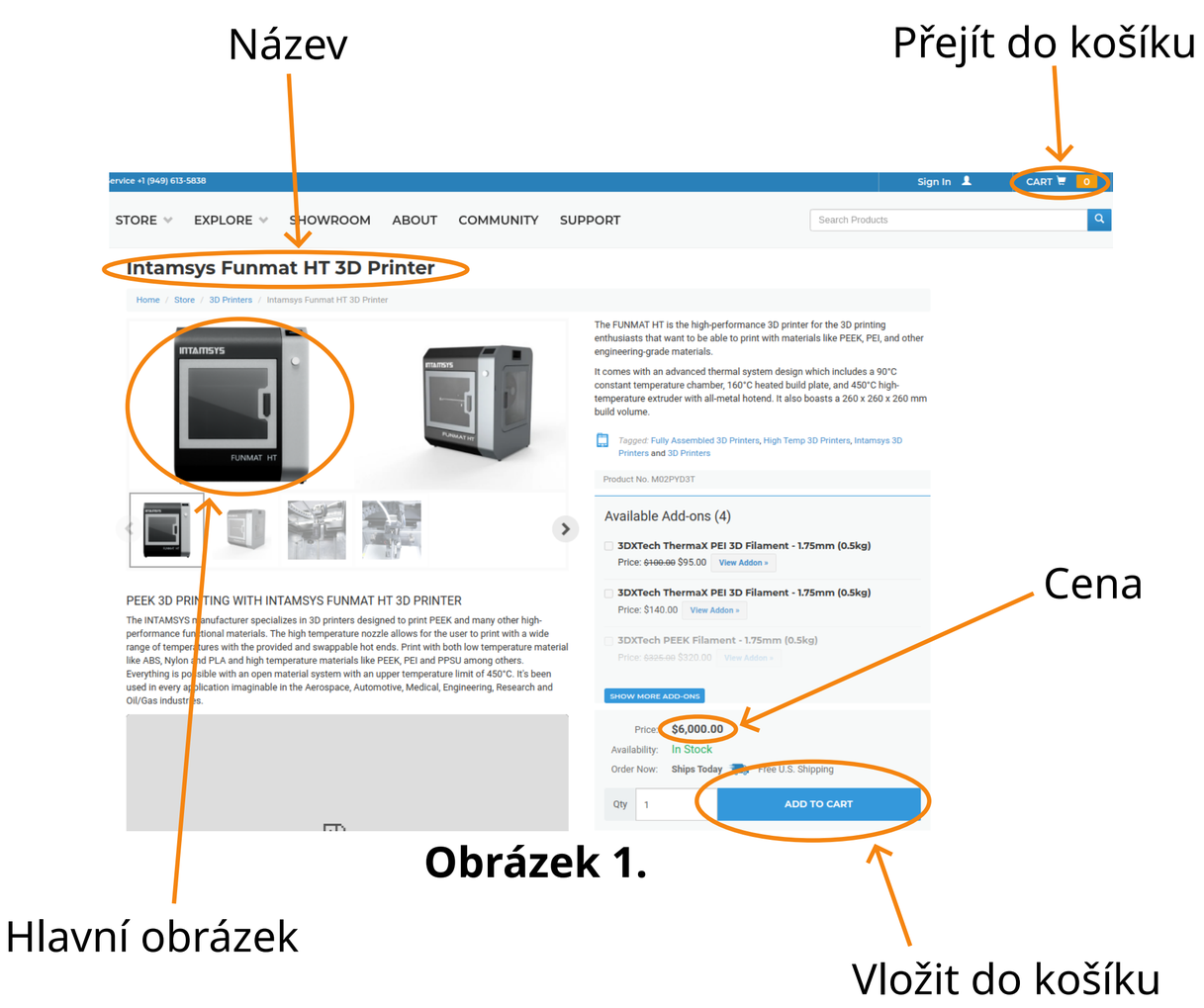
Extrakce informací z webových dokumentů pomocí grafových neuronových sítí

Autor: Josef Katrňák
Vedoucí: doc. Ing. Radek Burget, Ph.D.

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Motivace

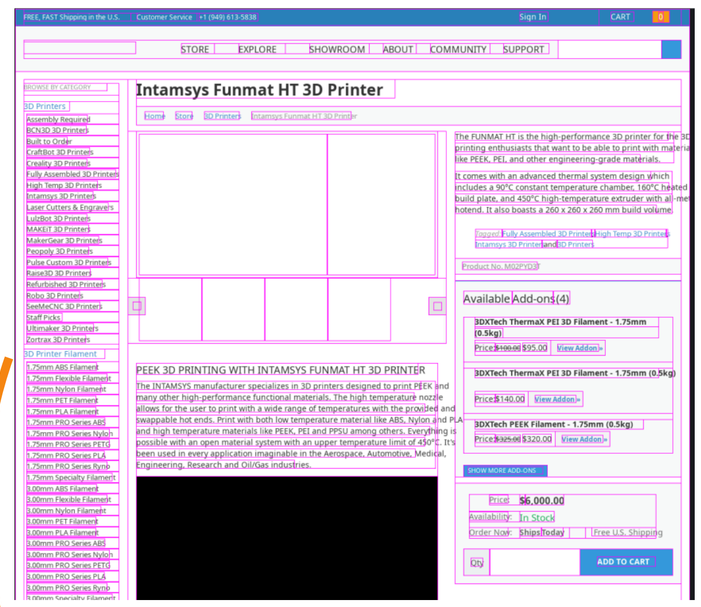
Identifikace a uložení důležitých údajů z webové stránky.



