

Seznam představuje nové jazykové modely pro práci s významem textů

17.4.2024 - Jakub Náplava | Seznam.cz

V Seznamu jsme v roce 2021 představili jazykový model Small-E-Czech (čti: smolíček), který se od té doby stal druhým nejpopulárnějším českým jazykovým modelem. V oblasti umělé inteligence a zpracování přirozeného jazyka je ale pořád co objevovat. A tak jsme pro vás nově vypiplali další malé jazykové modely, zaměřené na analýzu a interpretaci jazyka. Jaké jsou?

Malé jazykové modely, které vznikly v dílně seznamáckých výzkumníků, v přesnosti a schopnosti chápání češtiny představují nové současné state-of-the art. Dokonce překonávají i daleko větší modely z dílny OpenAI.

A proto jsme je s radostí už začlenili do některých našich klíčových služeb na Seznamu. Výrazně zlepšují kvalitu organického vyhledávání, poskytují přesnější a relevantnější featured snippets a pomáhají vylepšovat vyhledávání obrázků. Integrace těchto modelů do našich služeb umožní uživatelům získat ještě bohatší a přesnější výsledky.

Proč jsou seznamácké modely jedinečné?

Modely jsme postavili na osvědčené architektuře Transformer, o kterou se opírá naprostá většina nejmodernějších technologií ve zpracování přirozeného jazyka. Naše modely jsou jedinečné díky třem inovativním metodám učení, které jsme aplikovali.

Jde o RetroMAE (Masked Autoencoder for Retrospective Learning), unsupervised fine-tuning a multilingual-distillation. Tyto metody nám umožnily vytvořit modely, které excelují v porozumění českému jazyku, jsou schopné se efektivně učit z omezeného množství dat a dovedou se přizpůsobit různým kontextům.

Kromě uvedení modelů je nám potěšením podpořit i vědeckou a vývojářskou komunitu zveřejněním nového českého benchmarku. Tento benchmark je určen k testování a porovnávání výkonnosti českých jazykových modelů. Vedle benchmarku jsou k dispozici samozřejmě i natrénované modely na platformách jako GitHub a Hugging Face.

Chcete se o jazykových modelech v Seznamu dozvědět více? Přečtěte si náš odborný článek, který jsme v únoru prezentovali na AAAI konferenci ve Vancouveru.

Těšíme se na vaši zpětnou vazbu a hlavně na využívání našich modelů v různých aplikacích.

<https://blog.seznam.cz/2024/04/seznam-predstavuje-nove-jazykove-modely-pro-praci-s-vyznamem-textu>