1. Short Answer

   a. Flat clustering should be used when efficiency is needed in the clustering algorithm. Flat clustering also requires a defined number of documents, N, and a defined number of clusters, k. If a consistent algorithm is required, hierarchal clustering produces the most deterministic results. Its is also possible to use hierarchal clustering without knowing K, the number of clusters to form. This is because the algorithm produces a hierarchy of soft clusters, with documents belonging to parent, grandparent, etc. clusters.

   b.
      i. **K -means**
      ii. Initial seed selection upon beginning K-means
      iii. Number of clusters to form, K
      iv. Number of documents, N
      v. Number of iterations of reassignment and recomputation
      vi. Distance between document vectors and cluster centroids
      vii. Final RSS value (want to minimize)

      viii. **HAC**
      ix. Number of documents, N
      x. Which similarity algorithm is used (single link, complete link, centroid, group average)

2. Problem Solving

   a. doc_raw_tf = {term1: f, term2: f, etc}

      doc1_raw_tf = {carp: 1, swim: 1, water: 1}
      doc2_raw_tf = {horse: 1, neck: 1, land: 1, lung: 1}
      doc3_raw_tf = {lion: 1, lung: 1, land: 1}
      doc4_raw_tf = {elephant: 1, lung: 1, snout: 1}
      doc5_raw_tf = {dolphins: 1, swim: 1, water: 1, lung: 1}
      doc6_raw_tf = {seahorse: 1, swim: 1, water: 1, neck: 1}

      doc1_norm = {carp: 1/√3, swim: 1/√3, water: 1/√3}
      doc2_ norm = {horse: 1/2, neck: 1/2, land: 1/2, lung: 1/2}
      doc3_ norm = {lion: 1/√3, lung: 1/√3, land: 1/√3}
      doc4_ norm = {elephant: 1/√3, lung: 1/√3, snout: 1/√3}
      doc5_ norm = {dolphins: 1/2, swim: 1/2, water: 1/2, lung: 1/2}
      doc6_ norm = {seahorse: 1/2, swim: 1/2, water: 1/2, neck: 1/2}

Reassignment:

Distance doc1 -> doc3 = sqrt($(1/\sqrt{3} – 0)^2 + (1/\sqrt{3} – 0)^2 + (1/\sqrt{3} – 0)^2 + (0 - 1/\sqrt{3})^2 + (0 - 1/\sqrt{3})^2 + (0 - 1/\sqrt{3})^2$) = 1.41
Distance doc2 -> doc3 = sqrt($(1/2 – 0)^2 + (1/2 – 0)^2 + (1/2 - 1/\sqrt{3})^2 + (1/2 - 1/\sqrt{3})^2 + (0 - 1/\sqrt{3})^2$) = .919
distance doc1 > distance doc2. Doc3 will be clustered with doc2 centroid.

Distance doc1 -> doc4 = sqrt($(1/\sqrt{3} – 0)^2 + (1/\sqrt{3} – 0)^2 + (1/\sqrt{3} – 0)^2 + (0 - 1/\sqrt{3})^2 + (0 - 1/\sqrt{3})^2 + (0 - 1/\sqrt{3})^2$) = 1.41
Distance doc2 -> doc4 = sqrt($(1/2 – 0)^2 + (1/2 – 0)^2 + (1/2 – 0)^2 + (1/2 - 1/\sqrt{3})^2 + (0 - 1/\sqrt{3})^2 + (0 - 1/\sqrt{3})^2$) = 1.19
distance doc1 > distance doc2. Doc4 will be clustered with doc2 centroid.

Distance doc1 -> doc5 = sqrt($(1/\sqrt{3} – 0)^2 + (1/\sqrt{3} – 1/2)^2 + (1/\sqrt{3} – 1/2)^2 + (0 – 1/2)^2 + (0 – 1/2)^2$) = 1.19
Distance doc2 -> doc5 = sqrt($(1/2 – 0)^2 + (1/2 – 0)^2 + (1/2 – 0)^2 + (1/2 – 1/2)^2 + (0 - 1/2)^2 + (0 - 1/2)^2 + (0 - 1/2)^2$) = 1.22
Distance doc1 < distance doc2. Doc5 will be clustered with doc1 centroid.

Distance doc1 -> doc6 = sqrt($(1/\sqrt{3} – 0)^2 + (1/\sqrt{3} – 1/2)^2 + (1/\sqrt{3} – 1/2)^2 + (0 - 1/2)^2 + (0 - 1/2)^2$) = .919
Distance doc2 -> doc6 = sqrt($(1/2 – 0)^2 + (1/2 – 1/2)^2 + (1/2 – 0)^2 + (1/2 – 0)^2 + (0 - 1/2)^2 + (0 - 1/2)^2 + (0 - 1/2)^2$) = 1.22
distance doc1 < distance doc2. Doc6 will be clustered with doc1 centroid.

doc1 centroid -> (doc5, doc6)
doc2 centroid -> (doc3, doc4)

Recomputation:

For each term in cluster, sum norm. term values.
Then divide the value by num docs in the cluster.

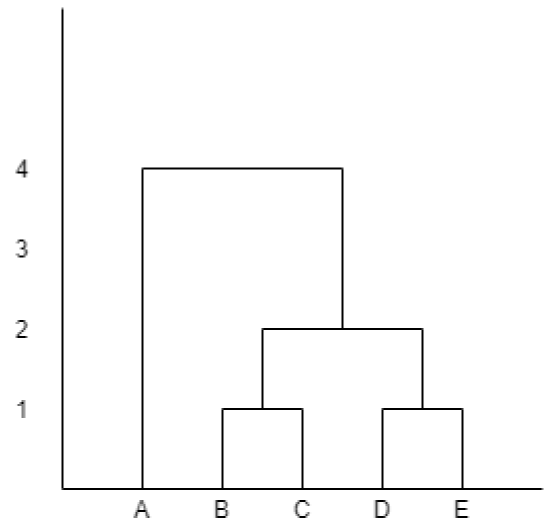
doc1 cluster = {carp: $(1/\sqrt{3} + 0 + 0)$, swim: $(1/2 + 1/2 + 1/\sqrt{3})$, water: $(1/2 + 1/2 + 1/\sqrt{3})$, dolphins: $(1/2 + 0 + 0)$, lung: $(1/2 + 0 + 0)$, seahorse: $(1/2 + 0 + 0)$, neck: $(1/2 + 0 + 0)$}
doc2 cluster = {horse: $(1/2 + 0 + 0)$, neck: $(1/2 + 0 + 0)$, land: $(1/2 + 1/\sqrt{3} + 0)$, lung: $(1/2 + 1/\sqrt{3} + 1/\sqrt{3})$, lion: $(1/\sqrt{3} + 0 + 0)$, elephant: $(1/\sqrt{3} + 0 + 0)$, snout: $(1/\sqrt{3} + 0 + 0)$}

Divide all vals by 3 because there are 3 documents in each cluster.

New centroid cluster 1= {carp: .192, swim: .525, water: .525, dolphins: .166, lung: .166, seahorse: .166, neck: .166}
New centroid cluster 2 = {horse: .166, neck: .166, land: .359, lung: .551, lion: .192, elephant: .192, snout: .192}

3. Problem Solving

a.



b.