ISCTE IUL **Instituto Universitário de Lisboa** Escola de Tecnologias e Arquitectura Dep. de Ciências e Tecnologias da Informação	PCL Ano letivo 2014/2015 1º Semestre **Trabalho 2** Versão 1.0

Trabalho **individual** ou em **grupo** (2 alunos) Prazo de entrega: **27 de Outubro de 2014**
Este trabalho deve ser submetido através de um formulário eletrónico, tal como o anterior (mais informações abaixo)

Construção de modelos de língua estatísticos – utilização prática

Dada uma palavra ou uma frase, pretende-se identificar automaticamente a língua em que está escrita. A identificação da língua é uma tarefa cujo sucesso depende da quantidade de dados utilizada para treinar o sistema. Neste trabalho deve usar sequências fixas de carateres, também conhecidas como n-gramas de carateres. Por exemplo, se considerarmos a frase "*sim senhor*" (10 carateres) e um modelo de Markov de 2ª ordem para sequências de letras, a probabilidade da frase ser de uma dada língua pode ser dada pela fórmula:

$$p(sim\ senhor) = p(s|<<) \times p(i|<s) \times p(m|im) \times p(\ |im) \times \ldots \times p(r|ho) \times p(>|or)$$

em que < e > são os símbolos de início e fim de frase, respetivamente. De forma evitar problemas de *underflow* e *overflow* deve utilizar logaritmos de probabilidades. Nesse caso, em vez da fórmula anterior, deverá usar a fórmula seguinte.

$$logprob(sim\ senhor) = log(p(s|<<)) + log(p(i|<s)) + \ldots + log(p(>|or))$$

Note que, como o logaritmo é uma função monotónica crescente (para bases superiores a 1, é claro), a língua mais provável continuará a ser a que maximizar o resultado da fórmula. Note também que a probabilidade $p(z|xy)$ para uma dada língua pode ser obtida com base em contagens de n-gramas a partir dos dados de treino dessa língua, de acordo com a fórmula seguinte:

$$p(z|xy) = \frac{contagem(xyz)}{contagem(xy)}$$

Por exemplo, se considerar a frase "*sim senhor*" teremos as seguintes contagens de bigramas e trigramas. Note que num modelo de trigramas temos de observar sempre os dois símbolos anteriores, por isso se consideram dois símbolos iniciais fictícios: <<; do mesmo modo é útil considerar o fim de sequência >.

bigrama	contagem
<s	1
si	1
im	1
m	1
␣s	1
se	1
en	1
nh	1
ho	1
or	1
r>	1

trigrama	contagem
<<s	1
<si	1
sim	1
im␣	1
m␣s	1
␣se	1
sen	1
enh	1
nho	1
hor	1
or>	1

Juntamente com o enunciado é fornecido um texto pertencente à lingua portuguesa, para o qual foram calculados os respetivos bigramas e trigramas. Este material encontra-se disponível no ficheiro `dados.zip`.

Tarefas a realizar

1. Construa um *corpus*, referente a cinco línguas.

   (a) Recolha até 100 frases reais em cada uma das línguas. Se pretender incluir o PT pode usar os dados fornecidos no material de apoio referentes à língua portuguesa.

   (b) Processe estes dados de forma a obter textos *normalizados*: separe as palavras das marcas de pontuação (pontos finais, vírgulas, pontos de exclamação, pontos de interrogação, aspas, ...); apague todos os símbolos usados para indicar início e fim de frase, "<" e ">" respetivamente.

   Observação. Para facilitar, não se preocupe com a conversão de numerais. Os comandos unix `sed`, `tr`, `grep`, `awk`, etc. facilitam muito esta tarefa.

2. Para cada língua, calcule o número de ocorrências de cada bigrama e trigrama

   Observação. Pode usar qualquer ferramenta disponível, ou fazer o seu próprio programa. Para facilitar a tarefa de avaliação por parte do docente, os ficheiros calculados devem seguir exatamente o formato dos exemplos fornecidos no material de apoio (dados.zip).

