IT Systems Engineering
Universität Potsdam

Captcha Service

Documentation

Potsdam, February 2017

Supervisor

Prof. Dr. Christoph Meinel,
Christian Bartz

Internet-Technologies and Systems Group

Abstract

The increasing numbers of bots, especially crawlers, within the World Wide Web has been a major concern for several years now. Over the years, different approaches to tackle bots were implemented and tested. One of the major solutions for dealing with bots in the recent years were Captchas. Originally, giving users specific tasks to solve, which bots would be unable to solve, was the main idea. Through distortion and other obstacles, Captchas were improved against algorithmic solutions.

The potential of million online users solving Captchas was quickly realized. Difficulties in identifying words or objects in images using computers, could be solved using the combined solutions of Captcha users. We implemented our own Captcha Service in order to allow researchers and scientists to get their own datasets labeled using online users.

Contents

1. Introduction

1.1. Motivation

Researchers and scientists are lacking the time and capacities to label their data, which is often further needed in order to advance other technologies. Using required authentication processes for online users, we are able to utilize huge amounts of free labor. Our main goal was building a straightforward service for researchers and scientists to allow precise data labeling.

Therefore, a simple integration for web services was also wanted.

2. Related Work

In preparation of creating an own Captcha service we searched the Internet for existing ideas and implementations of systems used for data labeling and machine learning purposes. The first popular approach was the Soylent Grid paper which was published in 2007 [1]. Although there were never any popular implementations of the ideas, the paper provided several different approaches for data labeling. They were mostly based around the idea of object recognition in images, e.g. by clicking on objects or drawing rectangles around it. Other proposals were directed to object recognition, where users had to name objects displayed in certain images.
Another paper which was published just a year later, dealt with text recognition and described the concept which was implemented in reCAPTCHA v1 [2]. The system was created by Google in order to solve problems in the "Google Book Project). Two different optical character recognition (OCR) algorithms were used to translate images of scanned pages into digital texts. The solving of Captchas was used to identify words which could not be deciphered clearly by the OCR algorithms. Because of its detailed documentation and its successful usage this method was ideal to be implemented and therefore the first Captcha type which is supported by our system.

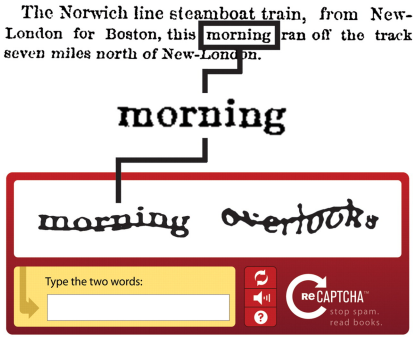

Figure 1: The reCAPTCHA approach explained in one picture

[1] http://vision.ucsd.edu/sites/default/files/icv2007.pdf
[2] http://science.sciencemag.org/content/321/5895/1465.full

3. Architecture

The architecture is mainly represented in the models.py data. It is designed for simple expandability and uses inheritance to simplify the introduction of new captcha types. An overview is given in the class diageramm in figure 2.

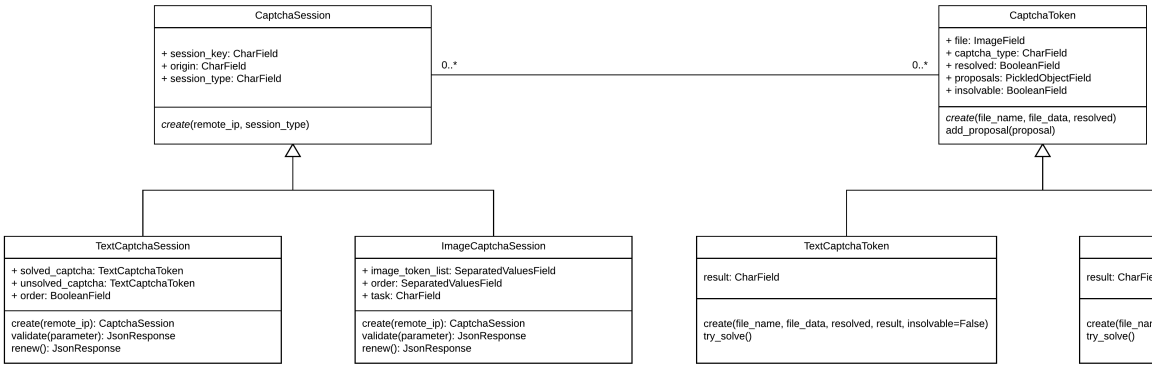

Figure 2: Class diagramm representing the classes used for the generation of captchas. TODO text, update

