# Product overlap calculator

Please create a script that calculates the overlap of products between 2 products. Script should be written in Python.

**What data do we have?**

Input file to work with is a csv file with 2 columns:

* Column 1: keyword
* Column 2: productid

**What needs to be calculated?**

For each keyword in list, please calculate how much products are same to each other keyword in list.

Keyword that will be checked against all other keywords = SourceKeyword (SK)

Keyword from all other keywords list with which we check overlap = ComparisonKeyword (CK)

For example:

For input file here <https://www.dropbox.com/s/1pzc3k9h9s1pkf5/20200213-ProductIdPerKeyword.csv?dl=0> we have the following Source Keyword with it’s product id’s.

|  |  |
| --- | --- |
| **SourceKeyword (SK)** | **ProductId** |
| robe salopette en jean | Z903085\_18M |
| robe salopette en jean | Z903085\_3Y |
| robe salopette en jean | Z903085\_6M |
| robe salopette en jean | Z903085\_9M |
| robe salopette en jean | Z903085\_10Y |
| robe salopette en jean | Z903085\_12M |
| robe salopette en jean | Z903085\_2Y |
| robe salopette en jean | Z903085\_3M |
| robe salopette en jean | Z903085\_4Y |
| robe salopette en jean | Z903085\_5Y |
| robe salopette en jean | Z903085\_6Y |
| robe salopette en jean | Z903085\_8Y |

We also have the following Comparison Keyword with it’s product id’s

|  |  |
| --- | --- |
| **Keyword** | **ProductId** |
| salopette en jean | W040076\_NAV-10 |
| salopette en jean | W040076\_NAV-14 |
| salopette en jean | W040076\_NAV-18 |
| salopette en jean | W040076\_WHI-6 |
| salopette en jean | Z903085\_18M |
| salopette en jean | Z903085\_3Y |
| salopette en jean | Z903085\_6M |
| salopette en jean | Z903085\_9M |
| salopette en jean | Z952331\_12M |
| salopette en jean | Z952331\_18M |
| salopette en jean | W040076\_NAV-12 |
| salopette en jean | W040076\_NAV-16 |
| salopette en jean | W040076\_NAV-6 |
| salopette en jean | W040076\_NAV-8 |
| salopette en jean | W040076\_WHI-10 |
| salopette en jean | W040076\_WHI-12 |
| salopette en jean | W040076\_WHI-14 |
| salopette en jean | W040076\_WHI-16 |
| salopette en jean | W040076\_WHI-18 |
| salopette en jean | W040076\_WHI-8 |
| salopette en jean | Z903085\_10Y |
| salopette en jean | Z903085\_12M |
| salopette en jean | Z903085\_2Y |
| salopette en jean | Z903085\_3M |
| salopette en jean | Z903085\_4Y |
| salopette en jean | Z903085\_5Y |
| salopette en jean | Z903085\_6Y |
| salopette en jean | Z903085\_8Y |
| salopette en jean | Z952331\_3M |
| salopette en jean | Z952331\_6M |
| salopette en jean | Z952331\_9M |

The outcome of the calculation I expect is a csv file with following columns below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SK** | **SKNbrProducts** | **CK** | **CKNbrProducts** | **NbrSKProductsInCK** |
| robe salopette en jean | 12 | salopette en jean | 31 | 12 |
| salopette en jean | 31 | robe salopette en jean | 12 | 12 |

For keyword **robe salopette en jean** ( 12 products), we found that comparison keyword **salopette en jean** (31 products) has 12 products that have exactly same productid’s.

For keyword **salopette en jean** ( 31 products), we found that comparison keyword **robe salopette en jean** (12 products) has 12 products that have exactly same productid’s.

The output file should be csv with 5 columns above and since our input file has 3112 unique keywords, I expect 9681432 lines in output file.