Building a Language Model

In this task, you will be building a language model from scratch, using the principles learnt in class. The task is as follows:

1. Select a corpus (MLRS - Maltese or BNC - English), carry out any necessary preprocessing and build a lexicon that covers the full corpus. Print the lexicon to a text file for easy reference.
2. Split the corpus into the different sets – training, test – the split should be *random*.
3. Build your unigram, bigram and trigram counts & models – the first version will be without any smoothing – we will call this the Vanilla version.
4. Next, build unigram, bigram and trigram using Laplace Smoothing, this will be the Laplace version.
5. Next, take all words in the training corpus that have a count of 1, and change them into the <UNK> token. Recalculate unigram, bigram and trigram counts and models. This version will be called the UNK version.
6. Finally build a function that can be passed the flavour of the model that we would like to use. This function will take that flavour and apply linear interpolation fixed as follows: trigram = 0.6; bigram = 0.3; unigram = 0.1.

Testing the models:

1. Given a sequence of words (1 or more), generate the rest of the sentence/sequence.
2. Given a sequence of words (1 or more), calculate its probability.

Testing parameters:
I should be able to choose either the individual models, e.g. laplace bigram, or else simply the flavor with linear interpolation.

Submission Requirements:

1. Code & Documentation – be concise – I don't want an explanation of the class notes. I want an explanation of your implementation choices.
2. Demo (arranged after submission)

Marking Scheme:
Preprocessing: 5 marks
Vanilla Version: 10 marks
Laplace Version: 10 marks
UNK Version: 10 marks
Interpolation: 10 marks
Generation Test: 10 marks
Probability Test: 10 marks
Documentation (justification of choices, etc.): 20 marks
Demo: 15 marks

Submission Date:
22nd March midnight

Important: This is an individual task, plagiarism will not be tolerated. Reference all sources used appropriately. Delayed submissions will result in a 10% deduction per day.

Demos will probably be held on 1st and 2nd April.