It consists of two main classes, the *CaptchaToken* and *CaptchaSession*. The class *CaptchaToken* represents a single image, that is part for a captcha, e.g. a single

word, that needs to be written down by the user in order to solve the captcha. The class *CaptchaSession* represents a complete captcha challenge a user has to solve, e.g. writting down the words shown on all images. Each type of captcha challenge provided by the service is represented by a subclass of *CaptchaSession* and *CaptchaToken*. Currently two kinds of Captchas, ImageCaptchas and TextCaptchas are supported.

All that needs to be done for implementing a new type of captcha challenge is to create a new subclass for *CaptchaToken* and *CaptchaSession* and implement specific functionality in these subclasses. Which methods and attributes need to be added in the new subclasses is listed in the "Attributes and Methods implemented in the subclass"-paragraph.

All instances of a *CaptchaToken* or *CaptchaSession* are saved in the `db.sqlite3`-Database.

3.1. CaptchaToken

TODO konsistenz kuze beschr u einleitung The class *CaptchaToken* represents a single image, that is part for a captcha, e.g. a single word, that needs to be written down by the user in order to solve the captcha.

Attributes and Methods implemented in the superclass

Attributes:

- `file`: Image, that is represented by the CaptchaToken.
- `captcha_type`: String, that defines the type of captcha the token can be used for. Currently "text" for Textcaptchas and "image" for Imagecaptchas are supported.
- `resolved`: Boolean, that indicates, if the solution for a *CaptchaToken* is known or not. A `0` means the token is unsolved and a `1` means the Token is solved.

- `proposals`: Dicitionary, that stores the possible solutions suggested by users of the captcha service and how often each solution was suggested.
- `insolvable`: Boolean that indicates, that a token is not solvable by clients of the captcha service.

Methods:

- `create( file_name, file_data, resolved)`: Responsible for basic configuration, that need to be done for all kinds of tokens, when they are created. Only used for supercalls in the `create()`-method of subclasses.
- `add_proposals(proposal)`: Adds a new suggested solution to the `proposals`-dictionary, or increments the counter for an allready suggested proposal.

Attributes and Methods implemented in the subclass

Attributes:

- `result`: Saves the correct solution for a token. Datatype differs between different subclasses, e.g. *TextCaptchaToken* saves a string and *ImageCaptchaToken* saves a boolean.

Methods:

- `create(file_name, file_data, resolved, result, insolvable=False)`: Resposible for configurating all attributes of the *CaptchaToken*. Returns a *CaptchaToken*.
- `try_solve`: Responsible for finding the correct solution for a *CaptchaToken* based on the values saved in the `proposals`-attribute.

3.2. CaptchaSession

TODO konsistenz kuze beschr u einleitung Represents an instance of a captcha challange, that needs to be solved by a certain client. A *CaptchaSessio*n consists

of multiple ImageTokens, that are chosen randomly in order to create different challanges dynamically. Each Sessions corresponds to one of the supported types of *CaptchaTokens*.

Attributes and Methods implemented in the superclass

Attributes:

- `session_key`: String, that serves as primary key to identify each session.
- `origin`: String, that holds the IP adress that requested the captcha challenge. It is used match requests made by the client to the corresponding session.
- `session_type`: String, that defines the kind of captcha challenge, the client has to solve. Currently "text" for Textcaptchas and "image" for Imagecaptchas are supported.

Methods:

- `create(remote_ip, session_type)`: Responsible for basic creation of a *CaptchaSession* of the requested type for the given IP adress. Only used for supercalls in the `create()`-method of subclasses.

Attributes and Methods implemented in the subclass

Attributes:

Each session needs to store the tokens, that were used for creating the session and additional information, that is needed for validating the answer given by the client. This can differ for every captchatype.

