UNIT 2

LAB 1

PROBLEM STATEMENT

Build the Unigram, Bigram and Trigram Language Models for the given corpus.

Deliverables: Three CSV files (Unigram.csv, Bigram.csv and Trigram.csv).

Format of the CSV files:

Unigram.csv should have one column for listing words, Word, along with Count, q_ML and log q_ML.

Bigram.csv should have two columns for listing words, Word1 and Word2 along with Count, q_ML and log q_ML.

Trigram.csv should have three columns for listing words, Word1, Word2 and Word3 along with Count, q_ML and log q_ML.

WORD1	WORD2	WORD 3	COUNT	q_ML	log q_ML

Submit to: course.nlp.2014@gmail.com.