(520|600).666 Information Extraction

Homework # 1

Due February 6, 2024.

1. Read Chapter 1 from the Jelinek book.

2. Write a one page summary (including any figures if you wish) of the article

   S. Young, "A Review of Large Vocabulary Continuous Speech Recognition," *IEEE Signal Processing Magazine*, pp 45-57, Sept 1996.

3. *Computer Exercise in Vector Quantization.* You will be given 100 2-dimensional points, $\{\mathbf{a}_i = (x_i, y_i),\ i = 1, 2, \ldots, 100\}$. You are to divide them into 3 sets using vector quantization based on Euclidean distance $d(\mathbf{a}_i, \mathbf{a}_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

   (a) Choose the three initial cluster-centers, ρ_k, $k = 1, 2, 3$, uniformly at random from the 1×1 square in which the 100 points are located.

   (b) Carry out the quantization process (cf Chapter 1) until no points change set membership.

   (c) Using 3 different colors, plot the resulting sets and their cluster-centers.

   Repeat the exercise several times, each with a different random choice of initial cluster-centers. Observe and report (i) the common *tendencies* and (ii) the occasional *outlier* behavior of the clustering algorithm.

4. *Product Quantization*: In class we discussed another quantization method called product quantization.

   (a) Describe the space and assignment complexity of K-Means vs. product quantization.

   (b) Using the above result, explain why this is useful.