In aceste doua saptamani ne-am focusat pe studiul sistemelor de image recognition si tehnologiilor existente de image labeling / adnotation si am parcurs cateva tutoriale precum Google ML Crash Course si Tensor Flow Neural Network Course. Am parcurs diferite lucrari de cercetare legate de:

- Indentificarea zonelor relevante din pagini web folosind DOM trees, insa acestea erau outdated, cele conprehensive fiind din ~2004.
- Annotated Transport Web Forms care creaza un model pentru automatizarea cumpararii de bilete pentru transport.

Concluziile la care am ajuns sunt ca pentru problema noastra nu exista corpusuri care sa poata fi folosite in antrenarea modelului precum si faptul ca ne-am putea folosi de service-ul de text recognition doar pentru partea de labeling a componentelor intr-un stadiu mult mai avansat al proiectului.

In urma identificarii problemei lipsei unui dataset pe care sa lucram, solutia pe care am gasit-o e urmatoarea: facem un webcrawler care sa salveze mai multe pagini cu caracteristici similare, apoi in paralel sa realizam

1. Clasificarea manuala a unor elemente de pe pagina (ex: partea de meniu, butonane, labeluri etc).
2. Dezvoltarea unui script: datele relevante de pe fiecare pagina vor fi extrase, apoi transformate in features relevante, ce pot fi folosite mai departe pentru antrenare. In legatura cu acest script am avea nevoie de sugestii: ne gandeam sa dezvoltam un script care sa tag-uiasca elemente folosind codul html din spate (sau chiar sa angajam indieni)

In saptamanile ce urmeaza vom incerca sa ne impartim urmatoarele task-uri:

- Dezvoltarea aplicatiei pentru adnotarea manuala a imaginilor de test si exportarea datelor in csv.
- Dezvoltarea unui webcrawler care sa faca screenshot si sa descarce codul html unui numar mare de pagini web. (scrapy + phantomJS)
- Dezvolarea unui script care sa automatizeze adnotarea elementelor folosind codul html.

In paralel cu crearea unui corpus vom incerca sa dezvoltam sistemul folosind CNN in Tensor Flow si sa il antrenam pe un dataset deja existent (ex: autonomous driving ) pentru o intelegere mai buna a funtionalitatii acestora.