

# Webová aplikace pro online web scraping

## Bakalářská práce

Jakub Drahoš  
Vedoucí práce: Mgr. Martin Podloucký

Fakulta informačních technologií  
České vysoké učení technické v Praze

18. 6. 2019

## 1 Úvod

## 2 Stávající řešení

## 3 Výsledná aplikace

## 4 Shrnutí

# Web scraping

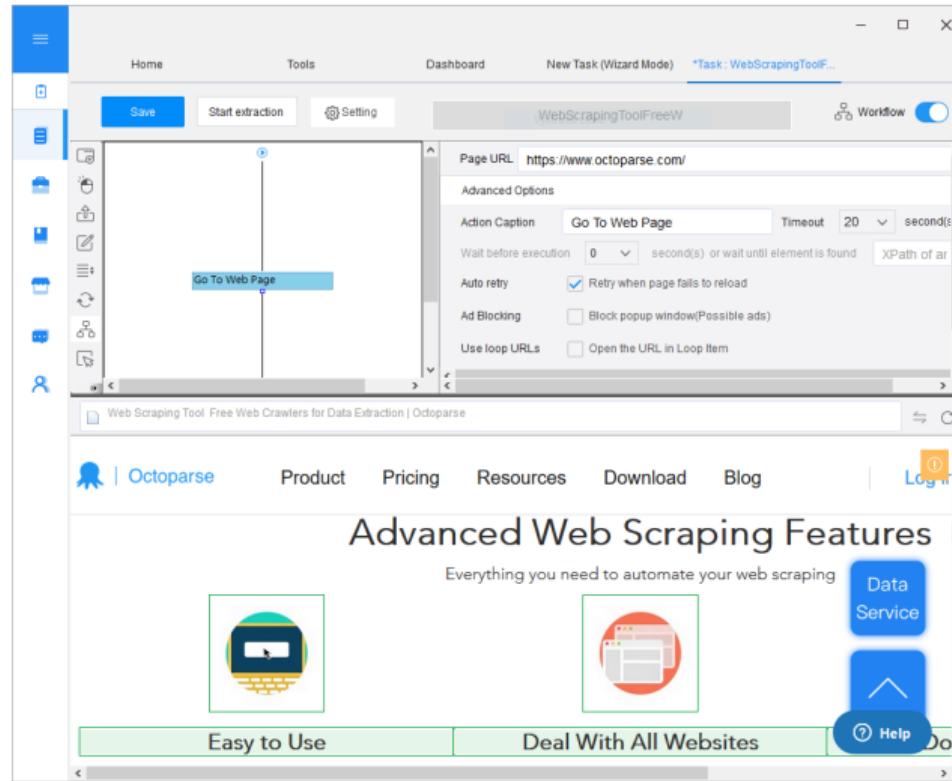
- = technika získávání dat z internetových stránek
- Většinou automatizované
- Např. marketingové společnosti nebo sledování produktů

# Cíle práce

- 1 Tvorba aplikace umožňující provádět web scraping**
- 2 Právní rešerše**
- 3 Analýza stávajících řešení**

# Nevýhody konkurence

- Složité a neintuitivní ovládání
- Nepřehledné grafické rozhraní
- Mnoho funkcionality na úkor uživatelského zážitku



**České vysoké učení technické v Praze**  **Fakulta informačních technologií**

[Fakulta](#) [Aktuálně](#) [Zájemci o studium](#) [Student](#) [Zaměstnanec](#) [Média](#)

[Hledání](#) [Vyhledat](#)

[Struktura](#) [Úřední deska](#) [Zahraniční styky](#) [Věda](#) [Historie](#) [Akademický senát FIT](#) [Partneři](#) [Kontakty](#)

[Novinky](#) [News](#)

[ČVUT](#)  [TOPUNIVERSITIES](#) [Časopis Bud FIT](#) [Anketa ČVUT](#)

[Facebook](#) [Instagram](#) [YouTube](#)

