**Similar Items: Plagiarism Detection**

**Initial analysis**

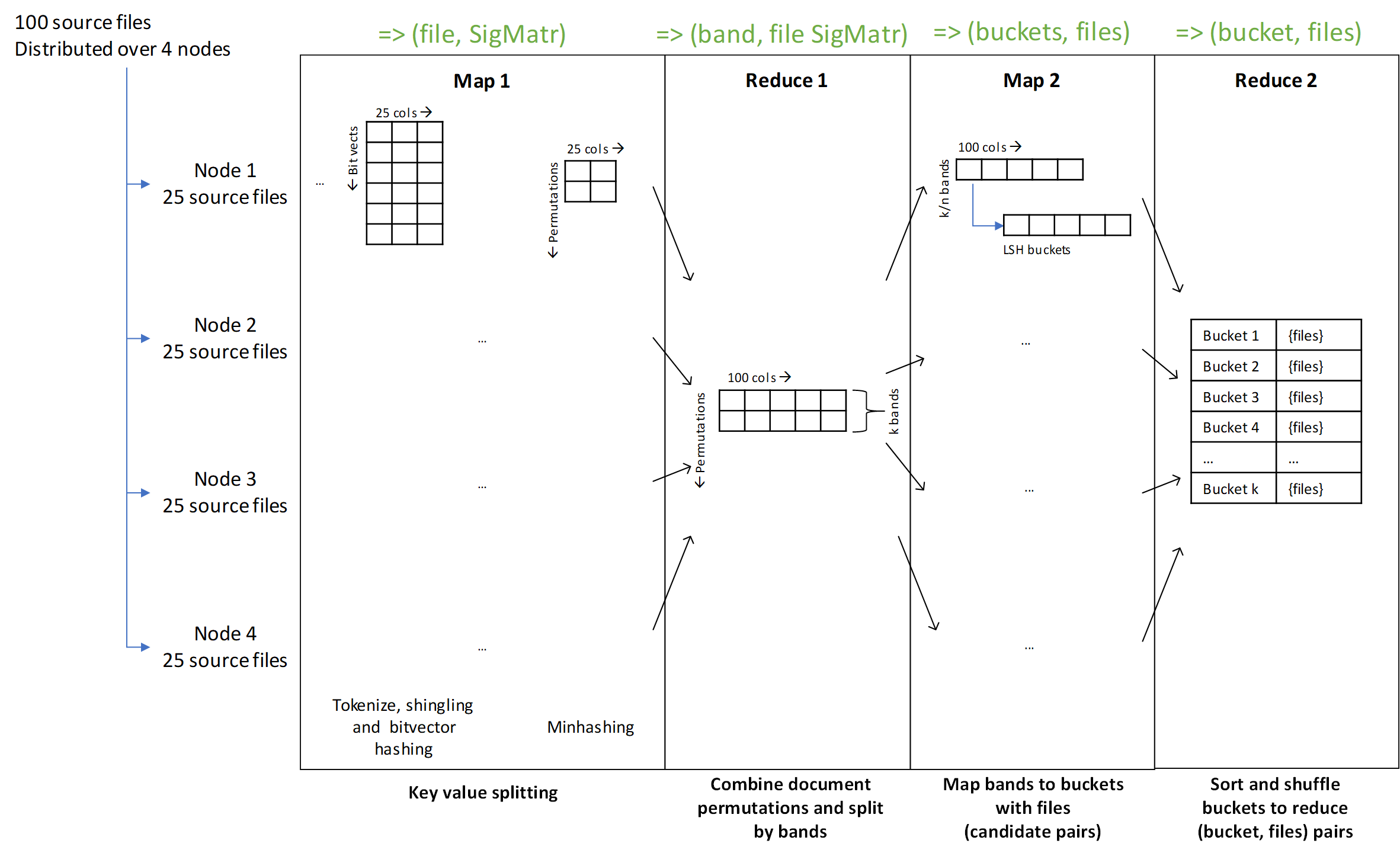
1. *How can we establish plagiarism in python source files? Keep in mind that students often tend to change solely the names of variables and functions within the code.*

* In de eerste plaats *tokenizen* we de source files. Met behulp van een Python language parser identificeren we woorden uit een source file die de naam van een variabele moeten voorstellen. Op die manier kunnen alle variabelen, klassen- en functienamen vervangen worden door “ID”-tokens. We zouden deze tokens dan als één karakter kunnen voorstellen (bijv. “$”). Zo voorkomen we dat de gelijkenis van de documenten negatief beïnvloed wordt door variabelen anders te noemen.
* Eens de broncode uniforme ID’s bevat, kunnen we deze opsplitsen in woorden van een bepaalde lengte die we vervolgens verder in k-shingles samenvoegen.

