ammico: A journey to an open-source software package

Inga Ulusoy
Scientific Software Center, Heidelberg University
The beginning: A consultation

- Delia emailed us about getting the data / extracting information from the data
- The data was based on a list of URLs, many of which were links to a wayback machine archive (images) and not actual html websites
- After an initial consultation, Delia proceeded with writing an initial script that would take screenshots of the websites.
- Delia contacted us again to help extract the information from the screenshots: Crop the screenshots automatically, extract the text from the remaining image, and carry out an image analysis for the image content. Delia had already tried a few libraries but could not get good results.
Support as a small-scale project
Support as a small-scale project

- Extract image on the image
- Extract the text: Only works if language is already known and if clearly legible
Support as a small-scale project

- First prototype was generated
- Quality of the data insufficient
- Automatic cropping of full pages inconsistent
- Text extraction working poorly as language not known / text does not stand out
Taking one step further: Open call proposal

- Short one-page proposal
- In communication with SSC: Feasibility and estimate of person-months
- Review and evaluation together with SSC advisory board

Research Aims and Impact

The ultimate goal of this project is to develop a computer-assisted toolset to investigate the content of disinformation campaigns worldwide. Empirically, we aim to map out the relationship between what is being shown in images and the accompanying text. Using original misinformation data that the International Fact-Checking Network collects from across the globe, our short-to-medium term empirical goals are to extract fact- and emotion-based text content, as well as to extract facial expressions and body posture data from the available posts. This toolset would enable us to conduct top class research on worldwide disinformation and its impact, and make a major contribution to the political science and communication research on the topic.

Intended Software Development

We are currently able to extract the posts from the internet as .png files, to separate the text from the accompanying visual content, and to translate the text in English. These python code files are available on Github. Our short-to-medium term targets are to complete the following five tasks: 1-improving the text extraction through better image transformation 2-improving the preparation of the text for the data analysis (e.g., text cleaning) 3-performing person and face recognition in images, facial expressions recognition, as well as the extraction of any other available individual characteristics (e.g., gender, clothes) 4-extraction of other non-human objects in the image 5-color analysis of the images
Distribution of small-scale and open call projects 2021-2023

Small scale projects:
- Life Sciences: 18.2%
- Natural Sciences: 18.2%
- Computer Science/Mathematics: 9.1%
- Digital Humanities: 9.1%
- Social Sciences: 9.1%
- Earth Sciences: 9.1%

Open call projects:
- Life Sciences: 27.3%
- Natural Sciences: 36.4%
- Computer Science/Mathematics: 9.1%
- Digital Humanities: 9.1%
- Social Sciences: 9.1%
- Earth Sciences: 9.1%
Open call project

Objectives July 2022:
- Text extraction and translation
- Text cleaning
- Person, face and emotion recognition, detection of gender, type of clothing
- Object detection
- Color (hue) detection

End of open call project June 2023:
- Text extraction and translation
- Text cleaning, summary, named entity recognition, sentiment analysis, topic analysis
- Person, face and emotion recognition, detection of gender, ethnicity, age
- Image captioning and visual question answering
- Color (hue) detection
- Multimodal search
- Cropping tool

Communication
Testing
Test data
Test results
Reference results

Initial requirements
Feasibility

Implemented features
Accuracy
ammico: An open call project
ammico: Text extraction and analysis
ammico: Image summary and VQA

Analysis type: summary_and_q
Model type: base
Analysis question: Are there politicians in the image?

Run Detector

<table>
<thead>
<tr>
<th>filename</th>
<th>../..data/Image_some_text/109237S_sp.png</th>
</tr>
</thead>
<tbody>
<tr>
<td>const_image_summary</td>
<td>a group of men sitting next to each other in a room</td>
</tr>
<tr>
<td>3_non-deterministic_summary</td>
<td>a group of men sitting in a row wearing suits, a group of men sitting on top of a blue couch, a group of men sitting in front of each other</td>
</tr>
<tr>
<td>Are there politicians in the image?</td>
<td>yes</td>
</tr>
</tbody>
</table>
ammico: Face and emotion detector

