30517 – Anonymizer

The problem

You have a text file with some data and want to send the file to a company to do some data analysis on it.

However, the file contains some privacy restricted data you don’t want to disclose

The problem

You have this file:

```
Name, Birthdate, Country, Degree
Joanna, 5-5-2000, US, Mathematics
Susan, 12-3-1995, UK, Physics
Francine, 11-28-1997, FR, Biology
Ping, 1-30-1996, CN, Chemistry
```

…

but want to send this file:

```
ID, Country, Degree
1, US, Mathematics
2, UK, Physics
3, FR, Biology
4, CN, Chemistry
```

…

Thinking general

In a real world scenario, this situation is going to repeat over and over, with possibly some minor changes.

The best solution is to write a function that is sufficiently flexible to adapt to some cases.

We call this function `anonymize()`

`anonymize()` arguments

- Input file (string)
- Output file (string with default = `'out.txt'`)
- Separator (string with default = `','`)
- New header (string with default = `'ID')`
- How many columns to “hide” (number) starting from the left

`anonymize()` interface

```
anonymize(f_in, k, sep=’,’, header=’ID’, f_out=’out.txt’)
```

How it works

Every line in the input is like

`Name, Birthdate, Country, Degree`

If we want to remove the first k columns, we need to find k-th occurrence of the separator and split the string at exactly that point.

That is, if k = 2 and the input line is

`Joanna, 5-5-2000, US, Mathematics`

we want to split the string into two substrings

`Joanna, 5-5-2000` and `US, Mathematics`

and write

`ID, US, Mathematics`

to the output

Flow of the function

1. Open the input file
2. Read the next line in the input:
3. If we are at the end of the input, STOP; otherwise GOTO 4
4. Split the line at the k-th occurrence of sep: left part, right part
5. Join the substitute header, sep, and right part into an output string
6. Print the output string to the output file
7. GOTO 2

split_at

- split_at(s, sep, k)
- Splits a string s in two pieces, at the k-th occurrence of separator sep
- s = the string to split
- k = occurrence
- sep = the separator
- Returns a tuple with two strings:
  - the substring of s before the k-th separator
  - the substring of s after the k-th separator