Vb “my\_list = { x \* x for x in range(10) }” => “$={$\*$for$inrange(10)}”   
=> 5-shingles: {$={$\*, ={$\*$, {$\*$f , $\*$fo, \*$for , …, (10)}}

1. *Explain how Minhash signatures can be computed in a MapReduce environment and argue about the correctness of your approach. Describe in detail how hashes are computed.*

* Indien alle bestanden gelijkmatig als chunks worden verdeeld over enkele compute nodes, kan het parallel berekenen van MinHashes voor problemen zorgen. De nodes kunnen vlot de tokenizing- en shingling-stappen uitvoeren. Bij de chunk-aanpak verliezen we echter de bundeling van een document. Je kan de MinHashing-stap enkel uitvoeren op bit-vectoren die op het volledige document zijn gebaseerd. Om de shingles naar behoren te hashen, zouden we daarom eerst alle chunks moeten defragmenteren in documenten. Deze stap lijkt ons omslachtig.
* Vermits het ontdekken van plagiaat wordt gebruikt binnen de context van schooltaken, lijkt het ons veroorloofd om te veronderstellen dat de documentgroottes niet erg variabel zijn. Dit betekent dat we documenten over nodes kunnen verdelen zonder een echt ongelijkmatige werkverdeling te veroorzaken. Alle voorgenoemde stappen kunnen dan op de nodes gebeuren zonder de omslachtige defragmentatiestap. Omdat alle documenten ‘heel’ blijven zullen de MinHash signaturen correct worden berekend.
* Om een goede Signature Matrix te bekomen, moeten we een aantal pseudo-random permutaties uitvoeren op de rij-(/bit-)indices van de gehashte bitvectoren van shingles. Deze permutaties worden gedaan via een aantal hashfuncties die de rijindices afbeelden op andere indices. Zulke hashfuncties kunnen bijvoorbeeld de vorm hebben van , met en distinct verschillende getallen voor elke functie en het aantal shingle-bits in de document input matrix.

1. *Present the end-to-end architecture of the pipeline, preferably by means of a visualization. Draw specific attention to the input and output of each of the MapReduce stages involved. Clearly indicate the key used during reduce phases and describe the logic of non-trivial map operations.*
2. De bronbestanden worden gelijkmatig verdeeld over de beschikbare compute nodes.
3. Elk bestand wordt gestreamd door een tokenizer die identifiers vervangt door bvb. `$` en whitespace negeert. N-shingles worden vervolgens opgesteld en gehasht naar een bitvector waarin elke bit aanduidt of de shingle uit de shingle-set voorkomt in het document.
4. Minhashing wordt toegepast en de Signature Matrix wordt opgesteld afhankelijk van een aantal pseudo-random gekozen hashfunctie, en dit voor elk document op de node.
5. In de volgende stap worden de Signature Matrices “samengevoegd” om voor elk bestand enkele banden uit de bitvector te kiezen voor verdere behandeling op elke node.
6. In deze banden worden mogelijke kandidaatparen beschouwd en de gelijkenis berekend. Paren waarvoor de kans op gelijk zijn boven een bepaalde grens ligt, worden naar eenzelfde bucket gehasht (Locality Sensitive Hashing).
7. In de voorlaatste stap worden de buckets van elke node opnieuw samengevoegd om een unieke set te creëren van elke file in elke bucket.
8. Ten slotte kan men de lijsten uit elke bucket nu samenvoegen om tot het finaal resultaat te komen.

In stap 1-3 worden *(file, Signature Matrix)*-paren aangemaakt in de mapping. Stap 4 maakt hiervan nieuwe *(band, file Signature Matrix)-*paren, waarbij een aantal banden per node verdeeld worden, samen met elke kolom (elke file). Stap 5 zorgt vervolgens dat deze files naar *(bucket, files)*-paren worden omgezet, die in stap 6 over elke node gelijkmatig verzameld worden, waarbij node 1 bijvoorbeeld aantal buckets krijgt die hij moet samenvoegen. In de laatste stap produceren we dan opnieuw *(bucket, files)*-paren, maar deze keer werden de lijst met bestanden reeds samengevoegd in de vorige stap. Elke bucket bevat een indicatie van de waarschijnlijkheid dat de bestanden in die bucket gelijkwaardig waren. Deze files kunnen vervolgens in meer detail vergeleken worden om na te gaan wat er precies in voorkomt dat gelijkaardig was aan de andere bestanden in diens bucket.

