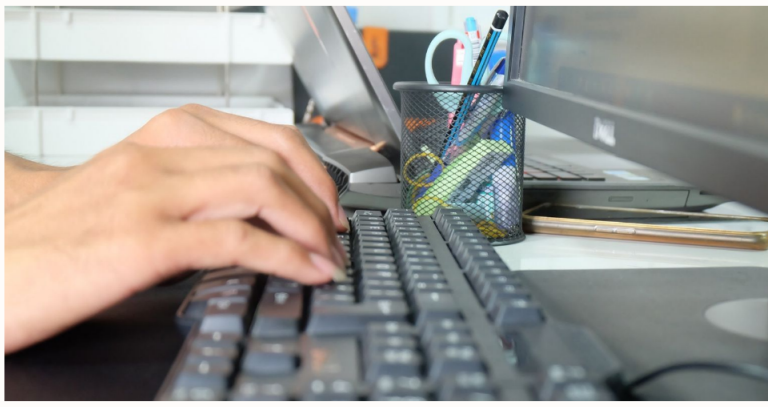
Nový algoritmus pomáhá MU vyhledávat plagiáty

Informační systém Masarykovy univerzity spustil nový algoritmus vyhledávání podobností.

Student | 4. října 2021 | Klára Petrovičová



Opisování, zhotovování napodobenin je neetická činnost a podvodné jednání, která nemá na vysoké škole svoje místo. Plagiáty mezi závěrečnými pracemi jsou nyní na MU odhalovány novým **algoritmem**, který vyvinul tým Informačního systému MU a **Theses.cz**.

Soubory v datovém úložišti v ISu jsou průběžně strojově analyzovány. Může se jednat o text seminární práce, eseje, prezentace, ale nejčastější a nejdůležitější texty na porovnání jsou závěrečné práce. Studenti si mohou sami algoritmus vyzkoušet a zkontrolovat, s jakými zdroji se jejich text shoduje, na kterých místech a na kolik procent je dokument podobný s ostatními. Stejnou kontrolu následně dělá vedoucí a oponent závěrečné práce, který tak může zhodnotit, jestli byl text opsán nebo ne, případně jestli byl dostatečně ozdrojován.

Vlastnosti nového algoritmu

Je zřejmé, že správná funkcionalita algoritmu na hledání podobností je velmi důležitá, a proto tým ISu MU a **Theses.cz** přišel s dalším vylepšením. Nově by měly být lépe odhaleny parafrázované texty, tedy takové, které student zkopíroval a lehce přeformuloval. Dále poskytuje nové funkce a modernější, uživatelsky přívětivější design. Vývojáři reagují především na podněty uživatelů a zkušenosti s novými způsoby opisování a pečlivým parafrázováním textů.

Nový algoritmus je schopen lépe poznat přeformulovanou informaci v delších pasážích. Naopak ignoruje kratší úseky, které se sice mohou podobat, ale často se jedná o používané definice, zaužívané víceslovné odborné termíny, předepsaná prohlášení v závěrečných pracích a podobné druhy textu, jejichž výskyt ve více dokumentech ještě neznamená plagiátorství.

Co se týče již zmíněného designu, oproti předchozí verzi také přibylo barevné rozlišení různých zdrojů podobných úseků. Navíc jsou skryty zdroje, které by oproti zobrazeným nepřidaly žádnou hodnotu v případě, že je dokument podobný na více textů. Tyto duplikované zdroje však lze zobrazit.

Další vítanou vlastností nového algoritmu je jeho rychlost. Navíc podobnosti lze zobrazit hned po vytvoření textové verze souboru, která je nutná pro spuštění algoritmu. Převést soubor do textové verze lze přímo v úložišti ISu.

Porovnání dokumentů napříč archivy

Systém hledá podobnosti jak napříč dokumentovým serverem IS MU, tak mezi závěrečnými pracemi škol, které jsou zapojené v systému **Theses.cz**. Dále mezi seminárními a závěrečnými pracemi v systému **Odevzdej.cz**, vědeckými publikacemi v systému **Repozitar.cz** a mezi dalšími dokumenty v informačních systémech provozovaných MU. Studentům tak neprojde ani odevzdání práce známého z jiné školy i přesto, že je databáze instituce neveřejná. Součástí systému je také časově a výpočetně náročnější algoritmus, který vyhledává zdroje z internetu. Právě kvůli jeho technickým omezením jsou archivy závěrečných prací používanější.

Porovnávané dokumenty mohou být v češtině, slovenštině a angličtině. Navíc mohou být zkoumány opakovaně a to například vyjde-li nová verze algoritmu, jako právě teď. Navíc se databáze neustále rozšiřuje o další zdroje, takže opakované prohledávání není zbytečné.

Algoritmus má smysl spouštět pro delší texty, kde má možnost odhalit podobnost s dalšími dokumenty. Ale může být spouštěn i pro odpovědníky, a to pokud se jedná o psané odpovědi, které mají více než 50 znaků.

Procento podobnosti

Jednou z nejviditelnějších informací po doběhnutí algoritmu je procento podobnosti. Toto číslo však nemusí jednoznačně určit, jestli je práce plagiátem nebo ne, ale upozorňuje na jisté podezření. Práci je vždy nutné posoudit odborníkem z daného oboru, který vyhodnotí, jestli je práce dostatečně ocitovaná a jakou závažnost podobnosti mají. Pokud autor prohlašuje nějakou část práce za svoji a procento a zobrazený text z archivu ukazují, že tomu tak není, může studentovi hrozit sankce.

Starý vs. nový algoritmus

Prozatím jsou v ISu k dispozici algoritmy oba. Přece jen fungují na trochu jiném principu, a proto mohou zobrazovat víceméně odlišné výsledky v krátkých úsecích a také v celkovém procentu podobnosti. Může se tedy vyplatit zkusit jak novou, tak starou verzi. O vypnutí staré verze budou správci ISu informovat.

Další informace o systému naleznete v **Nápovědě** a v Průvodci, který je odkazovaný přímo z aplikace, kterou najdete v možnostech dokumentu pod ikonou Vejce vejci (Vyzkoušejte novou verzi).