Assignment 1

due 25/2/2022

This assignment is based on real data and a real problem that cropped up during a recent genealogical research project. The raw data is culled from the website

`civilrecords.irishgenealogy.ie`

that holds Irish marriage records from the nineteenth century. The original records are based on dusty handwritten ledgers (see sample_page.pdf) but there is an online index, searchable by name etc. Unfortunately, while each record does enumerate various details of an individual's marriage, it omits one crucial detail: the name of the spouse! Our task is to identify a couple based on partial information about them: the woman's name (Mary Roche), the man's first name (Nicholas) and an approximate date of their marriage (1870 to 1885).

Results of searches (here a search for marriages involving women named "Mary Roche" in 1880) are returned as text in a fixed format as shown. The details of each individual marriage (the underlined bits such the individual's name, district, year and so on) vary from entry to entry, but are embedded in a fixed template (the spacing, line breaks and labels such as "Returns Year" etc.) that is precisely the same for all entries.

```
Marriage of MARY ROCHE
in 1880
Group Registration ID        N/R
SR District/Reg Area        Enniscorthy
Returns Year        1880
Returns Quarter        3
Returns Volume No        4
Returns Page No        276

Marriage of MARY ROCHE
in 1880
Group Registration ID        N/R
SR District/Reg Area        New Ross
Returns Year        1880
Returns Quarter        3
Returns Volume No        4
Returns Page No        349

. . .
```

The records of all the Mary Roches of interest (104 of them) are available in the text file `mary_roche.txt`, while those of the Nicholases of interest (919) are in `nicholas.txt`.

Notice the records does not name spouses directly, but we can identify likely couples by matching their record details i.e. if two records one from `mary_roche.txt`, the other from `nicholas.txt` share the same particulars they are likely to be a couple.

The name of the district and the year/quarter/volume/page refer to the precise location of the original marriage record in the old handwritten ledgers: husband and wife will both be listed (by name) in the ledger and will share the same year/quarter/volume/page numbers in the index.

Write a Python function `find_couples(brides, grooms)` that prints all pairs of matching records (records of two individuals, one from each list, that match up in terms of district, year, quarter, volume and page number). The results should be printed in the format shown. There should be seven matches in all.

```
Possible match!
NICHOLAS WALSH and MARY ROCHE in Wexford in 1876
Quarter = 1, Volume = 19, Page = 466

Possible match!
NICHOLAS HOWLETT and MARY ROCHE in New Ross in 1879
Quarter = 4, Volume = 4, Page = 375
. . .
```

You may find Python's regular expression machinery useful when "parsing" the two text files and extracting the details of the records contained therein. Try to frame a multiline regular expression that captures an entire marriage record and use finditer to extract the individual records from each file. Note that Python will concatentation two adjacent strings even without an explicit concatenation operator between them. This may be useful in composing a multiline regular express as a (raw) string.

Your submission must conform to the conditions specfied below.

1. Code must conforms tho the usual programming style strictures.
2. Code must adhere to specified naming: (1) the file is named `a1.py` and and the function `find_couples`.
3. No extraneous code should be included in the file i.e. code outside the `find_couples` function apart from any necessary imports and constant definitions. Any test code should be confined to a separate file (to to be submitted).
4. Code should not include:
   - Any input statements
   - Any hard-coded file path names