

050/520/600.666 Information Extraction from Speech and Text

Homework # 1

Due February 10, 2006.

1. Write a two 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.

2. *Computer Exercise in Vector Quantization*. You will be given 100 2-dimensional points, $\{\mathbf{a}_i = (x_i, y_i), i = 1, 2, \dots, 100\}$. 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, $\rho_k, k = 1, 2, 3$, 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 for a different random choice of the initial cluster-centers. Compare in a few words the sets obtained in the two trials.