Kutatói munka beszámoló
Commit Risk Assesment for CDH Components
2017/2018/II.

Mészáros András
Konzulensek: Wiandt Bernát, Kollár Nándor

Rövid beszámoló

Az idei félévben folytattam a kutatómunkát, amit még az előző félévben kezdtem. Az előző félévhez hasonlóan, idén is a Cloudera cégnél végeztem a kutatómunkát. A projekt során azt vizsgáltam, hogyan lehetséges előrejelezni azt, hogy az egyes commitok által végrehajtott forráskód modósítások bugjait miképp lehetne előrejelezni a korábbi commitok vizsgálatára támaszkodva.

Az előző félévből számos részeredményt felhasználtam az idei félévben. Ilyen például a Git commit-jira ticket kapcsolatból fakadó Commit típus jellemző és a már feldolgozott adatok (commitok és Jira ticket-ek) halmaza, de ide tartozik még a korábban kigyűjtött jellemzők, amelyek a gépi tanuláshoz szükségesek.

A féléves munka elején a már korábban megírt forráskódból egy framework-öt készítettem. A jelenlegi felépítés lehetővé teszi, hogy könnyen új modulokat lehessen hozzáadni a programhoz és a már elkészült programrészek módosítása is sokkal egyszerűbb. A Git commitok és a Jira ticket-ek szűrésére használt jellemzők is könnyen bővíthetők, új jellmezők könnyedén hozzáadhatók a már létezőkhöz és a felesleges jellemzők törlése is probléma nélkül végrehajtható.

Ezek után elkezdtem megvalósítani azon jellemzőket amelyeket már az előző félév során gyűjtöttem. Ilyen jellemzők például a commit méret, a hozzáadott sorok száma vagy a committer neve. Ezen jellemzők segítségével tudom leírni az adatokat az adatbázisban, illetve a jellemzők a gépi tanuláshoz is szükségesek.

Ezt követően elkezdtem, olyan commitok kézi vizsgálatát is amelyeket potenciális bug forrásként lehet megjelölni. Erre azért volt szükség mivel a gépi tanuláshoz szükséges egy tanulóhalmaz amelyet a tanuló algoritmus fel tud használni minták felismerésére az adatbázisban. Mivel a rendelkezésre álló adatok száma nagyon nagy (több mint 7000 commit komponensekként) ezért a tanulóhalmaz kézi összeállítása rengeteg időt és energiát venne igénybe.Ezért szükséges volt ezt a folyamatot automatizálni. Az automatizálási módszer kiválasztásához először irodalumkutatást végeztem, aminek eredményeképp a leggyakrabban használt módszert választottam. A folyamat egy egyszerű git blame-en alapul, amely ‘Bug’ típusú commitokból indul ki a változtatott sorokra alkalmazva a blamet.
Ezen automatizált folyamat javítása érdekében, filtereket készítettem, amelyek képesek kiszűrni azon kódváltoztatásokat a commitokból amelyek nem relevánsak számomra, ilyen kódváltoztatás a törölt fájl, a hozzáadott fájl vagy ha a változtatás egy teszt fájlon történt.

Összegzés, kitekintés

A féléves munka során elkészítettem a szükséges eszközöket ahhoz, hogy a következő félévben már tudjak analitikát végezni az adatokon.
Ezek az eszközök:

- Adatbázis
- Jellemzők
- Tanulóhalmaz

Ezek közül a legtöbb idő a tanulóhalmaz előállításával telt. Sajnos, egy jó tanulóhalmaz készítése nem triviális feladat, és sokkal több időbe telik mint ahogy gondoltam.
Az utolsó félévben már egy olyan rendszer fejlesztésén fogok tudni dolgozni, amely képes lesz egy új commitról eldönteni, hogy mekkora valószínűséggel tartalmazhat bugot. A rendszer a fejlesztők munkájának szeretné segíteni, azzal hogy felhívja a figyelmet az esetleges veszélyes commitokra. A végső döntés természetesen még mindig a fejlesztők kezében lesz, nem fog a rendszer helyettük döntést hozni.