**Implementation report**

**MapReduce dataflow**

1. Input files lezen

**Input:** (list\_of\_file\_paths)

* 1. Lijst van files in queue;

1. **Voor elke file:** K-shingles opstellen van tokens

Deze stap kan apart op elke node worden uitgevoerd en komt overeen met een Map-fase.

**Input:** (list\_of\_file\_paths)

**Intermediary:** (filename, file\_contents)

**Intermediary:** (filename, shinglelist: [shingles])

**Output:** (filename, multiset: {shingle\_hashes})

* 1. K factor bepaald door input parameter;
  2. Inlezen van een file
     + Omzetten naar tokens (reeds gebeurd)
  3. *NAME* tokens omzetten naar vb “$” in plaats van de naam zelf;
  4. Andere tokens opnieuw aggregeren met hun oorspronkelijke waarde;
  5. Elke *k* tokens samenvoegen in shingle en hashen naar de hash multiset van de ingelezen file;
  6. Goede hashfunctie nodig die groot genoeg is om collisions op shingles te vermijden  
     Vb MD5 of SHA256?

1. **Voor elke file:** Minhashing

Een node kan nu enkele files opvragen uit voorgaande output en eventueel zo veel mogelijk files nemen zodat ofwel de grootte van de benodigde Signature Matrix (#files\*#pHashes) plus (#files \* grootte van ieders multiset) nog steeds in diens geheugen past, of zodat elke node een gelijk aantal files heeft. Voor elk van de files wordt dan gelijktijdig de Signature Matrix samengesteld.

**Input:** (filename, multiset: {shingle\_hashes})

**Output:** (filename, [signature\_matrix\_col])

* 1. Permutatiehashes opstellen (zelfde hashes worden gebruikt op elke node)
     + Vb pHash = (value) => MD5(salt + id + value)   
       Met *salt* een vaste waarde, *id* het id van de permutatie en *value* de waarde die gehasht moet worden.
  2. Signature Matrix samenstellen
     + Normaal gebeurt dit door een permutatie te maken van de indices in de bitvector representatie van de shingle multiset. Het aantal mogelijke shingles is wellicht zo groot dat dit niet mogelijk blijkt. In plaats daarvan lopen we bij elke file enkel de hashes in de multiset af en kijken we welk van deze hashes een minimum waarde oplevert door er *pHash* op uit te voeren. Deze minima voor elke *pHash* vormen dan de Signature Matrix.
     + Voorbeeld implementatie:

# For each col/file in SignMatrix

**for** col **in** range**(**len**(**SignMatrix**[**0**])):**

# For each hashed shingle in the file's multiset

# This is the domain we are interested in

**for** x **in** hashed\_shingle\_multiset[col]**:**

# Iterate over permutation hashes (rows from SM)

**for** row**,** h **in** enumerate**(**pHashes**):**

hashval **=** h**(**x**)**

# Check if the new value is lower, and replace

**if** hashval **<** SignMatrix**[**col**][**row**]:**

SignMatrix**[**col**][**row**]** **=** hashval

1. **Herverdelen van Signature Matrices over nodes**: Locality Sensitive Hashing

Het LSH-proces in onafhankelijk van welke files of banden *b* beschouwd worden. We kunnen deze dan ook opsplitsen en op elke node een set van buckets opstellen die de gehashte banden bevat. Daarna kunnen de buckets van elke node gereduceerd worden naar één set van buckets die aangeven hoe groot de kans is dat files in dezelfde buckets zitten ook effectief gelijk zijn met een bepaalde kans *t* (threshold).

**Input:** (filename, [signature\_matrix\_col])

**Intermediary:** (band, (filename, signature\_matrix\_band))

**Output:** (band, buckets: [file\_names])

* 1. Kiezen van aantal banden en rijen per band (*b* en *n*)
     + Bepaling van deze factoren zorgen voor een verhouding tussen false-positives en -negatives
  2. Elke node stelt buckets op voor de LSH-hash functie
  3. Elke band van elke Signature Matrix per file kan apart beschouwd worden, dus deze kunnen verdeeld worden over alle nodes zodat elke node evenveel items heeft om te behandelen. De hashes in zo’n band worden samengevoegd en met de LSH-hash gehasht naar een bucket.
  4. Elke band krijgt zijn eigen buckets toegekend.

1. **Reduceren** van buckets over elke node

Als laatste stap voegen we alle buckets (per band) van elke node samen en krijgen we een set van verschillende lijsten. Elk item in dezelfde lijst had een band gemeenschappelijk met elk ander item en de kans dat deze items effectief gelijk zijn is minstens *t*, de kans die we in de vorige stap beschouwden.

