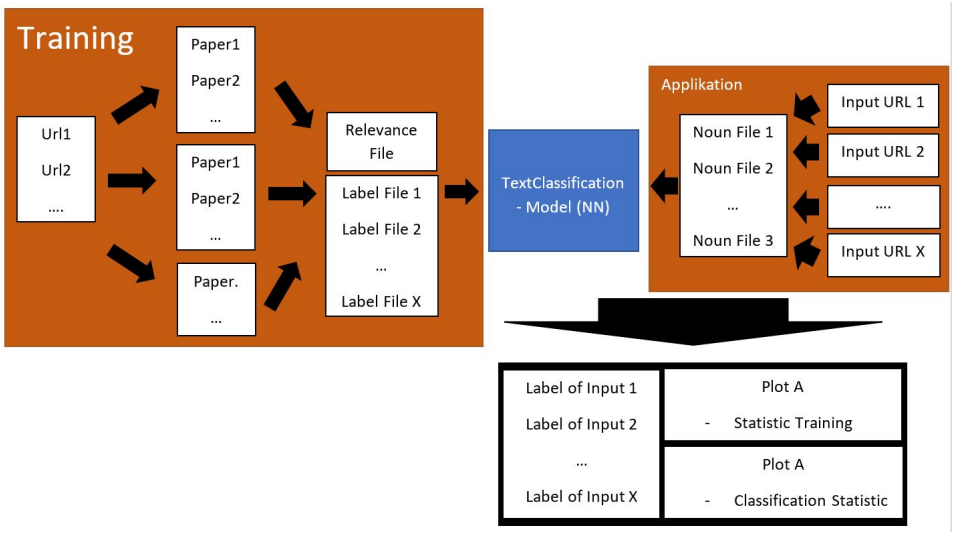
Data + Methods TODO(optional theoretical background)

- automatically retrieving papers from https://arxiv.org/
  - classified into five scientific areas

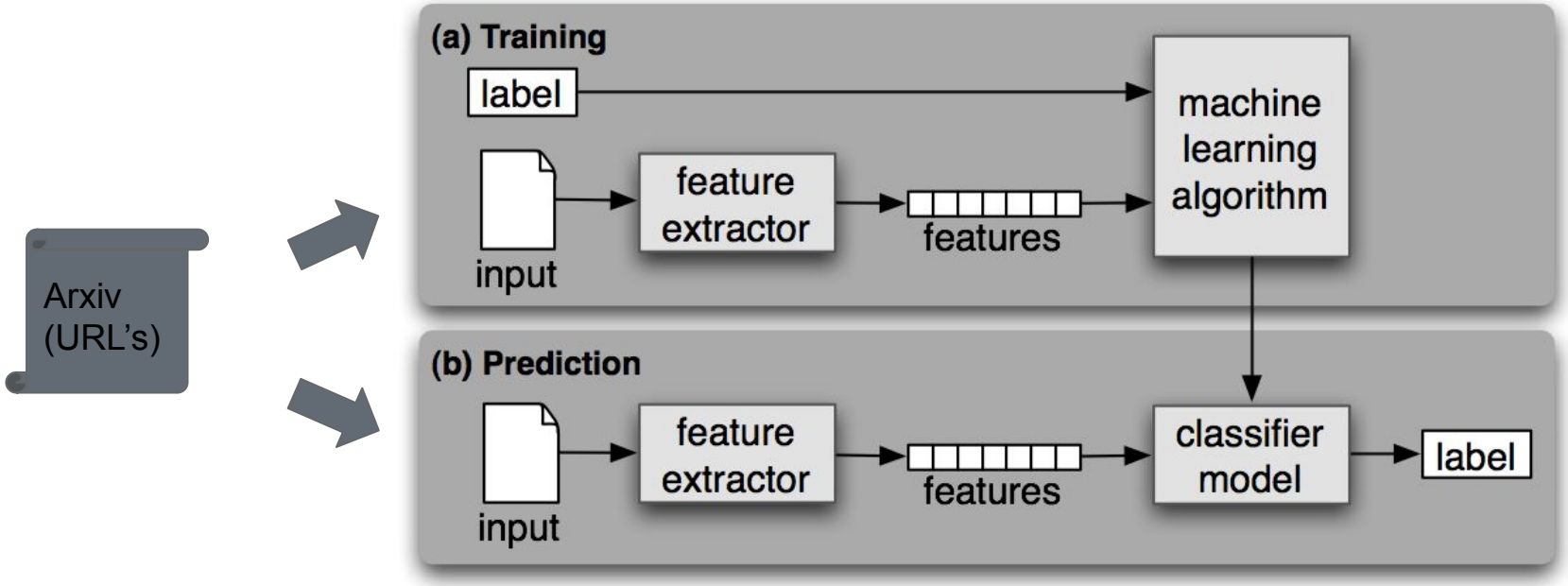
  - cs
    - https://arxiv.org/list/cs/pastweek?show=1000
  - q-bio
    - https://arxiv.org/list/q-bio/pastweek?show=1000
  - physics
    - https://arxiv.org/list/physics/pastweek?show=1000
  - eess
    - https://arxiv.org/list/eess/pastweek?show=1000
  - econ
    - https://arxiv.org/list/econ/pastweek?show=1000

Architecture (TODO adapt to AIR)

- Technologies
  - Python (Pycharm)
  - Neural Network
- Algorithms
  - Noun extraction
  - Neural Network
    - Loss Function: CrossEntropyLoss
    - Optimizer: SGD
- Libraries
  - Textblob
  - PyPdf2
  - torch
  - urllib.request
  - BeautifulSoup



Design



https://www.nltk.org/book/ch06.html

Results - Analysis and Interpretation (TODO Add Resulst!!)

Literature comparison:

- How I achieved 90% accuracy on a text classification problem with ZERO preprocessing
  - BERT sentence embeddings
  - used Spark NLP
  - 4 categories
  - Accuracy: 90%
- Text Classification with TF-IDF, LSTM, BERT: a comparison of performance
  - 5 categories
  - TF-IDF (97.9%)
  - Recurrent Neural Networks (94.6%)
  - Bert Language Model (96.6%)

- Test Accuracy
  - > 75%
- Comparison with literatur
  -
- Comparison with DL
  -

Literatur Vergleich Neurales Netzwerk

- Wie wurde ähnliches in Literatur umgesetzt + Welche Accuracy wurde erreicht?
  - https://towardsdatascience.com/how-i-achieved-90-accuracy-on-a-text-classification-problem-with-zero-preprocessing-6acfa96e8d2e
  - torch.NN Optimizer Adama Cross Entropy 90% accuracy by 5 Classees
  - https://medium.com/@claude.feldges/text-classification-with-tf-idf-lstm-bert-a-quantitative-comparison-b8409b556cb3
  - -,- TODO christina -> tabelle

Conclusion (incl. limitations/biases) TODO

bias: computerscience - da training sich an computerscience lehnt