Specifikáció

Mapreduce

Feladat:

Ez egy olyan feldolgozó algoritmus, ami hatalmas adathalmazok feldolgozására képes. Az algoritmus 3 részből áll. Az első a Map, amikor a megkapott adathalmazt feltöri kisebb halmazokra és ezt addig folytatja, amíg nem kap elemi adatokat, ami annyit jelent, hogy nem tudja tovább szétbontani, +ezután pedig ez az elemi adat lesz a kulcs érték, ami mellé tesz az algoritmus egy értéket. Így alkot kulcs érték párokat, amiket továbbít a Shuffle fázisnak. Ebben a fázisban a kulcs érték párokat rendszerezi, hogy a hasonló kulcs értékek egy helyen legyenek, a gyorsabb feldolgozás érdekében. Végül a keletkezett rendezett halmazokat átadja a Reduce fázisnak, amiben a hasonló kulcsoknak az értékeit összeszámolja és ezeket az értékeket adja vissza eredményként a felhasználónak. Ez az algoritmus azért jó, mert egyszerre több gépen is lehet futtatni és a végén összegezni az eredményeket, így sokkal gyorsabb működésre tehetünk szert.

Például, ha van egy szövegünk és szeretnék megszámolni, benne bizonyos szavak előfordulását. Az algoritmus először felbontja a szöveget szavakra és hozzápárosít egy 1-es értéket. Aztán az ugyan olyan szavakat egy listába gyűjtve átadja a Reduce-nak, ami az azonos szavak értékeit összeszámolja és egy értéket ad vissza, ami a szavak előfordulása lesz.

Program működése:

A program célja a két betűs szavak összeszámlálása lesz. A programnak, be lehet adni egy szöveget, vagy megadni egy olyan txt fájlnak a pontos nevét, ami a programmal egy mappában van. A megadott szövegből vagy a fájlból a két betűs szavakat fogja összeszámolni. A szöveget szóközök mentén fogja szétdarabolni és az egyéb speciális karaktereket pedig le fogja vágni a szavak végéről.

A felhasználónak, be lehet gépelni a szöveget, vagy kiválaszthatja, hogy fájlból olvassa be, de akkor a beolvasandó szöveget be kell rakni a program mappájába. Alapból ugye két betűs szavakat kell megszámolni, de ha nem csak két betűs a szó, akkor a program fogja feldarabolni két betűnkként. Ha véletlenül marad egy betű aminek nincs párja, akkor azt egyszerűen el fogja dobni.

Ha a program megkapja a szöveget, akkor a feldolgozás után kiírja, hogy melyik szóból mennyi található.