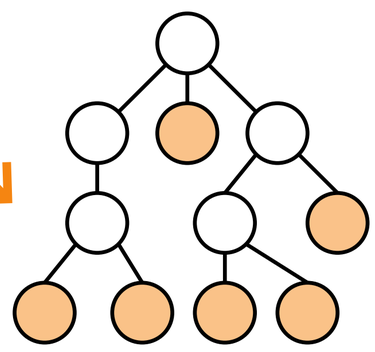
Obrázek 1.

Reprezentace

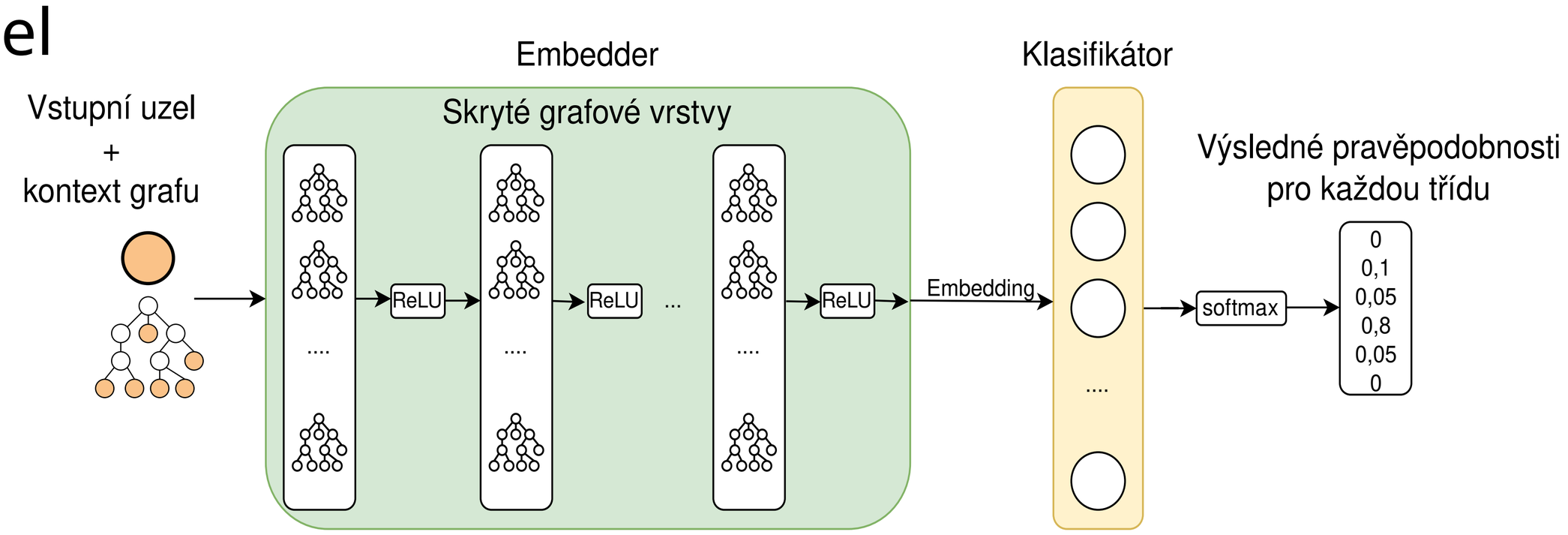
Převod zobrazené stránky na grafový model.



Obrázek 2.



Model



Obrázek 3.

Experimenty

Pro Embedder se ukázala jako nejlepší architektura grafová síť založena na *attention* mechanismu (GAT).

	Přesnost	F1	Přesnost nominace
MLP	0,8282	0,8768	0,2032
GCN	0,9618	0,9461	0,3280
GAT	**0,9660**	**0,9595**	**0,8581**

Tabulka 1.

S GAT architekturou a ztrátovou funkcí Cross Entropy Loss byla dosažena přesnost **0,9755** a skóre F1 **0,9737**. Stejná architektura, ale s váhovanou Cross Entropy Loss docílila přesnosti nominace **0,9285**.

	Přesnost	F1	Přesnost nominace
Cross Entropy Loss	**0,9755**	**0,9737**	0,6577
Weighted Cross Entropy Loss	0,8007	0,863	**0,9258**
Class-balanced Cross Entropy Loss	0,9731	0,9716	0,6505
Weighted Focal Loss	0,7566	0,8369	0,8978
Class-balanced Focal Loss	0,9141	0,9357	0,8695

Tabulka 2.

Excel@FIT 2023