# Software-Agenten im Internet

Florian Pircher
Technologische Fachoberschule
Oberschulzentrum Fallmerayer
Brixen, Italien

22. Dezember 2015

## Abstract

Menschen sind schon lange nicht mehr die einzigen Nutzer des Internets. Inzwischen werden über 50 % aller Webseiten-Aufrufe von autonomer Software getätigt. Diese auch als Bots bezeichnete Softwares agieren in den Schatten des Netzes. Unbemerkt indexieren sie Webseiten, verbreiten Spam, legen gefälscht Profile an oder versuchen in Datenbanken einzubrechen.
Hindernisse wie CAPTCHAs oder Honeypots galten bislang als effektive Gegenmittel, allerdings verhilft der rapide Fortschritt im Feld der Künstlichen Intelligenz modernen Bots auch derartige Barrieren zu durchbrechen. Im Bereich E-Mail-Spam findet ein unablässiger Kampf zwischen Spam-Bots und Klassifizierungsalgorithmen statt.
Im ersten Teil dieser Arbeit wird das Verhalten und Vermögen von Software-Agenten untersucht. Schwerpunkte bilden dabei die Themen Spam und Sicherheit. Der zweite Teil beschreibt die Anwendung des gewonnenen Wissens in Form der Entwicklung eines eigenen Bots der autonom durch das World Wide Web navigiert und anhand von maschinellem Lernen versucht CAPTCHAs zu knacken.

# Inhaltsverzeichnis

# 1 Einleitung

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

# 2 Beobachten und Verstehen

## 2.1 Spam

### 2.1.1 Definition

Unerwünschte Nachrichten im Internet, größtenteils Werbeangebote, Phishing-Attacken oder Übermittler von Schadsoftware, werden als Spam bezeichnet. Spam-Nachrichten stellen per definitionem eine leidige Kommunikationsform dar. Die alltäglichste Art, namhaft durch ihre Omnipräsenz im Leben mit dem Internet, stellt der E-Mail-Spam dar.

Der Begriff Spam entspringt dem gleichnamigen Dosenfleisch, welches im Zweiten Weltkrieg in großen Mengen an Soldaten verteilt wurde. Der Markenname SPAM (kurz für *Spiced Ham*) setzte sich auf diese Weise rasch im Vereinigten Königreich als Deonym für Frühstücksfleisch durch. Die britische Comedy-Gruppe Monty Python griff im Sketch *Spam*[1] der BBC Serie *Monty Python's Flying Circus* die Allgegenwärtigkeit des Fleischprodukts auf, weshalb dieses bis heute als Symbol für einen unerwünschten Überschuss fungiert.

### 2.1.2 Spam-Bekämpfung

Die Aufnahme von computergeneriertem Spam kann nicht restlos durch eine Blockade der Distribution, beispielsweise durch CAPTCHAs oder Sicherheitsfragen, abgewehrt werden. So können beispielsweise E-Mails versendet werden, ohne dass dies unter der Obrigkeit einer vom Empfänger vertrauten Instanz geschieht. Dieser Typ Spam muss auf Seiten des Adressaten identifiziert und beseitigt werden. Für jene Aufgabe eignen sich im Besonderen Klassifizierungsalgorithmen, welche qua Positiv- und Negativ-Beispielen erlernen Spam-Nachrichten ausfindig zu machen.

Klassifizierungsalgorithmen teilen genannte Beispiele in einen mehrdimensionalen Raum ein. Einzelne Aspekte, die sich von Nachricht zu Nachricht unterscheiden, beschreiben die verschiedenen Dimensionen des Raums. Neue Nachrichten werden ebenfalls in diesem Raum erfasst und mittels Clusteranalyse zugeordnet. Beschriebene Methoden sind den Feldern des Maschinellen Lernens und des Data-Mining angehörig.

[1] `https://www.youtube.com/watch?v=M_eYSuPKP3Y`