angry  happy  sad  surprised  disgust  fear  neutral

<table>
<thead>
<tr>
<th>filename</th>
<th>../..data/Image_some_text/1092375_spa.png</th>
</tr>
</thead>
<tbody>
<tr>
<td>face</td>
<td>Yes</td>
</tr>
<tr>
<td>multiple_faces</td>
<td>Yes</td>
</tr>
<tr>
<td>no_faces</td>
<td>6</td>
</tr>
<tr>
<td>wears_mask</td>
<td>Yes, Yes, Yes</td>
</tr>
<tr>
<td>age</td>
<td>28, 27, 31</td>
</tr>
<tr>
<td>gender</td>
<td>Man, Man, Man</td>
</tr>
</tbody>
</table>

EmotionDetector

Emotion threshold | Race threshold
50               | 50
ammico: Color detector

+ multimodal search, topic analysis, cropping
ammico: Technological readiness?

- Currently beta-testing (= testing by smaller group of end users)
- Released on PyPi (= install using pip)
- Preparing release on conda (=install using conda)
- Runs on linux, Windows, MacOS
- Runs on google colab and bwVisu (with limitations)
- Can be computationally expensive, especially VQA and multimodal search
- Text extraction via google cloud vision: You need a google key
- TRL 4 – technology validated in lab

Dumitrescu, D., Ulusoy, I., Andriushchenko, P., Daskalakis, G., Kempf, D., & Ma, X. (2023). AMMICO, an AI Media and Misinformation Content Analysis Tool (Version 0.2.0) [Computer software]. https://doi.org/10.31235/osf.io/v8txj
ammico: A timeline

- Email to SSC
- Consultation
- Small-scale project
- Testing of the prototype
- Preparation of open call proposal
- Consultation
- Open call project
- Testing of the software
- Production runs
- Publication
- Applications for third-party funding
- Search for collaborators
- Advertisement and distribution of the software package
Support offered by the SSC: Development & Sustainability

**Target**
Group leaders

**Aim**
- Support researchers with high-quality, sustainable research software that generates reproducible data
- Provide access to well-trained RSEs
- Provide a career option and a community to RSEs
- Adapt the latest software engineering standards in research software development

<table>
<thead>
<tr>
<th>type of support</th>
<th>small-scale project</th>
<th>open call project</th>
<th>third-party funded project</th>
</tr>
</thead>
<tbody>
<tr>
<td>time scale</td>
<td>1 week</td>
<td>1 – 12 months</td>
<td>&gt; 1 month</td>
</tr>
<tr>
<td>cost</td>
<td>free of charge</td>
<td>free of charge</td>
<td>determined through consultation</td>
</tr>
<tr>
<td>availability</td>
<td>throughout the year</td>
<td>proposal submission</td>
<td>deadline mid-June</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>depending on availability of resources</td>
</tr>
</tbody>
</table>
Support offered by the SSC: Teaching & Consultation

Aim
- Provide access to training in software engineering best practices
- Spread the use of best practices in the community
- Provide an easy, low-barrier entrance point for researchers, directly provide feedback on software quality or software-related questions

Target
PhD + Master students + Postdocs for training
All of these + group leaders for consultations

- Block courses
- Compact courses
- On-demand courses
- Guidelines
- Mentoring program
- Consultations
- Seminars
- Template repositories
Support offered by the SSC: Outreach & Communication

**Aim**
- Raise awareness for software engineering best practices
- Improve recognition of research software as scientific output
- Support and shape RSE career paths
- Support and enhance recognition of research software infrastructures
- Develop and keep official guidelines for research software management up-to-date
- Learn from and get involved in community actions
- Participate in national and international software engineering initiatives as a community service
- Contribute to open-source software
- Contribute to open science

**Target**
National initiatives
Funding agencies
Local institutions

- DFG
- HCDH
- Open Science AG
- SURESOFT
- de-RSE
- German Reproducibility Network
- KFD
- Carpentries
- University library
- ...
ammico: Also for your research?

- If you want to use ammico:

  https://github.com/ssciwr/AMMICO


- Or get in touch!
The SSC: Let us help you!

Contact us at

ssc@iwr.uni-heidelberg.de

Take a look at our offers at

https://www.ssc.uni-heidelberg.de