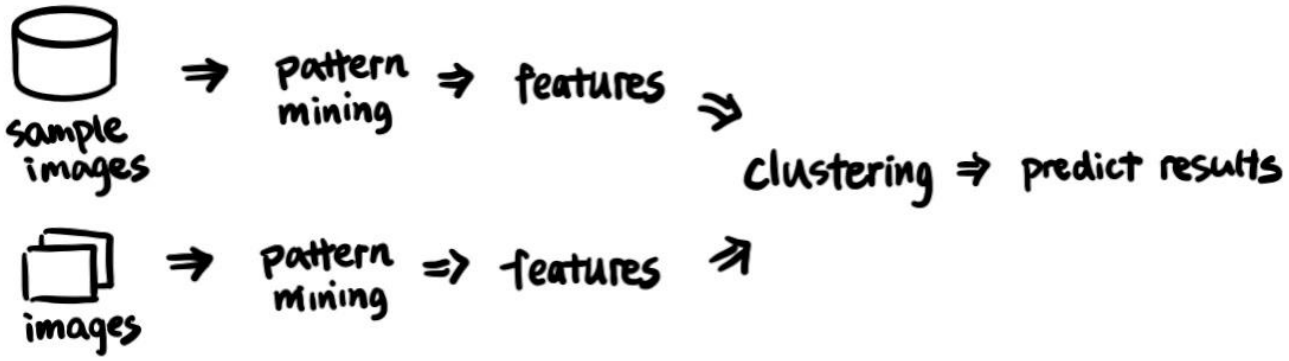
Question 1



We can use pattern mining and clustering to analysis x-ray images of lung disease, using pattern mining to get proper features of the texture information extracted from x-ray images, and apply clustering by inputting the features results to predict their disease types, such as K-Nearest Neighbors which assigns label decided from its K neighbors, voting for the greatest number of labels among these neighbors.

Question 2

1. H_0: the distance and the players are same distribution (independent).

	P.J. Tucker	Brook Lopez	Sum
0-3	166 ($\frac{321}{2000}$ x1000)	155 ($\frac{321}{2000}$ x1000)	321
3-10	101 ($\frac{194}{2000}$ x1000)	93 ($\frac{194}{2000}$ x1000)	194
10-16	14 ($\frac{59}{2000}$ x1000)	45 ($\frac{59}{2000}$ x1000)	59
16-3pt	8 ($\frac{28}{2000}$ x1000)	20 ($\frac{28}{2000}$ x1000)	28
3pt	711 ($\frac{1398}{2000}$ x1000)	687 ($\frac{1398}{2000}$ x1000)	1398
Sum	1000	1000	2000

	P.J. Tucker	Brook Lopez	Sum
0-3	166 (160.5)	155 (160.5)	321
3-10	101 (97)	93 (97)	194
10-16	14 (29.5)	45 (29.5)	59
16-3pt	8 (14)	20 (14)	28
3pt	711 (699)	687 (699)	1398
Sum	1000	1000	2000

$$x^2 = \frac{(166-160.5)^2}{160.5} + \frac{(155-160.5)^2}{160.6} + \frac{(101-97)^2}{97} + \frac{(93-97)^2}{97}$$
$$+ \frac{(14-29.5)^2}{29.5} + \frac{(45-29.5)^2}{29.5} + \frac{(8-14)^2}{14} + \frac{(20-14)^2}{14}$$
$$+ \frac{(711-699)^2}{699} + \frac{(687-699)^2}{699} = 22.54973$$

Using the chi-square test table, check column with the value in the row of $DF = (5-1)(2-1) = 4$, $p < 0.001$, which p value is less than the significance level 0.05. Therefore, the H_0 is false. The two players do not follow the same distribution.

2. Calculate the probability:

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) ln \frac{p(x)}{q(x)}$$
$$= 0.155 * ln \frac{0.155}{0.166} + 0.093 * ln \frac{0.093}{0.101} + 0.045 * ln \frac{0.045}{0.014} + 0.020 * ln \frac{0.020}{0.008} + 0.687 * ln \frac{0.687}{0.711} = 0.02898$$

x	q(x) P.J. Tucker	P(x) Brook Lopez
0-3	0.166	0.155
3-10	0.101	0.093
10-16	0.014	0.045
16-3pt	0.008	0.020
3pt	0.711	0.687
total	1	1

3. KL divergence measures the difference between these two probability distributions, which is 0.02898, that is these two distributions may be similar but are not same, and the Chi-Square Test has the value 22.54973, which also means the two players do not follow the same distribution.
   The Chi-Square Test value is large, which calculate whether the distance and person are correlated, while the KL divergence value is small, which measures the difference between these those two people. These two methods using different parameter.

