Big Data and Language Technologies

Term Projects

Organization

- Workload:
  - Leipzig: ~6 ECTS equivalent (total with lecture: 10)
  - Weimar: ~3 ECTS equivalent (total with lecture: 6)
- Groups:
  - Self-organized, 2 members (Weimar), 3 members (Leipzig)
  - You can use the #group-finding Discord channel
  - Must send topic preference using one email per group, with chosen topic and all members on CC, to Niklas+Lukas+Janek
- Formal requirements:
  - 8 page project report (including figures, excluding references), double column ACL style (LaTeX template will be available)
  - Code in Git repository with full commit history
  - Supplementary materials, e.g. datasets, models, evaluation results, visualizations, interactive demonstrations, … in a suitable format

L

Timeline

- Register your group/topic
  - until Wednesday, 18.05.2022, 22:00 if you propose an own topic (don’t need a full group yet)
  - we will be in touch with about your idea until the end of the week
  - until Sunday, 22.05.2022, 22:00 if you choose a preferred pre-defined topic
- Project exposé
  - until 20.06.2022, 22:00
  - 1-page description of your topic (i.e. research plan)
  - via email to Niklas+Lukas+Janek
- Project presentation
  - on 04.07.2022 during class, given by 1 group member of your choice
  - 5 minute presentation of your topic to other students
- Project report & supplementary materials
  - until 29.08.2022, 22:00
  - report in PDF format, link to Git repository, supplementary materials in a suitable format
  - via email to Niklas+Lukas+Janek

L

About the Projects

- Ambitious projects that aim to solve real problems from the research field of the Webis group
  - Use of cutting-edge tools; contribution to future technologies
  - Relatively large freedom of project goals and approaches
- Learning high standards of scientific writing
  - Ideal preparation for writing a thesis (methodical and perhaps topic-wise)
  - Close feedback and support from the supervisors
- Optional extracurricular activity: joint publication of papers / datasets

About the Groups

- Interdisciplinary groups are great
- Must agree on the topic, but other parameters must also match
  (e.g. technical skills may compliment each other)
- Team up with the other group members
  to exchange about the single-student course assignments coming up
  - giving each other feedback
  - getting in a working habit
  - getting to know each other better

L

Projects

Primer: Idiom Extraction

- Example: Opinion Mining
  - “IMHO, this is the best solution.”
  - Fixed figures of speech that indicates the purpose of an act of speech
  - Example paper: https://aclanthology.org/W13-4046.pdf
- Idioms give insight in aspects of society
- How can this be transformed into training data/evaluation? What are the tasks and expected derived insights?

Analyzing the Use of Idiomatic/Figurative Language in Web Data

Goals:

- From website data, patterns of idiomatic/figurative language should be extracted
- Such patterns could include: “x is defined as y”, “x is the y of field z”, analogies in general
- The extraction could be done using regular expressions. Additional cleaning will be required
- Analysis of the extracted dataset
- Training/finetuning a DL model on this data with the goal of refining/cleaning the dataset or learning abstractions on the used language
- A pipeline for streaming data from the Internet Archive web crawls into deep learning models is available

Focus: Not on the engineering (pipeline exists), but on defining an extraction process and interpreting the results.

References:

- https://github.com/niklasdeckers/web-archive-keras
- https://github.com/niklasdeckers/web-archive-keras/blob/master/examples/tools/regex_counter.py
- https://aclanthology.org/S12-1047.pdf

N

Statements About the Future

Goals:

- Use the Internet Archive pipeline to extract statements about the future
  (“in the future, we will have flying cars”, “in 10 years, nobody will be using the internet anymore”)
- We have access to ~10 years of Internet Archive data
- How can the extraction be made robust? How hard is it to increase precision?
- How can the concepts be visualized? Clustering?
- Sentiment analysis? Fact checking approaches?

Focus: Performance evaluations, checking bias, designing experiments. Visual analytics, data exploration.

References:

- https://github.com/niklasdeckers/web-archive-keras

N/L

Explicit Sentiment Statements

Goals:

- Search the Web Archive for patterns like “I love…”/”I hate…”
- Can the resulting dataset be used to train sentiment classification?
- How well do such statements conform with existing sentiment classification datasets?

Focus: Dataset cleaning, designing experiments, working with existing datasets and model architectures

References:

- https://github.com/niklasdeckers/web-archive-keras

N

Other Ideas for Language Patterns

- Desires: "I wish", "I would love", "it would be great if"
- Calls for action: "we should", "let's"
  - Combine this with the Lexicon of Verbal Polarity Shifters ("abandon", "avoid", ...)
- Uncertainty: "I don't know"/"nobody knows"/"I wish I knew"
- Definitions/explanations: "is defined as", "is the opposite of"

Website Template Induction for Data Extraction

Goals:

- Classify websites into genre categories such as blog, news, e-commerce, etc.
- Learn common “template” patterns in some of these categories for extracting metadata information.
- Extract information based on the learned templates (e.g. author, year, product names, prices etc.).

Focus: Large-scale training and application of deep learning models to web archive data for classification and extraction.

References:

- https://arxiv.org/abs/2202.00217
- https://storage.googleapis.com/pub-tools-public-publication-data/pdf/59f3bb33216eae711b36f3d8b3ee3cc67058803f.pdf

J

Constrained Language Generation

Goals:

- Explore ways to enforce occurrence of substrings / suffixes in autoregressive text generation
- E.g. modified beam search scoring, fine-tuning, prompt engineering, ...
- Compile a dataset / test battery to evaluate & compare approaches

Focus: Experiment design, testing different approaches. Building demo applications.

References:

- https://arxiv.org/abs/2110.15181

L

Text Reuse Detection using Contrastive Learning

Goals:

- Text Reuse is the reuse of a piece of text in another document through e.g. verbatim copying, rephrasing, summarization, …
- Use contrastive learning on top of pre-trained embeddings to construct a text reuse classifier for doc pairs
- Two levels of granularity possible: for a text pair (d_i, d_j) where text is potentially reused from d_i to d_j, predict a binary label; or predict the exact location of reused text in d_j

Focus: Dataset curation (fusion of existing datasets of small scale is necessary), model development, evaluation. Theoretical foundations on contrastive learning will be helpful.

References:

- https://lilianweng.github.io/posts/2021-05-31-contrastive/
- http://yann.lecun.com/exdb/publis/pdf/chopra-05.pdf
- http://ceur-ws.org/Vol-2723/long22.pdf

L/C

Propose Your Own

Write us an email with:

- What problem should be solved?
- Using which techniques?
- Using/creating what data?
- What are the deliverables?

L

Contact

- Janek: janek.bevendorff@uni-weimar.de
- Lukas: lukas.gienapp@uni-leipzig.de
- Niklas: niklas.deckers@uni-leipzig.de

L

Addendum

- We have open SHK/WHK position(s) at Temir
- Around 10h/w (can be more or less depending on preference)
- Backend development for the picapica.org web service
- Tech Stack: Golang, gRPC, Postgres, RabbitMQ, Kubernetes
- Experience preferable, but not required

If interested, or you know someone who is: email to lukas.gienapp@uni-leipzig.de