# Report 3

Task-urile propuse pentru aceste doua saptamani au fost:

- Dezvoltarea aplicatiei pentru adnotarea manuala a imaginilor de invatare si test
- Dezvoltarea unui webcrawler care sa faca screenshot unui numar mare de pagini web

Legat de dezvoltarea aplicatiei de labeling, am dezvoltat o solutie folosind libraria OpenCV. Functionalitatea acesteia consta in deschiderea unui screenshot si posibilitatea adnotarii manuale atat a meniurilor cat si a elementelor componente ale acestuia. Desi am investit timp in dezvolarea acestei aplicatii, am ales sa contiuam folosind tool-ul opensource numit labelImg (pe care il vom atasa). Acesta ne rezolva majoritatea problemelor legate de labeling, oferindu-ne o interfata grafica simpla in care putem sa manipulam screenshot-urile si sa exportam detaliile de care avem nevoie in diferite formate.

In legatura cu devoltarea webcrawler-ului, am incercat sa folosim Scrapy insa problema majora pe care am intampinat-o a fost lipsa existentei unei liste de site-uri pentru un domeniu specific (ex: e-commerce). Solutia cu care am venit noi a fost sa luam site-urile de pe Alexa top 1 million. Mai departe am folosit un headless browser (PhantomJS) pentru a downloada screenshoturi cu fiecare site in parte. Scriptul pe care l-am scris nu dadea un randament foarte bun, blocandu-se la 10-15 site-uri. Prin urmare am recurs la o alta solutie open source numita webscreenshot, scrisa in JS pentru PhantomJS. Chiar daca si aceasta solutie implica erori, ne faciliteaza munca lucrand pe batch-uri de aprox 100 de site-uri.

Pasul urmator este crearea data setului (filtrarea manuala a entry-urilor nepotrivite si labelingul elementelor relevante de pe pagina).

In paralel, am continuat sa rulam google object detection api pentru tensorflow pe niste exemple luate de pe internet sa vedem daca am putea

sa il folosim la problema noastra. Exemplul pe care l-am rulat este un dataset de carti de joc iar modelul de rnn poate sa detecteze cartile in realtime. Am ajuns la concluzia ca vom folosi acest model pentru a incerca sa invete elementele html (probabil un menu bar) dupa ce vom termina de creat datasetul cu site-uri web. Datasetul de carti este unul destul de mic (de 350 de imagini) si totusi modelul are rezultate bune asa ca speram ca va putea invata din datasetul nostru de dimensiuni reduse.

Screenshot tool: https://github.com/maaaaz/webscreenshot
Labeling tool: https://github.com/tzutalin/labelImg