Q1 (1 point)

a) (1/2 point)

- Flat clustering:
  - high efficiency (1/8 point)
  - need to decide a K before performing the cluster (1/8 point)

- Hierarchy clustering:
  - generates deterministic results and hierarchical structure (1/8 point)
  - K is unknown or hard to decide (1/8 point)

----

b) (1/2 point)

if they write at least 2 out of 4, they get the whole point

- K-means: (1/4 point)
  - Value of K, or the number of clusters
  - The number of iterations and stopping criteria.
  - Outliers.
  - Initialization and size of clusters.

- Hierarchical Agglomerative clustering factors: (1/4 point)
  - Metric for measuring similarity.
  - Clustering method used like complete-link, single-link, centroid, group average etc.
  - Outliers

Q2-1 (2 points)

(Raw/normalized term frequency vector) (1 points)

	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6
carp	$1/\sqrt{3} = 0.57$	0	0	0	0	0
dolphins	0	0	0	0	$1/2 = 0.5$	0
elephant	0	0	0	$1/\sqrt{3} = 0.57$	0	0
horse	0	$1/2 = 0.5$	0	0	0	0
land	0	$1/2 = 0.5$	$1/\sqrt{3} = 0.57$	0	0	0
lion	0	0	$1/\sqrt{3} = 0.57$	0	0	0
lung	0	$1/2 = 0.5$	$1/\sqrt{3} = 0.57$	$1/\sqrt{3} = 0.57$	$1/2 = 0.5$	0
neck	0	$1/2 = 0.5$	0	0	0	$1/2 = 0.5$
seahorse	0	0	0	0	0	$1/2 = 0.5$
snout	0	0	0	$1/\sqrt{3} = 0.57$	0	0
swim	$1/\sqrt{3} = 0.57$	0	0	0	$1/2 = 0.5$	$1/2 = 0.5$
water	$1/\sqrt{3} = 0.57$	0	0	0	$1/2 = 0.5$	$1/2 = 0.5$

(one iteration of K-means)

Euclidian distance from doc1 and doc2. (1/4 point)

doc1 and doc3: $\sqrt{2} = 1.41$	doc2 and doc3: 0.91
doc1 and doc4: $\sqrt{2} = 1.41$	doc2 and doc4: 1.19
doc1 and doc5: 0.91	doc2 and doc5: 1.22
doc1 and doc6: 0.91	doc2 and doc6: 1.22

Cluster Assignment: (1/4 point)

Doc1: Doc1, Doc2: Doc2, Doc3: Doc2, Doc4: Doc2, Doc5: Doc1, Doc6: Doc1

they could also have written it as:

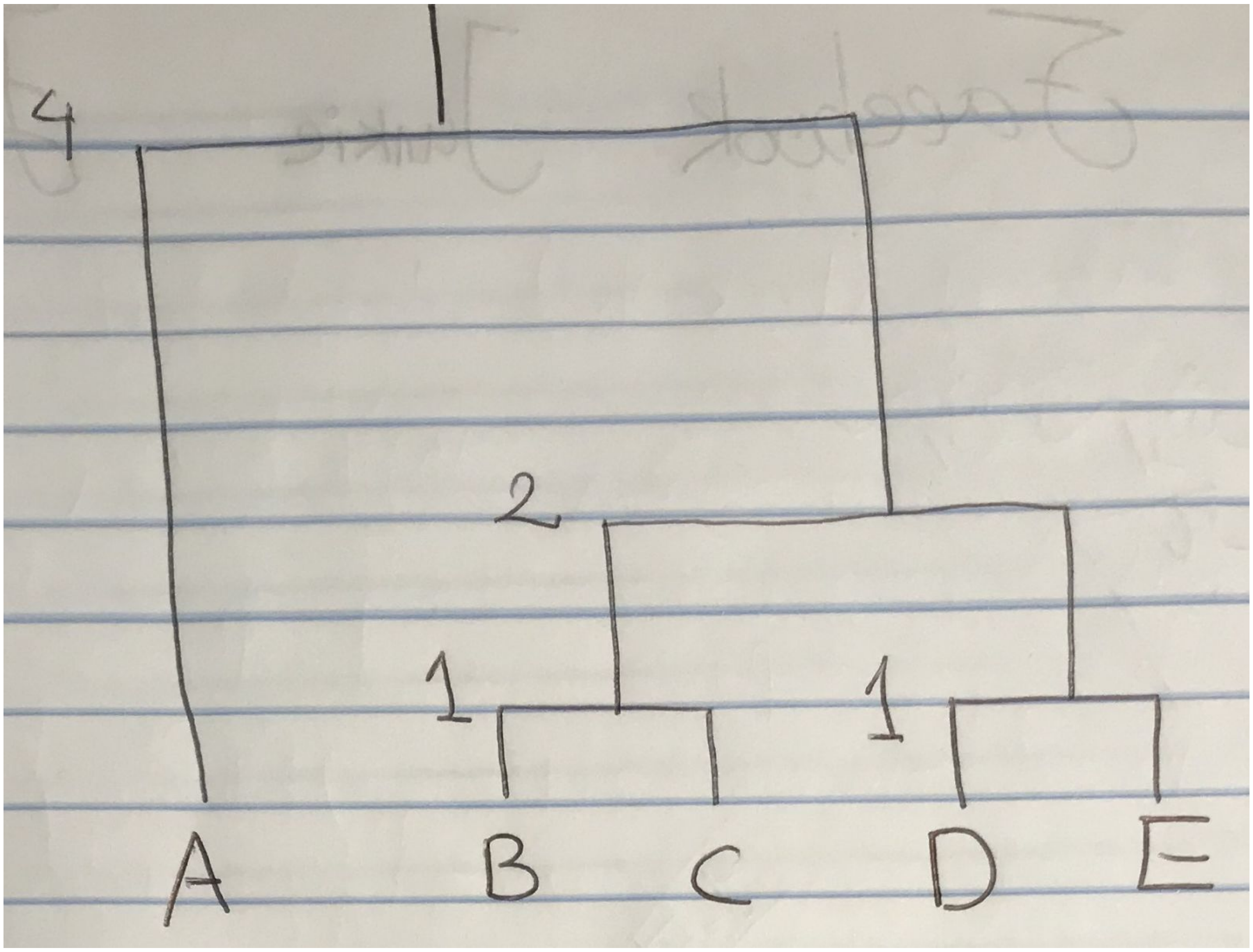
Cluster 1: Doc1, Doc5, Doc6
Cluster 2: Doc2, Doc3, Doc4

New Clusters: (1/2 point)

Cluster 1: Doc1, Doc5, Doc6
[0.19, 0.16, 0, 0, 0, 0, 0.16, 0.16, 0.16, 0, 0.52, 0.52]
Cluster 2: Doc2, Doc3, Doc4
[0, 0, 0.19, 0.16, 0.36, 0.19, 0.55, 0.16, 0, 0.19, 0, 0]

Q3 (2 points)

a) (1 point)



b) (1 point)