

Önálló projekt 1

Formális és program nyelvek elemzése gépi tanulási modellekkel

Sisák László Sándor

Főleg szoftverfejlesztéssel kapcsolatos problémák:

- Automatikus kódjavítás
- Kód kommentelés
- Szemantikai egyenértékűség eldöntése - clone detection
- Kódkeresés természetes nyelvű leírás alapján
- Kód természetes nyelvre fordítása

A nyelvfeldolgozás előfeladata a **vektor-reprezentáció (beágyazás)**.
Célunk, hogy hasonló szimbólumokhoz kis különbségű vektorokat rendeljünk.

A programkód reprezentálás sajátosságai

A programkód-beágyazást megnehezíti:

- Sűrűbb információ
- Strukturális információban gazdag
- Egy-egy függvény gyakran csak egyszer szerepel a programban
- A program megírója deklarálni szokta saját változókat, függvényeket
- Importálható kódcsomagok

A fentiek miatt nem érdemes természetes nyelvi modelleket a nyers forráskódon alkalmazni.

Absztrakt szintaxis fa - AST

Gyökeres fa, ami leírja a program működését. Csúcsaiban a programban használt szimbólumok vannak. (változók, függvények, operátorok, stb.)
Jól elkülöníti a strukturális információt.

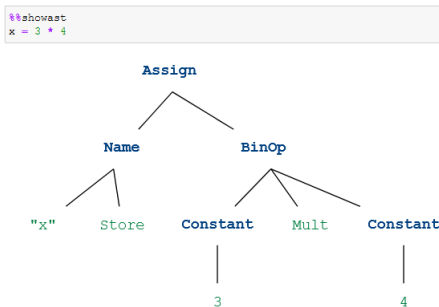


Figure: Példa AST-re

Data flow

Olyan gráf, melyen a változók közötti összefüggéseket ábrázoljuk. Ha hivatkozunk egy változóra, akkor a hivatkozásnak megfelelő csúcsból fut egy él a legutóbbi értékadásnak megfelelő csúcsba.

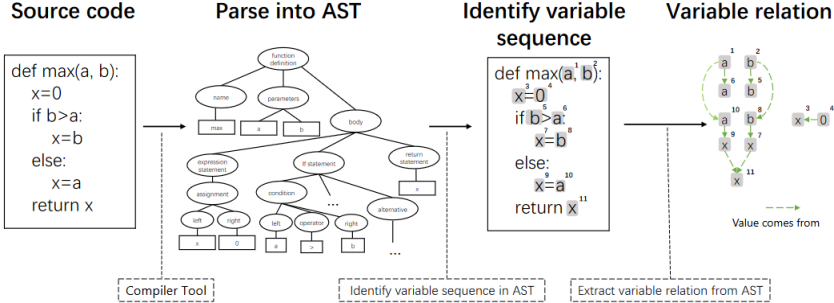


Figure: A data flow konstrukciója

Köszönöm a figyelmet!