**Mapa de tarefas**

Fora da programação:

**O1**- Gerar Corpus (treinamento + teste), gerar Corpus de teste de assunto

**O2**- Construir a Matriz de Confusão

**O3**- (Possivelmente) Construir um Spell Error para o vocabulário do nosso corpus (parte treinamento).

**O4**- Gerar nuvens de palavras para conjuntos de testes e treinamento

Na programação:

**I1**- Tokenizar o corpus (parte treinamento) - Vocabulário

**I2**- Contar ocorrências de cada palavra e dupla de palavras (array e matriz de ocorrência) – Relacionada a **I1**

**I3**- Computar probabilidades de ocorrência única e dupla (nesse segundo aplicando as duas suavizações exigidas add-1 e kneser-ney) (um array e duas matrizes de probabilidade – Unigram e Bigram)

**I4**- Calcular Perplexidade (teste)

FIM de Primeira e Segunda Partes da Atividade – Tarefa de fora dependente: **O1**

CORRETOR ORTOGRÁFICO

- Se palavra teste não **I5** PERTENCE ao Vocabulário:

- Elenque os Candidatos com a menor **I6** DISTÂNCIA de erro

**I7**- Calcular P(X|W).P(W) para cada Candidato – Usando a Matriz de Confusão para calcular P(X|W) (Proposta de adaptação: adicionar mais uma linha e coluna na Matriz de Confusão para o “carácter vazio” ilustrando as operações inserção e remoção – facilita o cálculo de P(X|W) sendo único para cada operação – diferente do expresso no slide, confesso não ter entendido direito:

P(X|W) = produtório para cada substituição da FREQUENCIA DA SUBSTITUIÇÃO / somatório das FRENQUENCIAS DA LINHA CORRESPONDENTE À LETRA CORRETA – valores contidos na matriz confusão)

- Seleciona a palavra de maior P(X|W).P(W)

TESTE DO CORRETOR

**I8**- Carrega o arquivos com as palavras erradas e suas correspondentes corretas

- Aplica CORRETOR ORTOGRÁFICO

- Conta ERROS e ACERTOS para computar a acurácia, que nesse caso será: ACERTOS / ERROS + ACERTOS.