TextCaptchaSession:

- `solved_captcha_token`: *TextCaptchaToken*, that is allready solved and is used as a control word for the session.

- `unsolved_captcha_token`: *TextCaptchaToken*, that is not solved and shall be solved by the client.
- `order`: Boolean indicating the order, in which the two tokens are displayed to the client.(0 -> solved unsolved 1 -> unsolved solved) It is needed to map the answers given by the client to the right tokens.

ImageCaptchaSession:

- `image_token_list`: List of *ImageCaptchaTokens*, where all tokens used for the session are saved.
- `order`: List of Booleans, that indicates which token in the `image_token_list` is solved. (0 -> unsolved, 1-> solved)
- `task`: String, taht saves the task for the *ImageCaptchaSession*, e.g. which objects should be detected in the images.

Methods:

- `create(remote_ip)`: Responsible for creating a *CaptchaSession* and returning the created session to the corresponding `view`.
- `validate(parameters)`: Responsible for validating the solution for a CaptchaSession and returning the created session to the corresponding `view`. The solution suggested by the client is included in the parameters. Returns whether the session is valid or not.
- `renew()`: Responsible for exchanging the *CaptchaTokens* of a *CaptchaSession*, to create a new challenge or the same session.

4. Image Distortion

The fact that images for text Captchas are provided by users makes it impossible to tell if those images are easy to recognize for bots and are therefore safe to be used as Captcha token. In order to complicate the recognition of the Captcha token the systems uses an image distortion algorithm which is automatically applied to all uploaded text Captcha tokens.
The image distortion algorithm consists of two steps: the drawing of a horizontal line and a wave transformation.
In the first step it places a horizontal line in the middle of the image which is colored with the dominant color of the whole picture. Afterwards this line will be transformed together with the rest of the image.
The frequency as well as the amplitude of the wave which will be applied to text are dependent to the height of the image. Furthermore the frequency depends on the width of the image so that one wavelength is at least as wide as the image itself. In addition to this both, the frequency and the amplitude, are modified by a random value in order make every transformation unique.
Everything that is shifted out of bounds will be cut off. Additionally the pixels which were located at the bottom and the top of the the original picture will be stretched out vertically to fill the space which was emptied due to the transformation.

5. Implementation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In erat mauris, faucibus quis pharetra sit amet, pretium ac libero. Etiam vehicula eleifend bibendum. Morbi gravida metus ut sapien condimentum sodales mollis augue sodales. Vestibulum quis quam at sem placerat aliquet. Curabitur a felis at sapien ullamcorper fermentum. Mauris molestie arcu et lectus iaculis sit amet eleifend eros posuere. Fusce nec porta orci.

Integer vitae neque odio, a sollicitudin lorem. Aenean orci mauris, tristique

luctus fermentum eu, feugiat vel massa. Fusce sem sem, egestas nec vulputate vel, pretium sit amet mi. Fusce ut nisl id risus facilisis euismod. Curabitur et elementum purus. Duis tincidunt fringilla eleifend. Morbi id lorem eu ante adipiscing feugiat. Sed congue erat in enim eleifend dignissim at in nisl. Donec tortor mauris, mollis vel pretium vitae, lacinia nec sapien. Donec erat neque, ullamcorper tincidunt iaculis sit amet, pharetra bibendum ipsum. Nunc mattis risus ac ante consequat nec pulvinar neque molestie. Etiam interdum nunc at metus lacinia non varius erat dignissim. Integer elementum, felis id facilisis vulputate, ipsum tellus venenatis dui, at blandit nibh massa in dolor. Cras a ultricies sapien. Vivamus adipiscing feugiat pharetra.Lorem ipsum dolor sit amet, consectetur adipiscing elit. In erat mauris, faucibus quis pharetra sit amet, pretium ac libero. Etiam vehicula eleifend bibendum. Morbi gravida metus ut sapien condimentum sodales mollis augue sodales. Vestibulum quis quam at sem placerat aliquet. Curabitur a felis at sapien ullamcorper fermentum. Mauris molestie arcu et lectus iaculis sit amet eleifend eros posuere. Fusce nec porta orci.

