**Slovenská technická univerzita v Bratislave   
Fakulta informatiky a informačných technológií  
Ilkovičova 2  
842 16 Bratislava 4**

**Vyhľadávanie informácií**

**Crawler – NHL Draft**

**Vypracoval: Matúš Ján Haľama / 102929 Ak. rok: 2023/2024**

**Čas cvičení: Pondelok 18:00 – 19:50 Cvičiaci: Ing. Igor Stupavský**

# Motivácia

Moja vášeň pre hokej a fascinácia okolo procesu NHL draftu ma viedla k vytvoreniu NHL Draft Crawler. Svet hokeja je plný talentovaných hráčov, ktorí vyrastajú a prechádzajú cez juniorské súťaže, pripravujúc sa na výzvu hrať v najvyššej lige. Po minulom roku, kde dominovalo Slovensko a tomto roku kde bol draftovaný nový generačný talent som sa rozhodol spraviť crawler na túto tému.

S výzvou vytvoriť takýto crawler som sa rozhodol vytvoriť nástroj, ktorý umožní získavanie a spracovanie dát o draftovaných hráčoch v jednoduchý a efektívny spôsob. Videl som príležitosť vytvoriť centralizovaný zdroj informácií, ktorý by uľahčil sledovanie hráčov. Môj cieľ je poskytnúť jednoduchý prístup k základným údajom o hráčoch, ako sú ich mená, tímy, ktoré ich draftovali, kolo, v ktorom boli vybraní, ich celkové umiestnenie a národnosť. Tieto údaje sú základné pre sledovanie hráčov a poskytujú pohľad na ich vývoj od juniorských súťaží až po vstup do profesionálneho hokeja. Pretože neznamená keď ste draftovaný skôr že ste hneď lepší a bežne sa stáva, že neskôr draftovaný sa lepšie uchytia ako tí v prom kole.

V konečnom dôsledku sa snažím vytvoriť nástroj, ktorý pridáva hodnotu komunite hokejových nadšencov.

# Prehľad súčasných riešení daného problému

V súčasnej dobe existuje niekoľko softvérových riešení, algoritmov k spracovávaniu informácií o hráčoch draftovaných do NHL. No nenašiel som softvér alebo stránku, ktoré by umožnili hľadať hráčov spôsobom roku draftu a poradia.

Jedným z významných online zdrojov je HockeyProspect.com, ktorý poskytuje hodnotenia a analýzy od skautov, zatiaľ čo EliteProspects ponúka rozsiahlu databázu hráčov z juniorských súťaží a profesionálneho hokeja. Okrem toho existujú aj špecializované softvérové nástroje, ako napríklad RinkNet Scouting Software, určené pre skautov a tímy NHL.

Existuje niekoľko softvérových nástrojov vyvinutých špeciálne pre skautov a tímy NHL. Tieto nástroje umožňujú skautom efektívnejšie zhromažďovať a spracovávať dáta o hráčoch. Príklady zahŕňajú RinkNet Scouting Software a iné podobné aplikácie.

<https://hockeyprospect.com/nhl-draft-scouting-reports/>

<https://www.eliteprospects.com/>

Existuje aj aplikácia simulátoru na Draft, kde sa môžete vžiť do kože scauta a nasimulovať si rok podľa vásho výberu:

<https://draftprospectshockey.com/draftsim/#/>

# Moje Riešenie

Moje riešenie sa skladá z 5 python scriptov.

## crawler.py

Tento kód implementuje jednoduchého webového crawlera v jazyku Python pomocou knižnice BeautifulSoup na analýzu HTML stránok. Cieľom je stiahnuť obsah webovej stránky, identifikovať odkazy na ďalšie stránky a rekurzívne ich preskúmať. Nachádza sa v ňom aj jednoduchý regex aby ukladal stránky len hráčov. Výstupom tohto scriptu sú HTML dáta v jednom .txt súbore

Stránka: <https://www.hockey-reference.com/draft/>

### Problémy:

Stránky majú ochranu proti crawlerom a preto som musel aj meniť stránku, ktorú crawlujem, kvôli zablokovanej IP. Preto bolo nutné naimplmementovať sleep() funkciu a posielať rôzne hlavičky v Requeste na stránku

## regex.py

Tento kód slúži na spracovanie HTML obsahu z uloženého súboru nadobudnutý z crawler.py. Následne prebehlo extrahovanie špecifických informácií o hráčoch draftovaných do NHL. Tieto informácie sú následne zapisované do CSV súboru.

### Problémy:

Problémy, sa vyskytli iba v dekódovaní UTF-8 charakterov pri niektorých menách.

## wikipedia.py

Tento kód využíva knižnicu PySpark na spracovanie dát pomocou Apache Spark. Jeho hlavným účelom je extrakcia informácií o hráčoch draftovaných do NHL z wikipédie, ktorý je v XML formáte. Pomocou UDF (User Defined Function) a regexov boli extrahované dáta o jednotlivých draftoch kde bola pridaná informácia o juniorských tímoch a národnosť hráčov, ktorá chýbala na predošlej stránke.

### Problém:

Wikipédia pre draft nemá jednotnú štruktúru. Taktiež do roku 1979 je nazývaná ako NHL Amateur Draft. Bolo potrebné viac regexov.

## merge.py

Tento kód vytvára spojený DataFrame z viacerých CSV súborov obsahujúcich informácie o hráčoch draftovaných do NHL. Následne prebieha spracovanie a úprava dát s cieľom odstrániť duplicity a zlúčiť informácie z vlastného crawlera. Tieto dáta sú následne uložené do .csv súboru.

## indexer.py

Tento kód vytvára index pre vyhľadávanie informácií o hráčoch draftovaných do NHL na základe údajov uložených v CSV súbore. Kombinuje indexovanie dát pomocou PyLucene s možnosťou interaktívneho vyhľadávania na základe rôznych kritérií. Používateľovi umožňuje definovať kritériá vyhľadávania a získava výsledky na základe zadaných podmienok.

# Vyhodnotenie

Podarilo sa mi vytvoriť menšiu databázu draftovaných hráčov a ich hľadanie podľa určitých kriterií. Testovanie prebehlo 4 unit testami, kde sa testovalo podľa jedného alebo viac kritérií. Indexer vrátil vždy podstatné dáta

Pri spojení s wiki sme obohatili 65.96% dát. Crawler na wiki bol určený na špecifické roky draftu pri čom každý mal vlastnú štruktúru a iný spôsob zapisovania dát.

Precision a Recall som nedokázal vypočítať. Indexer sa nedokázal dostať k TN.

# Spustenie/Inštalácia

Pre inštaláciu je potrebné mať stiahnutý Spark 3.5.0. Do tohto Sparku je nutné pridať aj spark-xml.jar.

Tiež je potrebné mať stiahnutú knižnicu : Pandas – pip install pandas  
 pySpark - pip install pyspark

Nainštalovaný docker a stiahnutý image - **coady/pylucene**