1. For this question, it is recommended that you use the GRM library and FSM or OpenFst libraries. In fact, try as much as possible to use the utilities of these libraries to answer the questions. However, you need to justify your responses and not just mention the library utilities used.

(a) Download the following training corpus $S$ and test corpus $\hat{S}$:
http://www.cs.nyu.edu/~mohri/asr10/train.txt

(b) Extract the vocabulary $\Sigma_1$ of $S$ and define a start and end symbol.

(c) Create the following language models:
- bigram back-off model;
- trigram back-off model;

Report for each of the weighted automata obtained
- the number of states;
- the number of transitions;
- the number of $\epsilon$-transitions;
- the number of $n$-grams found ($n = 2$ for bigram models, $n = 3$ for trigram models).

For these questions, you can use the utility `fsminfo` of the FSM library. You should however explain how you determine the number of $n$-grams.

(d) Randomly generate 100 sequences from the first model and compare the likelihood given by the two models to the sample formed by these sentences.

(e) Compute the perplexity of these models using the test corpus.

(f) Shrink both of these models with the option $-s4$. What are the perplexity estimates for these models.

2. Expectation-maximization: a computer disk manufacturer wishes to come up with a model of the time to failure of the disks it is designing. The probability of a disk failure at time $t \geq 0$ is assumed to follow an exponential model: $p_\theta[f(t)] = \frac{1}{\theta} \exp(-\frac{t}{\theta})$, where $\theta > 0$ is a parameter.
(a) Show that $\theta$ is the expected failure time.

To estimate the parameter $\theta$, the company selects $n_1$ disks $D_1, \ldots, D_{n_1}$ at random, monitors them, and determines the exact failure time of each: $t_1, \ldots, t_{n_1}$. It also selects $n_2$ disks $D_{n_1+1}, \ldots, D_{n_1+n_2}$. It does not observe the failure times $t_{n_1+1}, \ldots, t_{n_1+n_2}$ for these disks. But, it inspects them all at some time $t_0$ and registers their state: if disk $D_{n_1+i}$, $i \in [1, n_2]$, was still functioning, $s_{n_1+i} = 0$, otherwise $s_{n_1+i} = 1$.

(b) Give the expression of the full log-likelihood of the data, as a function of $t_1, \ldots, t_{n_1+n_2}$.

(c) Determine the probability of each unobserved variable $t_{n_1+i}$, $i \in [1, n_2]$, conditioned on the values of the observed variables $s_{n_1+i}$, for a given $\theta$.

(d) Derive an EM algorithm for estimating $\theta$: give its value at the $(k+1)$st iteration, $\theta_{k+1}$, as a function of its value at the $k$th iteration, $\theta_k$. It will be useful to use the notation $\overline{s} = \frac{1}{n_2} \sum_{k=1}^{n_2} s_{n_1+k}$ and $\overline{t} = \frac{1}{n_1} \sum_{k=1}^{n_1} t_k$.

(e) Choose 5 different initial values $\theta_0$ and for each report the value of $\theta$ obtained using this algorithm for $n_1 = 1000$, $n_2 = 500$, $t_0 = 50$, $\overline{t} = 75$, and $\overline{s} = .25$.

(f) Assume that $s_{n_1+i} = 0$ for all $i \in [1, n_2]$. Give a closed form expression of $\theta$ in that case.