Laboratório 4 | Parte 3

Nesta parte do laboratório foi-nos pedido para realizar um modelo de linguagem estatística. Decidiu-se então utilizar a linguagem de programação *Python* para resolver as questões propostas. Foram então criados 3 scripts.py que devem ser executados da seguinte ordem e da seguinte maneira.

1. python **ngrams.py** <input training file>
2. python **prob.py** <input test file>
3. python **prob_smoothing.py** <input test file>

Cada script de *Python* resolve a alínea correspondente. No entanto embora estes scripts estejam feitos para resolver qualquer ficheiro de entrada inicial o ficheiro **prob.py** e **prob_smoothing.py** têm em conta que foi executado o ficheiro **s1.txt** dado pela professora. Tendo por isso o script inicial, **ngrams.py**, gerado os ficheiros com o nome expectável para que o segundo e terceiro script funcionem naturalmente.

Todo o código encontra-se comentado e bem estruturado para que seja fácil a leitura e debug do mesmo.

A frase de teste o grupo escolheu foi "*<s> o rapaz foi ótimo no estúdio em abril admitiu a necessidade de aprender um monte de novos termos amorosos para experimentar com as jennifers de sampa </s>*", que embora produza uma probabilidade final muito reduzida, para ambos os modos sem smoothing como o modo add-one smoothing, esta nunca atinge o valor 0.

Apresenta-se abaixo, em suma alguns dos valores importantes desta parte, que não tornam dispensável a consulta dos scripts.

Tamanho do Vocabulário	35938
Número de Unigramas	430091
Número de Bigramas	405016

Tabela 1 – Valores obtidos na alínea 1

	Probabilidades Finais	
	Without Smoothing	**Add-One Smoothing**
s2.txt	4.79217097522e-10	5.16225623622e-19
s3.txt	3.77648268544e-51	4.69582884989e-105

Tabela 2 - Resultados finais das probabilidades pedidas