# Computer Science 1 — CSCI 1100
# Lab 8 — Sets

## Lab Overview

This lab uses sets to illustrate basic text processing. We will be working with the definitions of clubs from the Rensselaer Union. Our aim to use the definitions of these clubs to compare them and make recommendations using set processing.

To get started please download the file `lab08_files.zip` from the Course Materials section on Submitty. This file includes a few files that are descriptions of individual clubs such as `polytechnic.txt`, `wrpi.txt`, `gmweek.txt`, and `redarmy.txt`. There is also a bigger text file that contains all the clubs in the Union `allclubs.txt`.

For checkpoints 1 and 2, we will use the smaller files for testing. For checkpoint 3, we will work with the whole Union. In all parts, you can hardcode file names for simplicity and concentrate on the logic.

## Checkpoint 1: Sets of words

In this checkpoint all you have to do is a bit of data cleaning.

Write a program that reads the description of a single club. You can see that each of the example files has a single line which contains the name of the club and the description separated with a vertical line (`|`).

Now, write a function `get_words()` that takes as input the description part of a club as a string. Your function must construct and return a set containing all the words in the description based on the following process:

- remove all punctuation symbols: dot, comma, parentheses, and double quotes (`.,()"`) by replacing them with a space.
- make all words lowercase.
- keep only words with 4 or more characters that contain nothing but letters (`str.isalpha()` will get you there).

You must use a function for this part, it will become important for the remainder of the lab. For example, here is the set for `wrpi.txt`:

```
File wrpi.txt 33 words
{'effective', 'broadcast', 'local', 'programs', 'wrpi', 'located', 'bands',
 'alternative',  'miles', 'year', 'watts', 'affairs', 'radio', 'programming',
 'studios', 'special', 'first', 'floor', 'includes', 'live', 'days', 'events',
 'wide',  'campus', 'station', 'experimental', 'cultural',  'music', 'around',
 'public',  'simulcasts', 'sports', 'range'}
```

Note: words in sets have no ordering, so the words may be ordered differently in your set. All we care about is that it has the same words.

Once done, use your function to find the set of words for some of the input files and print the result. Test your code on a few of the files.

**To complete Checkpoint 1:** Show your code and output for several files to your TA or a mentor.

## Checkpoint 2: Comparing clubs

Copy your file from checkpoint 1 to a new file called `check2.py`. You are now going to compare two clubs using the code you have just written. This should be pretty easy.

Write a program that reads the first line of each of the two of the smaller files for different clubs. Process both files to compute the name and the words in description of the first and the second club.

Now, using this information print (use set methods to accomplish this):

- The words that are common in the description of the two clubs
- The words that are unique to the first club's description
- The words that are unique to the second club's description

For example, if we compare `wrpi` and `csa` we get (again, order of the words in the sets does not matter):

```
Comparing clubs wrpi and csa:

Same words: {'cultural', 'events'}

Unique to wrpi: {'effective', 'programs', 'music', 'programming',
'includes', 'public', 'days', 'first', 'miles', 'special', 'simulcasts', 'radio',
'located', 'range', 'watts', 'local', 'wide', 'wrpi', 'campus', 'around',
'alternative', 'experimental', 'live', 'sports', 'year', 'studios', 'floor',
'affairs', 'broadcast', 'station', 'bands'}

Unique to csa: {'helps', 'geographical', 'association', 'organization',
'movies', 'pride', 'gatherings', 'chinese', 'presents', 'which', 'include',
'community', 'adjust', 'friendship', 'them', 'various', 'festivals',
'advance', 'group', 'social', 'students', 'life', 'from', 'this', 'members',
'through', 'amongst', 'brings', 'rensselaer', 'welcomes', 'areas',
'culture', 'gathering', 'american', 'aspects'}
```

**To complete Checkpoint 2,** show your code to the TA or a mentor.

**Please come to lab for the last checkpoint.**