Integer vitae neque odio, a sollicitudin lorem. Aenean orci mauris, tristique luctus fermentum eu, feugiat vel massa. Fusce sem sem, egestas nec vulputate vel, pretium sit amet mi. Fusce ut nisl id risus facilisis euismod. Curabitur et elementum purus. Duis tincidunt fringilla eleifend. Morbi id lorem eu ante adipiscing feugiat. Sed congue erat in enim eleifend dignissim at in nisl. Donec tortor mauris, mollis vel pretium vitae, lacinia nec sapien. Donec erat neque, ullamcorper tincidunt iaculis sit amet, pharetra bibendum ipsum. Nunc mattis risus ac ante consequat nec pulvinar neque molestie. Etiam interdum nunc at metus lacinia non varius erat dignissim. Integer elementum, felis id facilisis vulputate, ipsum tellus venenatis dui, at blandit nibh massa in dolor. Cras a ultricies sapien. Vivamus adipiscing feugiat pharetra.Lorem ipsum dolor sit amet, consectetur adipiscing elit. In erat mauris, faucibus quis pharetra sit amet, pretium ac libero. Etiam vehicula eleifend bibendum. Morbi gravida metus ut sapien condimentum sodales mollis augue sodales. Vestibulum quis quam at sem placerat aliquet. Curabitur a felis at sapien ullamcorper fermentum. Mauris molestie arcu et lectus iaculis sit amet eleifend eros posuere. Fusce nec porta orci.

Integer vitae neque odio, a sollicitudin lorem. Aenean orci mauris, tristique luctus fermentum eu, feugiat vel massa. Fusce sem sem, egestas nec vulputate vel, pretium sit amet mi. Fusce ut nisl id risus facilisis euismod. Curabitur et elementum purus. Duis tincidunt fringilla eleifend. Morbi id lorem eu ante adipiscing feugiat. Sed congue erat in enim eleifend dignissim at in nisl. Donec tortor mauris, mollis vel pretium vitae, lacinia nec sapien. Donec erat neque, ullamcorper tincidunt iaculis sit amet, pharetra bibendum ipsum. Nunc mattis risus ac ante consequat nec pulvinar neque molestie. Etiam interdum nunc at metus lacinia non varius erat dignissim. Integer elementum, felis id facilisis vulputate, ipsum tellus venenatis dui, at blandit nibh massa in dolor. Cras a ultricies sapien. Vivamus adipiscing feugiat pharetra.Lorem ipsum dolor sit amet, consectetur adipiscing elit. In erat mauris, faucibus quis pharetra sit amet, pretium ac libero. Etiam vehicula eleifend bibendum. Morbi gravida metus ut sapien condimentum sodales mollis augue sodales. Vestibulum quis quam at sem placerat aliquet. Curabitur a felis at sapien ullamcorper fermentum. Mauris molestie arcu et lectus iaculis sit amet eleifend eros posuere. Fusce nec porta orci.

Integer vitae neque odio, a sollicitudin lorem. Aenean orci mauris, tristique luctus fermentum eu, feugiat vel massa. Fusce sem sem, egestas nec vulputate vel, pretium sit amet mi. Fusce ut nisl id risus facilisis euismod. Curabitur et elementum purus. Duis tincidunt fringilla eleifend. Morbi id lorem eu ante adipiscing feugiat. Sed congue erat in enim eleifend dignissim at in nisl. Donec tortor mauris, mollis vel pretium vitae, lacinia nec sapien. Donec erat neque, ullamcorper tincidunt iaculis sit amet, pharetra bibendum ipsum. Nunc mattis risus ac ante consequat nec pulvinar neque molestie. Etiam interdum nunc at metus lacinia non varius erat dignissim. Integer elementum, felis id facilisis vulputate, ipsum tellus venenatis dui, at blandit nibh massa in dolor. Cras a ultricies sapien. Vivamus adipiscing feugiat pharetra.Lorem ipsum dolor sit amet, consectetur adipiscing elit. In erat mauris, faucibus quis pharetra sit amet, pretium ac libero. Etiam vehicula eleifend bibendum. Morbi gravida metus ut sapien condimentum sodales mollis augue sodales. Vestibulum quis quam at sem placerat aliquet. Curabitur a felis at sapien ullamcorper fermentum. Mauris molestie arcu et lectus iaculis sit amet eleifend eros posuere. Fusce nec porta

orci.

