

Obsah na míru i zobrazení reklamy. Kde Seznam využívá strojové učení?

28.4.2025 - Vít Líbal | Seznam.cz

Domovskou stránku Seznamu denně navštíví okolo čtyř milionů uživatelů. Ač to nemusí být na první pohled úplně zřejmé, tato stránka je doslova napěchovaná novými technologiemi. A strojové učení (ML) a umělá inteligence (AI) mezi nimi hrají výsostnou roli.

Pokud byste se snažili vypátrat víc o ML v Seznamu, na prvním místě byste našli informace o technologiích vyhledávání, o skvělých službách jako jsou Mapy.com, Zboží.cz nebo Sreality.cz. Určitě byste zjistili i to, že Seznam vyvíjí – vlastní velký jazykový model (LLM) který už dnes pohání některé služby Seznamu.

V tomto článku se ale zaměříme na služby Seznamu, které jsou obvykle skryté, respektive se možná nezdají tak populární. Jde o služby **doporučování článků** a o **reklamní služby**. O to víc vás možná překvapí, jak sofistikovaný ML systém se pod kapotou těchto služeb skrývá.

Baví vás strojové učení? Skvělé! Přečtěte si, jak fungují naše AI sumarizace.

Služba doporučování

Obsah, který se uživateli na Seznam.cz nabízí, vybírá služba doporučování. Za personalizovaným doporučením obsahu se skrývá hned několik desítek ML modelů. Úlohy, které doporučovací ML modely řeší, jsou:

- výpočet optimálního počtu, řazení a velikosti tematických boxíků,
- sumarizace článků a jejich shluková analýza podle podobnosti obsahu,
- výběr a řazení nejvíce relevantních článků (databáze článků, ze které doporučujeme, obsahuje řádově statisíce položek),
- identifikace a výpočet charakteristik uživatele (tedy jeho profilu, včetně odhadu věku, pohlaví, zájmů),
- výpočet charakteristik článků a videí (charakteristiky jako téma, kvalita, míra „clickbaitovosti“ „lokálnost“ a lokalita, specifičnost atd.)

Pro každou charakteristiku má služba doporučování jeden dedikovaný ML model. Abyste si lépe představili, jaké úkoly v Seznamu pomocí ML modelů řešíme, pojďme se podívat na konkrétní příklady.

Předvýběr článků

Úkol:

V reálném čase (řádově desítky milisekund) vybrat ze všech položek v databázi (položka = článek, podcast, video nebo jiný obsahový formát) ty, které uživatele mohou zaujmout, tak aby zbylo maximálně 1000 položek.

Řešení:

Předvýběr se sestává ze super rychlého vyhledání nejpodobnějších položek. K tomu je nutné

předpočítat vektory článků a uložit je do speciální datové struktury umožňující rychlý přístup a vyhledání. Při předvýběru se potom vypočte podobnost mezi vektory článků a vektorem uživatele a provede se vyhledání nejpodobnějších položek – algoritmus pracuje se sub-lineární časovou složitostí.

Řazení článků

Úkol:

V reálném čase (řádově desítky milisekund) vyhodnotit relevanci až 1000 předvybraných položek tak, aby bylo možné je seřadit podle relevance pro každého uživatele v daný moment.

Řešení:

Relevance každé položky pro uživatele se na základě kompletních dat o položce a o uživateli napočte pomocí hluboké neuronové sítě Deep and Cross Network (DCN). DCN dokáže zachytit netriviální souvislosti mezi daty uživatele a daty položek. Model je trénován na historických datech a přetrénovává se na nových datech každých 5 minut.

Vyhodnocení „lokálnosti“ článků

Úkol:

Vyhodnocení „lokálnosti“ článku, videa, nebo jiné doporučované položky. Položka, kterou označujeme jako lokální, se vztahuje k nějaké specifické lokalitě, má potenciál zaujmout uživatele se zájmem o tuto lokalitu a chceme, aby služba doporučování tuto položku takovýmto uživatelům nabízela.