   Mais uma vez, se incluiu o PT, pode usar os bigramas e trigramas fornecidos no material de apoio referente à língua portuguesa

3. Crie um programa que carrega os 5 modelos de língua (bigramas e trigramas) e, depois de pedir uma frase, calcula a língua mais provável para essa frase utilizando as contagens de n-gramas. A seleção da língua mais provável consiste em calcular a probabilidade para cada uma das línguas e selecionar a língua que obtiver a maior probabilidade.

   Observação. O programa deve apresentar o valor calculado para cada uma das línguas avaliadas;

4. Crie uma segunda versão do seu programa na qual é usada uma forma de alisamento. Pode usar "Add-1" (também conhecida por Laplace), na qual
   $p(z|xy) = \frac{contagem(xyz)+1}{contagem(xy)+V}$
   Em que V é o tamanho do vocabulário, isto é, o número de letras diferentes

5. Teste os dois programas com 3 frases e comente os resultados obtidos.

6. Comente a viabilidade de desenvolver sistemas que detetem a língua correta.

Submissão

Para submeter o trabalho está disponível o seguinte formulário. Será necessário introduzir a palavra-passe definida anteriormente quando o aluno/grupo se inscreveu.

Submissão: http://pcl.dcti.iscte.pt/www/myscripts/Trabs/submeter_trab02.php

Deverá submeter um ficheiro `zip`, contendo:

- `opcoes.txt`: um ficheiro de texto com a descrição das opções tomadas (max. 1 página A4);
- `lingua1.txt`, `lingua2.txt`, ... : ficheiros de texto com os corpora recolhidos [tarefa 1a];
- `lingua1-nor.txt`, ... : ficheiros de texto com os corpora recolhidos normalizados [tarefa 1b];
- `lingua1.2gr`, `lingua2.2gr`, `lingua3.2gr`, ... : ficheiros com os bigramas obtidos para cada língua;
- `lingua1.3gr`, `lingua2.3gr`, `lingua3.3gr`, ... : ficheiros com os trigramas obtidos para cada língua;
- programas que identificam a língua, simples e com alisamento [tarefas 3 e 4];
- `testes.txt`: um ficheiro com as frases usadas para teste [tarefa 5];
- `resultado.txt`: o ficheiro com os resultados obtidos [tarefa 5];
- `viabilidade.txt`: o ficheiro de texto, não podendo exceder 1 página A4 [tarefa 6];
- `run.sh`, ou `run.bat`: ficheiro com os comandos usados para obter todos os resultados reportados;
- todo o restante código necessário à obtenção dos resultados apresentados

Pode realizar várias submissões, tendo em conta que uma submissão substitui a anterior.

Critérios de avaliação

Na avaliação serão tidos em conta os seguintes critérios:

- Independência do sistema operativo;
- Originalidade;
- Cumprimento de todos os requisitos;
- Correcção na construção dos corpora;
- Correcção das soluções propostas;
- Facilidade para proceder a alterações;
- Cumprimento de todas as regras de submissão. O não cumprimento de qualquer regra implica um desconto mínimo de 5 valores.
- Correcção ortográfica e sintáctica dos documentos submetidos para avaliação.

Política em caso de fraude

Os alunos podem partilhar e/ou trocar ideias entre si, sobre os trabalhos e/ou resolução dos mesmos. No entanto, o trabalho entregue deve corresponder ao esforço individual de cada grupo. São consideradas fraudes as seguintes situações:

- Trabalho parcialmente copiado
- Facilitar a copia através da partilha de ficheiros.

Em caso de detecção de algum tipo de fraude, os trabalhos em questão não são avaliados, sendo enviados à comissão pedagógica, que decide a sanção a aplicar aos alunos envolvidos.
Serão utilizadas as ferramentas *Moss* e *SafeAssign* para detecção automática de copias.