Integer vitae neque odio, a sollicitudin lorem. Aenean orci mauris, tristique luctus fermentum eu, feugiat vel massa. Fusce sem sem, egestas nec vulputate vel, pretium sit amet mi. Fusce ut nisl id risus facilisis euismod. Curabitur et elementum purus. Duis tincidunt fringilla eleifend. Morbi id lorem eu ante adipiscing feugiat. Sed congue erat in enim eleifend dignissim at in nisl. Donec tortor mauris, mollis vel pretium vitae, lacinia nec sapien. Donec erat neque, ullamcorper tincidunt iaculis sit amet, pharetra bibendum ipsum. Nunc mattis risus ac ante consequat nec pulvinar neque molestie. Etiam interdum nunc at metus lacinia non varius erat dignissim. Integer elementum, felis id facilisis vulputate, ipsum tellus venenatis dui, at blandit nibh massa in dolor. Cras a ultricies sapien. Vivamus adipiscing feugiat pharetra.

6. Solving Algorithm

In order to provide a value for the researchers which add data to the Captcha service, the system has to label the uploaded images. This becomes possible due to the solving algorithm, which determines the label based on the given user inputs.
In case of text Captchas the algorithm needs at least three users which solved the Captcha correctly. If three or more suggestions match, the image is marked as solved and labeled accordingly. However the token is identified as unsolvable if there are six or more proposals but no more than two of them match. This approach relatively similar to the concept reCAPTCHA uses. In a paper [3] that was published it was stated, that in most cases three human resolutions are enough to label the image reliably.
The method for labeling image Captchas is similar to the one used for texts. The main difference is the fact that the proposals for these are limited to *true* and *false*, are they suiting the specified task or not. Therefore the algorithm checks if at least four resolutions match and also declares a token as unsolvable if more the six suggestion are given but failed to produces four that match. It was de-

[3] http://science.sciencemag.org/content/321/5895/1465.full

cided to raise the bar for labeling a picture from three to four, because it is more likely to falsely select an image due to a wrong click.

7. Evaluation

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In erat mauris, faucibus quis pharetra sit amet, pretium ac libero. Etiam vehicula eleifend bibendum. Morbi gravida metus ut sapien condimentum sodales mollis augue sodales. Vestibulum quis quam at sem placerat aliquet. Curabitur a felis at sapien ullamcorper fermentum. Mauris molestie arcu et lectus iaculis sit amet eleifend eros posuere. Fusce nec porta orci.

Integer vitae neque odio, a sollicitudin lorem. Aenean orci mauris, tristique luctus fermentum eu, feugiat vel massa. Fusce sem sem, egestas nec vulputate vel, pretium sit amet mi. Fusce ut nisl id risus facilisis euismod. Curabitur et elementum purus. Duis tincidunt fringilla eleifend. Morbi id lorem eu ante adipiscing feugiat. Sed congue erat in enim eleifend dignissim at in nisl. Donec tortor mauris, mollis vel pretium vitae, lacinia nec sapien. Donec erat neque, ullamcorper tincidunt iaculis sit amet, pharetra bibendum ipsum. Nunc mattis risus ac ante consequat nec pulvinar neque molestie. Etiam interdum nunc at metus lacinia non varius erat dignissim. Integer elementum, felis id facilisis vulputate, ipsum tellus venenatis dui, at blandit nibh massa in dolor. Cras a ultricies sapien. Vivamus adipiscing feugiat pharetra.Lorem ipsum dolor sit amet, consectetur adipiscing elit. In erat mauris, faucibus quis pharetra sit amet, pretium ac libero. Etiam vehicula eleifend bibendum. Morbi gravida metus ut sapien condimentum sodales mollis augue sodales. Vestibulum quis quam at sem placerat aliquet. Curabitur a felis at sapien ullamcorper fermentum. Mauris molestie arcu et lectus iaculis sit amet eleifend eros posuere. Fusce nec porta orci.