Řešení:

Klasifikaci lokálnosti provádí malý jazykový model postavený na základě modelu retromae-small-cs trénovaný s důrazem na sémantické embeddingy a dotrénovaný na datech, která pomáhají tvořit velký jazykový model – SeLLMa 70B od Seznamu. Tato kombinace nám umožňuje využít výhody a výkon velkého jazykového modelu s náklady na provoz malého jazykového modelu.

Doporučování bez ML? Ani náhodou

Každý den projdou službou doporučování statisíce článků, videí, podcastů, fotogalerií a dalších obsahových formátů v češtině i jiných jazycích. Bez personalizovaného doporučování by každý uživatel viděl domovskou stránku Seznamu s totožným ručně vybraným obsahem. Pro velkou část uživatelů by obsah stránky nebyl zajímavý, a proto by neměli důvod na ní trávit čas, natož se vracet.

Nejen obsah, ale i reklama na míru

Reklamní systém (RS) slouží především inzerentům, kteří za reklamy Seznamu platí. Klíčové je reklamy umísťovat tak, aby uživatele v ideálním případě zaujaly. Podobně jako u doporučování je i pro reklamu nutné pracovat s identitou uživatele, předpočítávat jeho profil a charakteristiky zobrazovaných internetových stránek.

Mezi RS a službou doporučování je řada rozdílů: v reklamě je nutné předpočítávat také charakteristiky inzerentů a jejich reklamních kampaní, požadavků na vyhodnocení umístění reklam je o celý řád více než požadavků na doporučování. Reklamy navíc soutěží o umístění na stránkách prostřednictvím aukce.

Na provozu RS se podílejí desítky ML modelů. Příklady úloh, které ML modely řeší jsou:

- predikce prokliku,
- predikce konverzí,
- stanovení optimálních cenových prahů aukce,
- optimalizace výše bidu reklamy pro aukci,
- automatizace výroby reklamních kampaní,
- vymezení publik pro cílenou reklamu,
- automatické cílení reklamy,
- klasifikace témat obsahových stránek,
- klasifikace nákupních zájmů uživatele,
- detekce lokalit zájmu uživatele,
- detekce podvodných kliků a další.

Abychom ještě lépe přiblížili využití ML v RS, pojďme prozkoumat konkrétní využití několika uplatňovaných ML modelů.

Predikce konverzí

Úkol:

V reálném čase (do 50 milisekund) předpovědět, s jakou pravděpodobností povede uživatelův proklik na reklamu ke konverzi. Úloha je vzhledem k vysoké řidkosti konverzních dat (jen cca čtyři prokliky ze sta jsou konverzní), vysoké diverzitě typů a počtů konverzí a vzhledem k velkému zpoždění konverze po kliknutí (konverze může nastat s mnohadenním zpožděním, proto započítáváme i ty konverze, které nastanou až 30 dnů po reklamním prokliku) velmi náročná.

Řešení:

Predikci toho, zda konverze nastane, respektive predikci počtu konverzí provádějí modely logistické regrese a Poissonovy regrese. Prediktory se trénují na datech za období několika měsíců a s novými daty se přetrénovávají každý den.

Optimalizace bidování

Úkol:

Určení optimální výše bidů v online aukcích tak, aby reklamní kampaň inzerenta obdržela maximální množství konverzí.

Řešení:

Pro optimalizaci počtu konverzí využíváme kontrolní zpětnovazební smyčku, která vypočte a automaticky nastaví správnou výši bidu pro každou reklamní kampaň zvlášť na základě dat o uživateli, o reklamě, o reklamním prostoru a na základě počtu došlých i očekávaných konverzí.

Klasifikace online obsahu

Úkol:

Pro každou ze stovek milionů internetových stránek, které navštěvují uživatelé Seznamu, chceme určit její téma, tedy zájmovou kategorii. Pouze s kvalitní znalostí témat každé internetové stránky, kde doporučujeme články a kam vydáváme reklamu, dokážeme zajistit, aby doporučené články, nebo zobrazená reklama byly pro uživatele relevantní a zajímavé.

