Full Name: _____

EEL 4750 / EEE 5502 (Fall 2019) – Code #01 **Due Date:** **Sept. 4, 2019**

Question #1: *(1 pts)* How many hours did you spend on this homework?

Question #2: *(10 pts)* *The Inner Product*

One of the most important operations in signal processing, statistics, and machine learning is the inner product. In signal notation, the inner product between length-N signals $x[n]$ and $y[n]$ is

$$s = \sum_{n=0}^{N-1} x[n]y[n] \ .$$

In a linear algebra notation, the inner product of two length-N vectors is

$$s = \mathbf{x}^T\mathbf{y} \ ,$$

where $\mathbf{x}$ and $\mathbf{y}$ are real-valued (i.e., not complex) vectors. In MATLAB, this is expressed as

```
s = x'*y                    % Compute the inner product of x and y
```

We will be using the inner product throughout the course. In this coding problem, we will use the inner product to create a simple search engine.

Before we do that, let's establish some underlying theory.

(a) Show that when $y[n] = x[n]$, the inner product is the energy of $x[n]$, defined as E_x.

(b) Consider the following two "metrics of similarity"

$$c_1 = \sum_{n=0}^{N-1} x[n]y[n] \qquad , \qquad c_2 = \frac{\sum_{n=0}^{N-1} x[n]y[n]}{\sqrt{\sum_{n=0}^{N-1} |x[n]|^2}\sqrt{\sum_{n=0}^{N-1} |y[n]|^2}} \qquad ,$$

Determine c_1 and c_2 when $y[n] = ax[n]$ and $y[n] = -cx[n]$. Assume a is a real number.

(c) Assume $x[n]$ and $y[n]$ can only contain 1's and 0's across all n. Under this condition, show that c_1 is the count of all locations where 1 is found in both $x[n]$ and $y[n]$.

(d) Consider the signals

$$x[n] = \delta[n] + \delta[n-1]$$
$$y_1[n] = \delta[n] + \delta[n-1] \qquad\qquad y_2[n] = \delta[n-1]$$
$$y_3[n] = \delta[n] + \delta[n-1] + \delta[n-2] \qquad\qquad y_4[n] = \delta[n-1] + \delta[n-3]$$

Compute c_1 and c_2 for $x[n]$ with $y_1[n]$, $y_2[n]$, $y_3[n]$, and $y_4[n]$. (8 values in total)

(e) Describe the advantages and disadvantages for using c_1 or c_2 as a metric of similarity.

Side Note: The value c_2 is often referred to as the *correlation coefficient* between $x[n]$ and $y[n]$. You may know this as the R-value that is often measured for linear regression (i.e., the correlation coefficient between the fit line and the data).

Question #3: *(10 pts) Creating a Search Engine*

In this problem, we will create a simple search engine using the inner product and its properties, discussed in Question # 2. From the downloaded zip file, retrieve the file called `2019_eee5502_code01_q2.mat`. The file contains three variables: a cell `vocabulary`, a cell `documents`, and matrix `counts`.

The cell `vocabulary` is a list of 4436 English words from the given documents. The cell `documents` is a list of 1734 text fragments from old 1980's text-based adventure games. The matrix `counts` has a size of 1734×4436 and contains the frequency of each word across 1734 text fragments.

(a) Write a MATLAB script that uses c_1 and c_2 as metrics of similarity. Specifically, compute the similarity between each row of `counts`, each of which corresponds to a document / text fragment, and a corresponding search term (hint: it should have a very similar form as each row of counts). The document that is the most similar, or best matching, with your search term will maximize c_1 or c_2.

(b) Submit the two matched documents for c_1 and c_2 given the search term:

```
an angry wizard resembles a dragon
```

(ignore words not in `vocabulary`). Also provide the values of c_1 and c_2 for this search.

(c) Submit the two matched documents for c_1 and c_2 given the search terms:

```
the angry wizard resembles a dragon
```

(ignore words not in `vocabulary`). Also provide the values of c_1 and c_2 for this search.

(d) Do (b) and (c) yield different results? Why or why not? Do the results from c_1 or c_2 seem more reliable for your search engine?

(e) The zip file contains the function `get_search_term`. Run `get_search_term` with your UFID as a parameter to retrieve your unique search terms. Submit the matched text fragments corresponding to best c_1 and c_2. Also, provide the corresponding c_1 and c_2 values.