Integer vitae neque odio, a sollicitudin lorem. Aenean orci mauris, tristique luctus fermentum eu, feugiat vel massa. Fusce sem sem, egestas nec vulputate vel, pretium sit amet mi. Fusce ut nisl id risus facilisis euismod. Curabitur et elementum purus. Duis tincidunt fringilla eleifend. Morbi id lorem eu ante

adipiscing feugiat. Sed congue erat in enim eleifend dignissim at in nisl. Donec tortor mauris, mollis vel pretium vitae, lacinia nec sapien. Donec erat neque, ullamcorper tincidunt iaculis sit amet, pharetra bibendum ipsum. Nunc mattis risus ac ante consequat nec pulvinar neque molestie. Etiam interdum nunc at metus lacinia non varius erat dignissim. Integer elementum, felis id facilisis vulputate, ipsum tellus venenatis dui, at blandit nibh massa in dolor. Cras a ultricies sapien. Vivamus adipiscing feugiat pharetra.Lorem ipsum dolor sit amet, consectetur adipiscing elit. In erat mauris, faucibus quis pharetra sit amet, pretium ac libero. Etiam vehicula eleifend bibendum. Morbi gravida metus ut sapien condimentum sodales mollis augue sodales. Vestibulum quis quam at sem placerat aliquet. Curabitur a felis at sapien ullamcorper fermentum. Mauris molestie arcu et lectus iaculis sit amet eleifend eros posuere. Fusce nec porta orci.

Integer vitae neque odio, a sollicitudin lorem. Aenean orci mauris, tristique luctus fermentum eu, feugiat vel massa. Fusce sem sem, egestas nec vulputate vel, pretium sit amet mi. Fusce ut nisl id risus facilisis euismod. Curabitur et elementum purus. Duis tincidunt fringilla eleifend. Morbi id lorem eu ante adipiscing feugiat. Sed congue erat in enim eleifend dignissim at in nisl. Donec tortor mauris, mollis vel pretium vitae, lacinia nec sapien. Donec erat neque, ullamcorper tincidunt iaculis sit amet, pharetra bibendum ipsum. Nunc mattis risus ac ante consequat nec pulvinar neque molestie. Etiam interdum nunc at metus lacinia non varius erat dignissim. Integer elementum, felis id facilisis vulputate, ipsum tellus venenatis dui, at blandit nibh massa in dolor. Cras a ultricies sapien. Vivamus adipiscing feugiat pharetra.Lorem ipsum dolor sit amet, consectetur adipiscing elit. In erat mauris, faucibus quis pharetra sit amet, pretium ac libero. Etiam vehicula eleifend bibendum. Morbi gravida metus ut sapien condimentum sodales mollis augue sodales. Vestibulum quis quam at sem placerat aliquet. Curabitur a felis at sapien ullamcorper fermentum. Mauris molestie arcu et lectus iaculis sit amet eleifend eros posuere. Fusce nec porta orci.

Integer vitae neque odio, a sollicitudin lorem. Aenean orci mauris, tristique luctus fermentum eu, feugiat vel massa. Fusce sem sem, egestas nec vulputate vel, pretium sit amet mi. Fusce ut nisl id risus facilisis euismod. Curabitur et

elementum purus. Duis tincidunt fringilla eleifend. Morbi id lorem eu ante adipiscing feugiat. Sed congue erat in enim eleifend dignissim at in nisl. Donec tortor mauris, mollis vel pretium vitae, lacinia nec sapien. Donec erat neque, ullamcorper tincidunt iaculis sit amet, pharetra bibendum ipsum. Nunc mattis risus ac ante consequat nec pulvinar neque molestie. Etiam interdum nunc at metus lacinia non varius erat dignissim. Integer elementum, felis id facilisis vulputate, ipsum tellus venenatis dui, at blandit nibh massa in dolor. Cras a ultricies sapien. Vivamus adipiscing feugiat pharetra.

8. Future Work

With the thought in mind of building an easy expandable service, the logic consequence would be on focusing on different Captcha types. In the process, key factors such as access for disabled users can be tackled, e.g. by implementing audio Captchas. Another aspect would be expanding the web interface. The option of downloading solved and unsolved Captchas can be specialized by selecting specific upload times or certain time spans. A feedback of the labeling progress within a task would also be another great tool.

A. Appendix

Appendix