Řešení:

Webové stránky klasifikujeme do standardní IAB taxonomie zájmových oblastí pomocí sofistikovaného jazykového modelu založeného na BERT architektuře – kombinuje předtrénovaný dist-mpnet model s vlastní MLP vrstvou optimalizovanou s využitím regularizačních technik a mixup

augmentace.

Příprava kvalitních trénovacích dat probíhá zcela automatizovaně. Nejprve se identifikují potenciální kandidáti pro jednotlivé kategorie na základě vektorové podobnosti textů. Následně jsou tito kandidáti prověřeni pokročilým jazykovým modelem použitím techniky „LLM as a Judge“ (LLMaaS). Tento postup zajišťuje konzistentně vysokou kvalitu klasifikace online obsahu bez ohledu na průběžně se měnící obsah a bez nutnosti manuální přípravy anotací.

Online reklama je bez ML nepředstavitelná

Reklamní systém Seznamu vyhodnocuje statisíce požadavků na zobrazení reklamy za sekundu. Online reklama bez ML by možná fungovat mohla, ale jen velmi špatně. Znamenalo by to neadresnou reklamu, každý uživatel by dostal reklamu buďto tu samou, nebo vybranou náhodně, bez ohledu na jeho profil – ta by byla pro uživatele v převážně nerelevantní. Zatížení online prostoru reklamou by pravděpodobně bylo dramaticky vyšší – aby inzerenti dosáhly stejného zásahu své reklamy, bylo by nutné vydat násobně více online inzerce (na stejném množství online obsahu).

A nešlo by to úplně bez reklamy? Rozhodně ne. Je to právě reklama, která financuje online obsah, online služby a aplikace, které všichni denně – a mnohé zcela zdarma – využíváme. Především díky reklamě Seznam uživí špičkové výzkumné týmy a může si dovolit investovat velké finanční prostředky do vývoje nových technologií včetně ML modelů i zmíněných vlastních LLM technologií. Snad o Seznamu víte, že část těchto investic se vrací ML komunitě.

Práce v Seznamu? Velké výzvy i prostor pro růst

Reklamní a doporučovací systémy patří svou složitostí, velikostí a rychlostí dat, které zpracovávají, k nejsofistikovanějším systémům s obrovskými nároky na provoz a údržbu systémů i dat. Pro všechny, kteří se podílejí na vytváření a na provozu těchto systémů, to znamená nepřetržité řešení mnoha náročných výzev. Na druhou stranu to také vytváří výjimečnou příležitost pro schopné vývojáře a výzkumníky ML a AI – pokud dokážete přispět k vývoji reklamních systémů, nepochybně si poradíte s jakoukoliv další ML doménou.

Chcete být součástí týmu doporučování? Mrkněte na výzkumnickou pozici, kterou právě obsazujeme. A prozkoumejte i další příležitosti v oblasti ML, které v Seznamu nabízíme.

Práce našich výzkumníků a vývojářů je organizovaná podle agilních metod. Výzkumný tým zadaný problém analyzuje, prozkoumá na úrovni publikovaných SOTA (State-Of-The-Art) znalostí ML postupů a technologií a navrhne řešení, které odevzdá vývoji ve formě funkčního prototypu. Vývoj prototyp implementuje do produkční pipeline a zajistí, že výsledný produkt splňuje všechny náročné provozní požadavky. Každý vývojář i výzkumník dostává dostatečný prostor pro osobní rozvoj a dbá se na systematické vzdělávání.

Zábavná a kreativní práce plná výzev s velkým přínosem pro koncové uživatele, neformální a příjemné pracovní prostředí a jedinečné možnosti profesního růstu – to jsou hlavní důvody, proč se do Seznamu hlásí noví výzkumníci a vývojáři a proč v Seznamu rádi zůstávají.

<https://blog.seznam.cz/2025/04/obsah-na-miru-i-zobrazeni-reklamy-kde-seznam-vyuziva-strojove-uceni>