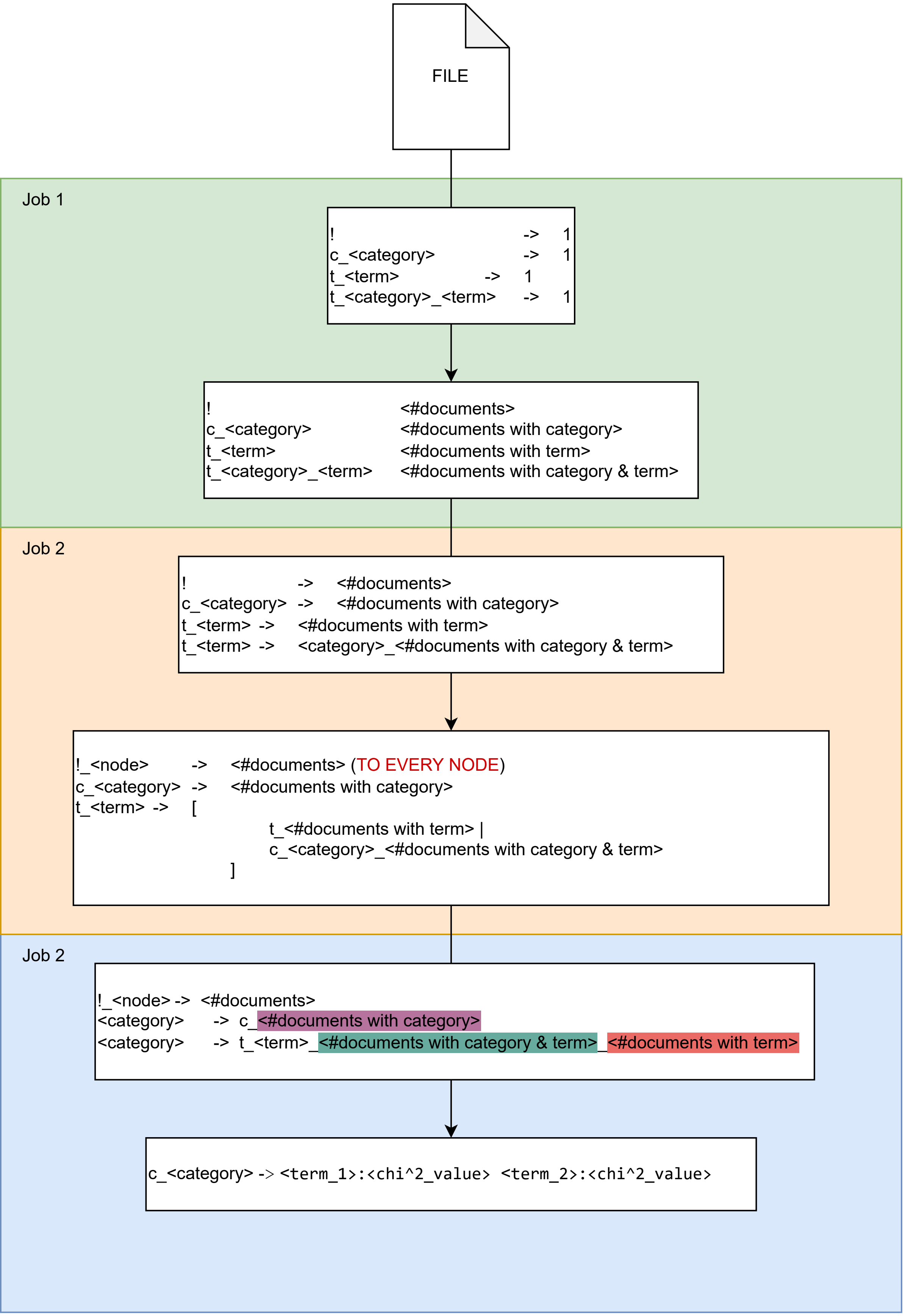

FILE

Job 1

```
!                          ->    1
c_<category>               ->    1
t_<term>                ->    1
t_<category>_<term>        ->    1
```

```
!                         <#documents>
c_<category>              <#documents with category>
t_<term>                  <#documents with term>
t_<category>_<term>       <#documents with category & term>
```

Job 2

```
!               ->    <#documents>
c_<category>  ->    <#documents with category>
t_<term>  ->    <#documents with term>
t_<term>  ->    <category>_<#documents with category & term>
```

```
!_<node>        ->    <#documents> (TO EVERY NODE)
c_<category>  ->    <#documents with category>
t_<term>  ->    [
                      t_<#documents with term> |
                      c_<category>_<#documents with category & term>
                ]
```

Job 2

```
!_<node> ->  <#documents>
<category>     ->  c_<#documents with category>
<category>     ->  t_<term>_<#documents with category & term>_<#documents with term>
```

```
c_<category> -> <term_1>:<chi^2_value> <term_2>:<chi^2_value>
```

Count

- Total Docs
- Docs with category
- Docs with term
- Docs with category X term

Copy key/values:

- counted documents
- counted categories

Map the counted term and the counted category X term to the term

Mapper:

Copy key/values:

- counted documents
- counted categories

Map counted term and the counted category X term to the category

Reducer:
First read total number documents and store it in variable.
Then calculate chi square values.
Every value necessary is mapped to the category (see calculations)

Calculations

A ... number of documents in c which contain t => <#documents with category & term>
B ... number of documents not in c which contain t => <#documents with term> - <#documents with category & term>
C ... number of documents in c without t => <#documents with category> - <#documents with category & term>
D ... number of documents not in c without t =>
    <#documents> - <#documents with category> - <#documents with term> + <#documents with category & term>