# **Grabex** Analytická dokumentace

## Významy pojmů

### „Hlavní kategorie“ (main category)

Základní kategorie na hlavní stránce.

Obsah obrázku text, snímek obrazovky, mapa

Popis byl vytvořen automaticky

Seznam kategorií se nachází v <div class="mainCategories">

### „Kategorie“ (category)

Kategorie na libovolné podřízené stránce s kategoriemi na libovolné úrovni zanoření, tj. www.firmy.cz/<kategorie>/...

Obsah obrázku text, snímek obrazovky, diagram, mapa

Popis byl vytvořen automaticky

Příklady:  
www.firmy.cz/Remesla-a-sluzby  
www.firmy.cz/Vse-pro-firmy/Sluzby-pro-firmy/Administrativni-sluzby

Seznam kategorií se v HTML vždycky nachází v <div class="categories"> nehledě na to, do jaké hloubky jsme zanořeni.

### „Popis“ (description)

Obsah obrázku text, snímek obrazovky, Písmo, číslo

Popis byl vytvořen automaticky

### „Basic info“

Obsah obrázku text, snímek obrazovky, číslo, Písmo

Popis byl vytvořen automaticky

### „Labels“

Obsah obrázku text, snímek obrazovky, číslo, Písmo

Popis byl vytvořen automaticky

### „Štítky“ (tags)

Obsah obrázku text, snímek obrazovky, Písmo

Popis byl vytvořen automaticky

## Algoritmus

Program při spuštění provede přečtení webu firmy.cz (scrape), a uloží data všech firem do databáze.

1. Program nejprve přečte všechny hlavní kategorie,
2. potom přečte každou hlavní kategorii (projde celé stránkování), a načte URL všech firem.
3. Z každého URL firmy (detailu firmy) načte popis, interní ID, všechny základní informace, kategorie a štítky.

Uvedená data uloží do databáze. Každý scrape se přidá do databáze jako nová data pod novým ID scrape.