```
"""
Clusters Bob Ross paintings by features.

By Walter Hickey <walter.hickey@fivethirtyeight.com>

See http://fivethirtyeight.com/features/a-statistical-analysis-of-the-work-of-bob-ross/
"""

import numpy as np
from scipy.cluster.vq import vq, kmeans, whiten
import math
import csv

def main():

    # load data into vectors of 1s and 0s for each tag
    with open('elements-by-episode.csv','r') as csvfile:
        reader = csv.reader(csvfile)
        reader.next() # skip header
        data = []
        for row in reader:
            data.append(map(lambda x: int(x), row[2:])) # exclude EPISODE and TITLE columns

    # convert to numpy matrix
    matrix = np.array(data)

    # remove colums that have been tagged less than 5 times
    columns_to_remove = []
    for col in range(np.shape(matrix)[1]):
        if sum(matrix[:,col]) <= 5:
            columns_to_remove.append(col)
    matrix = np.delete(matrix, columns_to_remove, axis=1)

    # normalize according to stddev
    whitened = whiten(matrix)
    output = kmeans(whitened, 10)

    print "episode", "distance", "cluster"

    # determine distance between each of 403 vectors and each centroid, find closest neighbor
    for i, v in enumerate(whitened):

        # distance between centroid 0 and feature vector
        distance = math.sqrt(sum((v - output[0][0]) ** 2))

        # group is the centroid it is closest to so far, set initally to centroid 0
        group = 0
        closest_match = (distance, group)

        # test the vector i against the 10 centroids, find nearest neighbor
        for x in range (0, 10):
            dist_x = math.sqrt(sum((v - output[0][x]) ** 2))
            if dist_x < closest_match[0]:
                closest_match = (dist_x, x)

        print i+1, closest_match[0], closest_match[1]

if __name__ == "__main__":
    main()
```