Laboratorium 7

Sprawozdanie

Jan Jędrzejewski 239529

Zadanie 1

wikipedia.org

- Wikipedia dopuszcza tylko swoich robotów internetowych: **Israbot** i **Orthogaffe**. W tym przypadku szybkie crawlowanie nie gra roli.
- Pozostałe przypadki pozwalają na powolne crawlowanie statycznych dokumentów - czyli każdy bot z ustawionym opóźnieniem, może dokonywać crawlowanie wikipedia.org, z odpowiednimi ograniczeniami na dokumenty dynamiczne oraz na dokumentację REST API

p.lodz.pl

- Strona politechniki łódzkiej nie dopuszcza scrapowania dokumentów związanych z logowaniem, rejestracją, panelem admina i profilem użytkownika
- Pozostałe dokumenty oraz dane, javascript, css, photos (jpg, jpeg, gif, png, svg) są dopuszczone do scrapowania.

facebook.com

- Wymagana jest pisemna zgoda od Facebooka na pozyskiwanie danych z ich witryny. Na tej pisemnej zgodzie zostanie również określony limit, do którego trzeba się stosować.
- Boty z tylko konkretnymi zakazanymi ścieżkami: Applebot, baiduspider, Bingbot, Googlebot, Googlebot-Image, msnbot, Naverbot, Screaming Frog SEO Spider, seznambot, Slurp, teoma, Twitterbot, Yandex, Yeti
- Wyselekcjonowane boty (witryn komunikacyjnych oraz z zawartościami medialnymi - grafiki i widea) mają dostęp do zawartości wiedo pod ścieżką /videos. Discordbot, Telegrambot, Screaming Frog SEO Spider, LinkedInBot, Googlebot, facebookexternalhit, Bingbot, Pinterestbot.

youtube.com

- Ograniczenia te same dla wszystkich botów, oprócz tych powiązanych z Google: Mediaparnters-Google-*.
- Youtube, wyklucza ścieżki związane z filmami, komentarzami, chatem textowym, wynikami wyszukiwania, społeczności użytkowników i inne.

linkedin.com

- Podobnie jak w przypadku facebooka, wymagana jest pisemna zgoda od Linkedin na pozyskiwanie danych z ich witryny. Na tej pisemnej zgodzie zostanie również określony limit, których stron można scrapować.
- Dozwolone ścieżki są związane z blogami sprzedaży i nauki, pomocą związaną ze stroną, ustawieniami strony