Question 3

1. The numbers of unique tokens in D1: 4558
   The numbers of unique tokens in D2: 2311

2. Top 100 most frequent tokens in D1:
   {'people': 0.071, 'black': 0.06, 'police': 0.049, 'dont': 0.043, 'amp': 0.038, 'one': 0.038, 'like': 0.038, 'white': 0.037, 'trump': 0.036, 'anonymous': 0.035, 'please': 0.03, 'get': 0.029, 'youranoncentral': 0.027, 'man': 0.026, 'protests': 0.024, 'see': 0.023, 'make': 0.022, 'protest': 0.022, 'antifa': 0.022, 'say': 0.021, 'america': 0.02, 'right': 0.02, 'didnt': 0.02, 'want': 0.019, 'fuck': 0.019, 'know': 0.018, 'need': 0.018, 'realdonaldtrump': 0.018, 'last': 0.018, 'night': 0.017, 'peaceful': 0.017, 'live': 0.016, 'years': 0.016, 'cops': 0.016, 'would': 0.016, 'got': 0.016, 'lives': 0.016, 'back': 0.015, 'fucking': 0.015, 'woman': 0.015, 'never': 0.015, 'help': 0.015, 'twitter': 0.015, 'created': 0.015, 'thats': 0.015, 'yall': 0.014, 'racism': 0.014, 'isnt': 0.014, 'wont': 0.014, 'still': 0.014, 'video': 0.014, 'the\r': 0.014, 'said': 0.014, 'many': 0.014, 'stop': 0.014, 'support': 0.014, 'another': 0.013, 'time': 0.013, 'protesting': 0.013, 'cant': 0.013, 'let': 0.013, 'money': 0.013, 'going': 0.012, 'every': 0.012, 'really': 0.012, 'government': 0.012, 'understand': 0.012, 'youre': 0.012, 'looting': 0.012, 'always': 0.012, 'way': 0.012, 'system': 0.012, 'saying': 0.012, 'protesters': 0.012, 'group': 0.012, 'think': 0.011, 'house': 0.011, 'come': 0.011, 'love': 0.011, 'today': 0.011, 'first': 0.011, 'george': 0.011, 'arrested': 0.011, 'violence': 0.011, 'criminals': 0.011, 'much': 0.01, 'everyone': 0.01, 'it\r': 0.01, 'protestors': 0.01, 'issue': 0.01, 'tear': 0.01, 'girl': 0.01, 'day': 0.01, 'stand': 0.01, 'hope': 0.01, 'around': 0.01, 'family': 0.01, 'create': 0.01, 'impunity': 0.01, 'allows': 0.01}
   Top 100 most frequent tokens in D2:
   {'covid19': 0.729, 'pandemic': 0.26, 'testing': 0.246, 'protesters': 0.242, 'centers': 0.239, 'closing': 0.238, 'punish': 0.238, 'theyre': 0.236, 'kissychalamets': 0.233, 'weaponizing': 0.233, 'attempt': 0.233, 'sil\r': 0.233, 'healthcare': 0.111, 'providers': 0.095, 'many': 0.087, 'protest': 0.084, 'care': 0.084, 'need': 0.082, 'telling': 0.078, 'insurance': 0.078, 'plans': 0.078, 'tipptipphorray': 0.077, 'recommend': 0.077, 'deny': 0.077, 'get': 0.068, 'like': 0.053, 'two': 0.051, 'country': 0.044, 'new': 0.04, 'love': 0.038, 'blood': 0.035, 'still': 0.034, 'one': 0.034, 'charge': 0.034, 'cases': 0.034, 'day': 0.034, 'thought': 0.033, 'would': 0.033, 'away': 0.031, 'anyone': 0.031, 'gone': 0.03, 'tested': 0.03, 'amp': 0.029, 'hasnt': 0.029, 'protested': 0.029, 'police': 0.029, 'bryanna711': 0.028, 'downtown': 0.028, 'mom': 0.028, 'testin\r': 0.028, 'elemental': 0.028, 'folks': 0.028, 'people': 0.027, 'first': 0.027, 'bill': 0.027, 'thats': 0.027, 'fordm': 0.026, 'reading': 0.026, 'complying': 0.026, 'replies': 0.026, 'stayatho\r': 0.026, 'gotten': 0.024, 'patients':

0.023, 'may': 0.022, 'cnn': 0.022, 'reports': 0.022, 'city': 0.021, 'president': 0.021, 'april': 0.021, 'clots': 0.021, 'symptom': 0.021, 'began': 0.021, 'got': 0.02, 'deaths': 0.02, 'jheil': 0.02, 'tinderbox': 0.02, 'dont': 0.02, 'start': 0.02, 'around': 0.02, 'emerged': 0.02, 'mysterious': 0.02, 'coronavirussoon': 0.02, 'hearing': 0.02, 'pati\r': 0.02, 'want': 0.019, 'period': 0.019, 'seeing': 0.019, 'protests': 0.018, 'today': 0.018, 'notgeauxgabby': 0.018, 'protestors': 0.018, '214': 0.018, 'incubation': 0.018, 'signs': 0.018, 'illness': 0.018, 'keeps': 0.018, 'leading': 0.018, '103yearold': 0.018, 'woman': 0.018, 'celebrates': 0.018}

3. I use the WordCloud package in python and create these two word-cloud figure by calling WordCloud() function. I set the background color to white and use function generate_from_frequencies() to show words with frequencies. The larger font size of a word represents a higher frequency.

   The word-cloud figure for D1:

   The word-cloud figure for D2:

4. I separate two components using photoshop and create word clouds in each component and then put them together. The two sets of keywords are in two different color set, the words in D1 is set with ‘Purples’ color map and the words in D2 is set with ‘Greens’