## Letní IT kurz pro studenty středních škol

Ve spolupráci se Stanford University pořádáme 8.–19. července intenzivní 14-denní kurz programování. Kurz určený sítědolákařům s názvem **Introduction to Computer Science** je jedinou takovou akcí v Evropě. Cílem je vzbudit u studentů zájem o programování a naučit je základy závaznou formou. Výuka bude probíhat na naší fakultě a povídají ji profesori a studentiště lektori z obou institucí. Kurz bude v anglickém jazyce a přihlášky je možné podávat do 15. května. Učastníci kurzu platí pouze zápisné 1 000 Kč.

Celý článek o kurzu ICS 2019  
Tisková zpráva o kurzu ICS 2019

## Robot Pepper předával diplomy

Humanoidní robot Pepper se stal 22. března součástí slavnostního ceremoniálu v Betlémské kapli, kde se aktivně zapojil do promoci bakalářských a magisterských absolventů. Robotu pro tuto událost naprogramovali první studenti FIT v moderně vybavené Laboratoři inteligentních vestavěných systémů. Fakulta je jediná v České republice, která robota Pepper využívá také právě pro výuku.

Tisková zpráva o předávání diplomů robotem Pepper

## LAW FIT 2019: Digitální realita

Konference o právu a IT LAW FIT se uskuteční 20. května.

**Scrapper**

[Rows](#) [Columns](#)

[?](#)

**Basics**

**Text search**

**CSS selector**

[Preview](#) [Download](#) 

# Implementace

- Implementováno jako rozšíření do prohlížeče Google Chrome
- Napsané v jazyce JavaScript
- Určené především ke statickému vytěžování stránek

The screenshot shows a web browser displaying the official website of the Faculty of Information Technology (FIT) at the Czech Technical University (ČVUT). The URL in the address bar is [fit.cvut.cz](http://fit.cvut.cz). The page title is "Telefony".

The main navigation menu includes links such as "Fakulta", "Aktuálně", "Zájemci o studium", "Student", "Zaměstnanec", "Média", "Struktura", "Úřední deska", "Zahraniční styky", "Věda", "Historie", "Akademický senát FIT", "Partneři", and "Kontakty".

The left sidebar contains a list of links under the heading "Zaměstnanci":

- > Fakturační údaje
- > Zaměstnanci
- > Úřední hodiny podatelny
- > Děkanát
- > Katedry
- > Vedení FIT
- > Webmaster
- > Helpdesk
- > Vyhledávání

The main content area displays a grid of letters (B, C, Č, D, E, F, G, H, CH, J, K, L, M, N, O, P, R, Ř, S, Š, T, V, Z, Ž) each with a corresponding contact entry below it.

For letter "B":

- Ing. Zdeněk Balák zdenek.balak@fit.cvut.cz +420-22435-7988  
T9:341b
- Bc. Maksym Balatsko balatmak@fit.cvut.cz  
TH:A-1256, TH-A-1258

For letter "C":

- oddělení pro spolupráci s průmyslem - programátor balatmak@fit.cvut.cz

For letter "D":

- katedra aplikované matematiky - výzkumný a vývojový pracovní

# Výhody oproti konkurenci

- Jednoduché na používání
- Přehledné grafické rozhraní
- Extrakce během pár okamžiků

# Shrnutí

- Právní rozbor problematiky
- Nedostatky konkurenčních nástrojů
- Vytvoření aplikace se zaměřením na uživatelský zážitek

# Dotazy

Otázka č. 1: „*Plánujete na projektu dále pracovat a otevřeně rozšíření distribuovat? Pokud ano, pod jakou licencí a proč?*“

# Dotazy

Otázka č. 2: „*Bylo by možné (a případně jak) jednoduše rozšířit Váš Scraper o crawling například výčtem adres či označením typu odkazů k následování tak, aby bylo zachováno intuitivní rozhraní (viz sekce 4.4)? Jak byste případně postupoval?*“