**UFRPE - Pró-Reitoria de Pesquisa e Pós-Graduação**
**Programa de Pós-Graduação em Informática Aplicada (PPGIA)**


**Disciplina: MT – Prof. Rinaldo Lima**
**N-Grams**
**Lista de Exercícios (Individual)**
Data de Entrega: 18/04/2018 - Enviar ao email: rjlima01@gmail.com

## 2ª. LISTA DE EXERCÍCIOS

1. Dado o seguinte techo de um corpus:

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> I am Sam </s>
<s> I do not like green eggs and Sam </s>
```

Use o modelo de bigram para calcular as seguintes probabilidades:
   a. P(Sam|am)
   b. P(Sam|and)
   c. P(am|I)
   d. P(do|I)

2. Considerando as seguintes tabelas de probabilidades de bigramas, calcule a probabilidade da frase *P(I want chinese food)* em cada caso.

a.

| | i | want | to | eat | chinese | food | lunch | spend |
|---|---|---|---|---|---|---|---|---|
| **i** | 0.002 | 0.33 | 0 | 0.0036 | 0 | 0 | 0 | 0.00079 |
| **want** | 0.0022 | 0 | 0.66 | 0.0011 | 0.0065 | 0.0065 | 0.0054 | 0.0011 |
| **to** | 0.00083 | 0 | 0.0017 | 0.28 | 0.00083 | 0 | 0.0025 | 0.087 |
| **eat** | 0 | 0 | 0.0027 | 0 | 0.021 | 0.0027 | 0.056 | 0 |
| **chinese** | 0.0063 | 0 | 0 | 0 | 0 | 0.52 | 0.0063 | 0 |
| **food** | 0.014 | 0 | 0.014 | 0 | 0.00092 | 0.0037 | 0 | 0 |
| **lunch** | 0.0059 | 0 | 0 | 0 | 0 | 0.0029 | 0 | 0 |
| **spend** | 0.0036 | 0 | 0.0036 | 0 | 0 | 0 | 0 | 0 |

b.

| | i | want | to | eat | chinese | food | lunch | spend |
|---|---|---|---|---|---|---|---|---|
| i | 0.0015 | 0.21 | 0.00025 | 0.0025 | 0.00025 | 0.00025 | 0.00025 | 0.00075 |
| want | 0.0013 | 0.00042 | 0.26 | 0.00084 | 0.0029 | 0.0029 | 0.0025 | 0.00084 |
| to | 0.00078 | 0.00026 | 0.0013 | 0.18 | 0.00078 | 0.00026 | 0.0018 | 0.055 |
| eat | 0.00046 | 0.00046 | 0.0014 | 0.00046 | 0.0078 | 0.0014 | 0.02 | 0.00046 |
| chinese | 0.0012 | 0.00062 | 0.00062 | 0.00062 | 0.00062 | 0.052 | 0.0012 | 0.00062 |
| food | 0.0063 | 0.00039 | 0.0063 | 0.00039 | 0.00079 | 0.002 | 0.00039 | 0.00039 |
| lunch | 0.0017 | 0.00056 | 0.00056 | 0.00056 | 0.00056 | 0.0011 | 0.00056 | 0.00056 |
| spend | 0.0012 | 0.00058 | 0.0012 | 0.00058 | 0.00058 | 0.00058 | 0.00058 | 0.00058 |

c. Explique o motivo da diferença dos valores das probabilidades em (a) e (b)

3. Considere o seguinte extrato de um corpus:

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> Sam I like </s>
<s> Sam I do like </s>
<s> do I like Sam </s>
```

Assumindo o modelo de *bigrama*, qual é a mais provável palavra a ser predita logo em seguida em cada caso abaixo?

a. <s> Sam ???
b. <s> Sam I do ???
c. <s> I am Sam ???
d. <s> Do I like ???

Qual das seguintes frases obtém a maior probabilidade usando o mesmo modelo acima?

e. <s> Sam I do like </s>
f. <s> Sam I am </s>
g. <s> I do like Sam I am </s>

4. Usando o *Brown Corpus* disponível no NLTK, calcule a frequência de todos os unigramas e exiba os 50 primeiros mais frequentes. Considere apenas o vocabulário, isto é, *word type*.
   a. Quais as classes gramaticais destas 50 palavras?
   b. Gere um histograma para a distribuição acima e discuta seu resultado.

5. Usando sentence splitting e tokenization para *português* implemente um programa que receba como entrada o corpus de notícias **NoticiasPortugues.zip** e realize:

   a. a divisão de sentenças
   b. converta todas as palavras para minúscula
   c. tokenize as frases
   d. calcule a probabilidade de todos os unigrams presentes neste corpus, $P(w_i)$. Considere como unidade de unigrama, todos os tokens distintos (types ou vocabulário)
   e. Implemente uma função que dada uma frase, contendo *N* tokens, seja retornada a probabilidade desta frase usando o modelo de unigrama.

      Ex.: P("eu li uma má notícia ontem") = P (Eu, li, uma, má, nóticia, ontem)
      = **P(eu) * P(li) * P(má) * P(notícia) * P(ontem)**

   OBS.: Para o cálculo das probabilides, use o artificio que transforma a multiplicação de probabilidades numa soma de logarítimos.

Questão desafio (opcional)

6. Implemente uma versão baseada em bi-gramas para o exercício (5) acima.

---

**Material extra de apoio para os exercícios**

Ver os links no arquivo de **Links.txt**.