**Input:** (band, buckets: [file\_names])

**Output:** (file\_name, candidates: [files])

* 1. Alle overeenkomstige buckets van dezelfde band over elke node kunnen worden samengevoegd;
  2. Indien items voorkomen in dezelfde bucket, is de kans dat ze *similar* zijn, bepaald door de *b* en *n* factors bij het kiezen van banden met een threshold *t*. Deze items kunnen dan ook nader vergeleken worden om de echte gelijkwaardigheid te verifiëren.
  3. Voor elke file gaan we na met welke andere file deze in een bucket zat en stellen we een lijst op van die files.

1. Explain how your final implementation differs from the one presented in the initial analysis. Elaborate on the computation of the Minhash signatures and argue the correctness of your approach.
   * In onze analyse zijn we er impliciet vanuit gegaan dat de karakteristieke matrix (of bitvector) in het geheugen van een node past. Vermits dit uiteraard niet het geval is, moesten we een nieuwe minhashing-aanpak zoeken.
   * Bij onze eerste poging maakten we gebruik van het feit dat iedere shingle door een hashfunctie werd omgezet naar een hash. Dit betekende dus dat als we het volledige beeld van de hashfunctie controleerde, we alle mogelijke shingles hadden gecontroleerd (zonder hierbij alle shingles in memory te laden). Uiteraard zouden ook heel veel niet relevante hashwaardes onderzocht worden. Veronderstellende dat we MD5 als hash-functie gebruikten zouden we dan 2128 verschillende hashes moeten controleren, hetgeen uiteraard niet haalbaar was.
   * Na het minhashing-algoritme uit de cursus opnieuw te bestuderen, konden we een verband vinden tussen de gehashte shingle en de hash van die gehashte shingle. I.p.v. het hele beeld van MD5 te onderzoeken, gaan we nu enkel de shingles van één file beschouwen. Iedere hashfunctie gaat dan alle shingles hashen. Elke hashfunctie creëert dan een verzameling van gehashte ‘shingle-hashes’. Uit die verzameling wordt dan de kleinste waarde geselecteerd als entry in de signatuurmatrix.
   * Correctheid:
     + In het boek en in de slides wordt het minhashing-algoritme voorgesteld alsof de karakteristieke matrix beschikbaar is. Je kan dan die matrix rij per rij afgaan. Indien er 1’en voorkomen in een bepaalde kolom van een rij controleer je de corresponderende hashwaarde voor die rij op die kolom en vervang je de waarde in de signatuurmatrix indien de hashwaarde van de huidige entry groter is.
     + Voor een enkele file is het echter niet nodig om de volledige karakteristieke matrix ter beschikking te hebben indien je zijn kolom binnen de signatuurmatrix wilt berekenen. De aanpak uit het boek zorgt ervoor dat je in een enkele iteratie door de karakteristieke matrix de volledige signatuurmatrix kan berekenen.
     + In feite bestaat de kolom van een file binnen de signatuurmatrix uit niets meer dan de kleinst mogelijke waardes die de hashfuncties konden outputten voor de verzameling van shingles uit een bestand.
     + Het komt daarom op hetzelfde neer om enkel hashes te berekenen van de ‘shingle-hashes’ en vervolgens voor iedere hashfunctie de kleinste outputwaarde te selecteren.
2. Present the final architecture of the pipeline, preferably by means of a visualization. Draw specific attention to the input and output of each of the MapReduce stages involved. Clearly indicate the key used during reduce phases and describe the logic of non-trivial map operations.
   * <text>
     + <more text>
3. Discuss the results produced by your implementation, describe potential improvements and issues encountered along the way
   * De output resulteert in een bestand dat op elke lijn een filenaam heeft staan, gevolgd door een lijst met andere filenamen.
   * De kans dat een kandidaat key-value paar effectief een voldoende gelijkenissen vertonen hangt af van het aantal banden en het aantal rijen per band dat we kozen namelijk 100 permutaties die verdeeld werden over 20 banden met 5 rijen elk.
   * Mogelijk zou een bepaalde file ‘A’ die gelijkenissen vertoont met een file ‘B’, niet per se willen zeggen dat ‘B’ gelijkenissen met ‘A’ vertoont. Bijv. 80% van file A komt voor in file B, maar dit wil dus niet zeggen dat 80% van file B in file A voorkomt. In onze output zou dit niet mogen voorkomen omdat files in dezelfde bucket steeds bij elkaar worden toegevoegd.
   * Na het berekenen van de gelijkaardige files, kunnen ze nader worden vergeleken om precieze verschillen en overstemmingen vast te stellen.
   * Eventueel zouden de signatuurmatrices van elke file als tussenresultaat kunnen opgeslagen worden zodat bij het vergelijken van nieuwe files de oude files niet opnieuw geparset zouden worden
     + <more text>