# *HF specifikáció/dokumentáció*

A HF projektünk egy program implementálása, mely egy tudásbázis alapján megbecsüli, hogy egy adott mű mennyire hasonlít a tudásbázisban szereplő szerző munkáira.

Az általunk használt algoritmus összefoglalója megtalálható az ‘Algoritmus dokumentáció’ nevű fájlban. Alapvetően abból az ötletből indultunk ki, hogy miután szakmabeli fórumokat olvasva (pl. stackoverflow releváns topicjai) megtudtuk, hogy a PCA algoritmus egy sokdimenziós mátrixot képes kisebb dimenziójúra redukálni, és ezek a pontok jól reprezentálhatják egy szerző munkásságát, ha redukció helyett az alapmátrix-szal dolgozunk pontosabb képet kaphatunk egy szerzőről. A mátrix a tudásbázisban szereplő művek 50 leggyakoribb szavainak előfordulását tartalmazza, százalékos reprezentációban.

A mátrix reprezentációjára van a kódban egy PCA algoritmus implementálva, de tapasztalatunk szerint ez sokkal jobban összemossa az adatokat, mint pl. az eredeti mátrix első három oszlopa (ezek a mátrix felépítéséből adódóan a három leggyakoribb szó előfordulásait mutatják a művekben).

A tudásbázis a megadott adatok alapján, egy 50 dimenziós térben kiszámítja a művek ‘súlypontját’, majd a komparálandó mű ettől vett távolsága alapján mond egy becslést a stilisztikai egyezésre. Az ismert művek átlagtávolságát vesszük alapul, majd ezt egy értékkel ellátva (mi 85 százalékot definiáltunk erre, ezt tesztsorozatok alapján adtuk meg, így egy 80 százalék fölötti egyezés intuitív módon elég nagy egyezésnek számít, mint ahogy ezt elvárnánk) a többi értéket egyenes arányossággal állapítjuk meg.

Az általunk használt tudásbázishoz használt fájlok, majd az összehasonlított szövegek megtalálhatók a fingerprint és a texts mappában, a tudásbázist Shakespeare ismertebb drámái adják. Tapasztalatunk szerint, amennyiben a stílus/műfaj nem gyökeresen eltérő, a szerző saját munkái 70 százalék fölött, a nem tőle származó művek pedig 50-55 körül várhatók.

Implementáltunk egy stopword szűrést is, de tapasztalatunk alapján az általunk használt módszerhez releváns információt nyújt ezek figyelembevétele. Bár a saját munkák becslése így 90 százalék körülire ugrott, a nem a szerzőtől származó írások becslései drasztikusan megugrottak, és nagyon közeli (5-10 százalékos különbségnél kisebb) eredményeket mutatnak, sőt, más nyelvű munkák is meglepően közelinek tűnnek (Mikes vs. Brontë: 60 százalék fölött).