LÍNGUA NATURAL 2014/2015

Mini-Projecto Nº 2 — MP2

A realizar:	☐	individualmente	☒	**em grupo**
Local de trabalho:	☐	aula prática	☒	**casa (TPC)**
Local de entrega:	☐	aula teórica	☒	**submissão electrónica**

Data limite entrega: **até às 12:00 (meio dia) do dia 27/Out**

OBJECTIVOS OPERACIONAIS

Construção de modelos de língua estatísticos. Sua utilização prática.

ENUNCIADO

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 considerar 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`|<<) * p(`i`|<`s`) * p(`m`|`si`) * p(`␣`|`im`) * ... * p(`r`|`ho`) * 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 das probabilidades. Nesse caso, em vez da fórmula anterior, deverá usar a fórmula:

> logprob(`sim␣senhor`) = log(p(`s`|<<)) + log(p(`i`|<`s`)) + log(p(`m`|`si`)) + log(p(`␣`|`im`)) + ... + log(p(`r`|`ho`)) + log(p(>|`or`))

A probabilidade p(`z`|`xy`) pode ser obtida com base em contagens de n-gramas a partir dos dados de treino dessa língua:

> p(`z`|`xy`) = contagem(`xyz`) /⁄ contagem(`xy`)

Por exemplo, se considerar a frase "`sim␣senhor`" (10 carateres) tem as seguintes contagens de bigramas e trigramas:

bigrama	contagem
<`s`	1
`si`	1
`im`	1
`m␣`	1
`␣s`	1
`se`	1
`en`	1
`nh`	1

trigrama	contagem
<<`s`	1
<`si`	1
`sim`	1
`im␣`	1
`m␣s`	1
`se␣`	1
`sen`	1
`enh`	1

ho	1
or	1
r>	1

nho	1
hor	1
or>	1

Execute as seguintes tarefas:

1. Construa um corpus referente a cinco línguas:
   a. Recolha até 100 frases reais em cada uma das línguas;

   b. Processe estes dados de forma a obter textos "normalizados": separando 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. Para facilitar, não se preocupe com a conversão de numerais.

      Nota: Os comandos "grep", "awk", "sed" e "tr" (unix) facilitam muito esta tarefa.

      Nota: Pode usar o corpus fornecido no material de apoio referente à língua portuguesa.

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

   Nota: Pode usar qualquer ferramenta disponível, ou fazer o seu próprio programa.

   Nota: Para facilitar a tarefa de avaliação, os ficheiros calculados devem seguir exatamente o formato dos exemplos fornecidos no material de apoio (pt.zip).

   Nota: 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.

   Nota: O programa deve listar o valor calculado para cada uma das opções (línguas) avaliadas;

4. Crie uma segunda versão do seu programa com as seguintes modificações

   a. Use uma forma de alisamento: pode usar "Add-1" (também conhecida por Laplace), dada por:

      $p(z|xy) = (\text{contagem}(xyz)+1) / (\text{contagem}(xy)+V)$

      em que "V" é o número de bigramas mais 1.

5. Teste os 2 programas para 3 frases. Comente os resultados obtidos.

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

SUBMISSÃO

Submeta no Fenix, agrupamento Mini-Projecto, um ficheiro zip (o nome do ficheiro deve ser formado por concatenação de “MP2-” com o número do grupo e com extensão “.zip”) que deve conter:

- um ficheiro de texto (com o nome “opcoes.txt”) com a descrição das opções tomadas, não podendo exceder 1 página A4;
- os ficheiros com o corpus referentes às 5 línguas (“lingua1.txt”, “lingua2.txt” ... “lingua5.txt”) [tarefa 1.a];
- os ficheiros com o corpus normalizado (“lingua1NOR.txt”, “lingua2NOR.txt” ... “lingua5NOR.txt”) [tarefa 1.b];
- os ficheiros com os bigramas (“lingua1.bigrama”, “lingua2.bigrama” ... “lingua5.bigrama”) [tarefa 2];
- os ficheiros com os trigramas (“lingua1.trigrama”, “lingua2.trigrama” ... “lingua5.trigrama”) [tarefa 2];
- programa que seleciona a língua [tarefa 3];
- programa que seleciona a língua fazendo alisamento [tarefa 4];

- um ficheiro com as frases usadas para teste (“testes.txt”) [tarefa 5];
- o ficheiro com os resultados obtidos (“Resultado.txt”) [tarefa 5];
- o ficheiro de texto (“viabilidade.txt”), não podendo exceder 1 página A4 [tarefa 6];
- um ficheiro de texto ("run.sh", ou "run.bat") com os comandos usados para obter todos os resultados reportados;
- todo o código necessário à obtenção dos resultados apresentados.

Sempre que possível, todos os ficheiros devem conter a identificação do grupo e dos alunos participantes na elaboração deste trabalho.

Critérios de avaliação

Na avaliação serão tidos em conta os seguintes critérios:

1. Independência do sistema operativo;
2. Originalidade;
3. Cumprimento de todos os requisitos;
4. Correção na construção dos corpora;
5. Correção das soluções propostas;
6. Facilidade para proceder a alterações;
7. Cumprimento de todas as regras de submissão. O não cumprimento de qualquer regra implica um desconto mínimo de 2 valores. Se os programas não respeitar o formato de entrada indicado, ocorrerá uma penalização extra de 8 valores (em 20);
8. Correcção ortográfica e sintáctica dos documentos submetidos para avaliação.

Código de honra na Universidade de Stanford (http://www.stanford.edu/dept/vpsa/judicialaffairs/guiding/honorcode.htm)

The Honor Code is the University's statement on academic integrity written by students in 1921. It articulates University expectations of students and faculty in establishing and maintaining the highest standards in academic work:

1. The Honor Code is an undertaking of the students, individually and collectively:
    1. that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;
    2. that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.
2. The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.
3. While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

Examples of conduct that have been regarded as being in violation of the Honor Code include:

- Copying from another’s examination paper or allowing another to copy from one’s own paper
- Unpermitted collaboration
- Plagiarism
- Revising and resubmitting a quiz or exam for regrading, without the instructor’s knowledge and consent
- Giving or receiving unpermitted aid on a take-home examination
- Representing as one’s own work the work of another
- Giving or receiving aid on an academic assignment under circumstances in which a reasonable person should have known that such aid was not permitted

In recent years, most student disciplinary cases have involved Honor Code violations; of these, the most frequent arise when a student submits another’s work as his or her own, or gives or receives unpermitted aid. The standard penalty for a first offense includes a one-quarter suspension from the University and 40 hours of community service. In addition, most faculty members issue a "No Pass" or "No Credit" for the course in which the violation occurred. The

standard penalty for multiple violations (e.g. cheating more than once in the same course) is a three-quarter suspension and 40 or more hours of community service.