# Adverse media screening

Tím 8

**Členovia:** Jakub Hlavačka, Dominik Horváth, Jakub Müller, Táňa Poláková, Dávid Silady, Adam Šípka
**Vedúci:** Richard Marko

**Kontakt na tím:**
tim8.fiit.stuba@gmail.com

## Tím

Náš tím tvoria perspektívni ľudia, ktorých záujmy pokrývajú rôzne oblasti informatiky. Väčšina členov nášho tímu aktuálne pôsobí v zamestnaní, ktoré im dáva cennú informatickú prax. Dávid Silady a Dominik Horváth sú kolegovia vo firme IBM, kde pracujú na pozíciach Package specialist a DevOps inžinier. Automatizácia procesov, Docker, Kubernetes, Bash alebo Python sú častou súčasťou ich práce. Táňa Poláková má ako hlavnú náplň práce zber a predspracovanie veľkého množstva dát (využíva MySQL, numpy), vďaka ktorým bude môcť navrhnúť model neurónovej siete, ktorý nahradí aktuálne riešenie. Jakub Hlavačka v práci často využíva technológie Scrapy a Selenium na scrapovanie údajov z internetu.

Naše bakalárske práce sa zaoberali dátovou analýzou, umelou inteligenciou, technológiou blockchain a webom.

Veľkou hodnotou (tak, ako aj ostatní členovia) v našom tíme je Dominik Horváth, ktorého bakalársky projekt bol zameraný na dátovú analýzu falošných správ súvisiacich s pandémiou COVID-19, a jeho práca bola jednou z prvých, ktoré sa venovali tejto oblasti. Daná tematika sa priamo dotýka témy na tímový projekt, o ktorý máme najväčší záujem.

Táňa Poláková a Dávid Silady sa na svojich bakalárkach zaoberali neurónovými sieťami (pracovali s knižnicami TensorFlow a PyTorch). Dávid odhadoval hĺbku z fotografií pomocou neurónových sietí, kde sa zoznámil najmä s generovaním a spracovaním dát, rovnako ako so základmi konvolučných sietí.  Táňa trénovala neurónovú sieť, ktorej výstup nezávisel od usporiadania prvkov vo vstupe, pričom vstup boli medicínske dáta.

Práca Jakuba Hlavačku bola zameraná na využitie technológie blockchain v IoT zariadeniach, kde nadobudol prvé väčšie znalosti so systémami ako Docker a Hyperledger Fabric.

Jakub Müller má zo svojej bakalárskej práce skúsenosti z oblasti tvorby webových stránok a UX testovania, keďže jeho práca bola zameraná na používateľmi vnímanú dôveryhodnosť v online spravodajstve.

Adam Šípka v bakalárke vytvoril webový crawler (použil framework Selenium), ktorý zaznamenával a ukladal metadáta o YouTube videách. Vďaka tomu získal skúsenosti so získavaním dát a následne aj s ich spracovaním a analýzou.

## Motivácia

Projekt Adverse Media Screening (v skratke AMS) sa zaoberá overovaním osôb prostredníctvom článkov zverejnených internetovými médiami. Jedná sa o riešenie, ktoré umožňuje previerku kriminálneho pozadia fyzickej alebo právnickej osoby. Vstupom do výsledného projektu je meno osoby a na výstupe sa nachádzajú internetové články, ktoré o tejto osobe hovoria. Tento projekt môže byť prínosom pre spoločnosti, kde integrita ich potenciálnych zamestnancov je kľúčová pre spoľahlivý a bezpečný chod. Do tejto kategórie spadajú banky, poisťovne, vládne organizácie a podobne.

## Realizácia

Na projekte je potrebné riešiť hneď niekoľko zaujímavých a rôznorodých informatických úloh.

Riešenie bude spočívať z webovej aplikácie, nezávislého API servera a webového scrapera. Webová časť riešenia bude pozostávať z klientskej aplikácie v ReactJS, webového servera v NodeJS/Express a PostgreSQL databázy. API server bude realizovaný pomocou frameworku Flask a bude slúžiť ako poskytovateľ zozbieraných a indexovaných dát. Ako indexer bude použitý Elasticsearch, zatiaľ čo Scrapy využijeme ako scraper.

Webová časť riešenia bude slúžiť ako jednoducho dostupné užívateľské rozhranie. V ňom plánujeme implementovať niekoľko základných i rozšírených funkcionalít, ako napr. prístup k archivovaným stránkam. API server bude zároveň slúžiť pre verejnosť a bude realizovaný nezávisle od webového servera.

Jednu z najväčších výziev predstavuje získanie a spracovanie dát vo forme internetových článkov a následné indexovanie pre rýchle vyhľadávanie osôb. Internetové články budú pochádzať predovšetkým z oblasti Európskej únie, preto je však potrebné vytvoriť riešenie, ktoré bude fungovať na viacerých jazykoch.