CS2400 Homework 6, Email Addresses
40 Points

Search engine companies, like Google, often search internet websites for the purpose of data collections. Programs, called web scrapers, collect information for the purpose of indexing the sites and collecting other information such as email addresses, phone numbers, etc. You've been asked to access a website and only extract email addresses from it. The site file often has the extension ".html". See below for an example site that you can use.

Email addresses are tagged within the site as follows (note that other formats may be included):

```
<a href="mailto:bob@ohio.edu"> Send email </a>
<a href="mailto:bob.smith@ace.ohio.edu"> Send email </a>
<a href="mailto:bob@bob-cats.ohio.edu"> Send email </a>
```

Where "`bob@ohio.edu`" is the email address, "`bob`" is the user name, and "`ohio.edu`" is the domain name.
Write a program that processes a website file and extracts all the email addresses from the site and stores the emails in *parallel* arrays or vectors (emails, users, domains). If you're using arrays, you may assume that the number of emails will not exceed **1000**. You only need to extract email addresses that conform to the tag formats specified above.
Output the following the number of lines process and the number of unique emails extracted to the screen:

```
51 lines processed
20 emails found
```

Write a function that outputs the data to a file as follows:

```
Email                          user           domain
-------------------------------------------------------------
bob@ohio.edu                   bob            ohio.edu
bob.smith@ace.cs.ohio.edu      bob.smith      ace.cs.ohio.edu
cs2400@gmail.com               cs2400         gmail.com
bob@bob-cats.ohio.edu          bob            bob-cats.ohio.edu
```

Read the input file one line at a time (**hint: use getline**) and process it. Note that a line may have more than one email address. Process lines until the end of the input file is reached. For each email address extracted, split it into a *user* and a *domain*. Use three arrays to store the email addresses, users, and domains. **Before storing the email into the arrays, make sure the email is not in the array already**. The array of emails should only contain unique email addresses. Your program should only output unique email addresses to the output file.

The name of the input and output file names must be provided at the command line. For example:

```
./a.out website.html output.txt
```

CS2400 Homework 6, Email Addresses
40 Points

Report errors if the number of arguments is incorrect or the either file is not accessible.

You may use any function or library discussed in class or in the chapters we covered from your textbook. Do not use any other libraries or functions.

Hints:

- `splitEmailAddress`: A function that splits an email address and returns the two parts of it as reference parameters.
- `isFound`: A search function for an array of strings.
- `getLineEmails`: A function that takes a string and extracts all the emails from a single line into the parallel arrays and check for uniqueness. You may want to call `splitEmailAddress` each time.

How to get a sample data file

- Browse to the website
- https://www.ohio.edu/engineering/about/people/departmental-listing.cfm#ElectricalEngineeringandComputerScience
- Save the source code of the file
  - In Chrome: Right click on the page background, select save as "html" file. Choose a name for the file.
  - In Safari: Right click on the page background, select "Save Page as" and make sure the "Page Source" is selected. Choose a name for your file.
- A sample file is provided with the assignment.

Grading:
Programs that contain syntax errors will earn zero points.
Programs that do not include functions will also earn zero points.
Programs that use global variables other than constants will earn zero points.

Your grade will be determine using the following criteria:

- Correctness (36 points)
  - (7 points) Processing each line
  - (7 points) splitting the email
  - (7 points) Correct number of emails retrieved
  - (5 points) Verifying unique email addresses
  - (5 points) Output is clear and as requested.
  - (5 points) Error checking for command line arguments and files
- Style & Documentation (4 points)

Follow the coding style outline on GitHub:

https://github.com/nasseef/cs2400/blob/master/docs/coding-style.md