Building a Language Model

The first project will be related to building a language model from scratch. In the first instance, the first thing that is required is that you prepare a corpus (choose only one) and carry out the necessary pre-processing.

Maltese Corpus: http://mlrs.research.um.edu.mt/index.php?page=downloads
(Baby) British National Corpus: http://ota.ox.ac.uk/desc/2553

First set of tasks:

1. Extract the selected corpus
2. Build frequency counts for n-grams

Things to keep in mind:

- Try to make your code modular. Of course, initially hard coding stuff is ok. But ideally, evolve your hard-coding to generic, modular code
- You might also want to hard code a specific input string so that you can ensure that your frequency counts are correct.
- Consider computational issues – check how long it takes to read the corpus, build frequency counts, how well it scales up, amount of RAM that the system ends up using, etc.
- When reading the files, you might want to consider using one file initially. But already start investigating using all files in the corpus. Remember the larger your corpus, the better the models will be.
- Think ahead – Once you get the frequency counts sorted, remember that like in any machine learning set up, we will be needing a training set and a test set. So you will eventually need to split your corpus into two distinct parts. More about this at a later stage, but it is good that you start thinking about this eventuality.