One Ring to rule them all

One System to solve them all

Jan Rygl, rygl@fi.muni.cz

NLP Centre, Faculty of Informatics, Masaryk University

Hot issues

- False product reviews
- Using false information about age or gender in discussion forums and chats
- Machine translation submitted as human expert translation
- Purchase of school essays during educational process
- …

State of the art

- To solve any of mentioned problems, combine:
  - the biggest possible data,
  - style-markers extractors,
  - machine learning,
- and optimize (select good style-markers, feature representation, machine learning parameters, …)

What is wrong

- Research is task dependent:
  - Advancement in one field doesn't not influence other fields

- Research is closed:
  - Companies, even universities are not willing to share their software
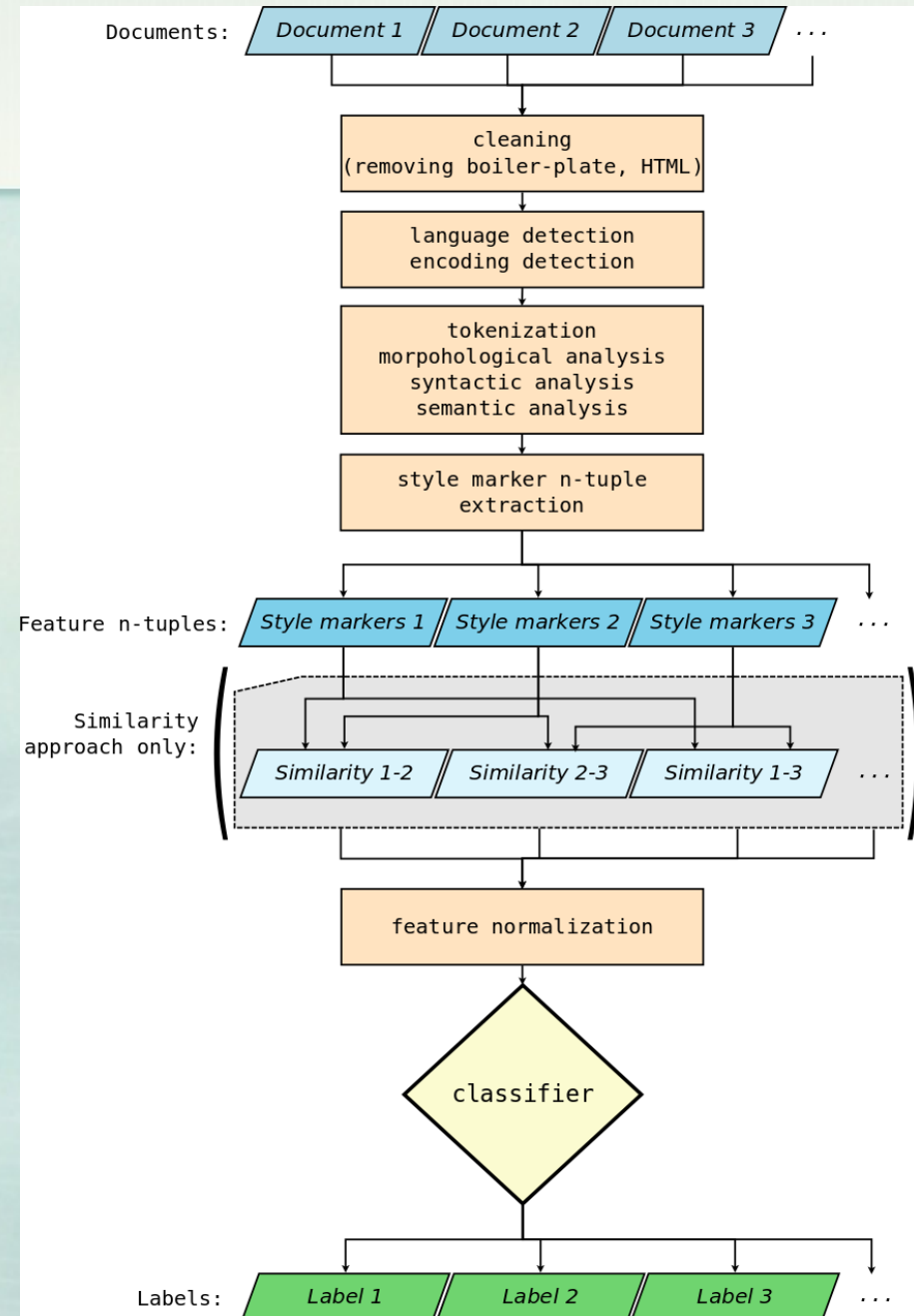
How to solve it

- Make research universal and modular:
  - Advancement in one field affects all fields

- Mysteriousness is always issue
  - Sharing know-how damages efficiency of the system

One system

- Schema:
  - Preprocessing
  - Stylometry
  - Machine learning
    - One model per author vs similarity approach

Progress

- Collect annotated data:
  - automatically from Internet (classic corpora are not useful: authors are mixed and texts are not bounded)
- Create and evaluate style-markers:
  - We have 15 categories containing about 20 000 style-markers
- Improve machine learning techniques:
  - Strong focus on double-layer technique

Thank you for attention