LÍNGUA NATURAL 2016/2017

Mini-Projecto Nº 2 — MP2

A realizar:	☐ individualmente	☒ **em grupo**
Local de trabalho:	☐ aula prática	☒ **casa**
Local de entrega:	☐ aula teórica	☒ **submissão electrónica**
Data limite entrega:	**até às 12:00 (meio dia) do dia 7/Nov**	

OBJECTIVOS

Aprender a construir e utilizar modelos de língua estatísticos no processamento de língua natural.

ENUNCIADO

Pretende-se identificar o autor de um texto usando o conhecimento previamente extraído de um corpus de textos de vários escritores.

a. Tendo em conta a coleção de textos referentes a 7 autores portugueses (disponibilizada em "treino.zip"):

   1. Normalize todos os textos para que a pontuação tenha sempre um espaço à direita e à esquerda (esta é a única restrição que tem de respeitar);

   2. Calcule os unigramas e bigramas, sem e com alisamento (qualquer estratégia de alisamento é aceite) para os textos de cada um dos autores.

      ATENÇÃO:

      Pode usar qualquer ferramenta para calcular os ficheiros de unigramas e bigramas (por exemplo, ngram-count, srilm toolkit, ...);

      Para facilitar a tarefa de avaliação, os ficheiros calculados devem apresentar uma de duas sintaxes:
      - contagem por linha (ver os ficheiros “unigramasDEMO.txt” e “bigramasDEMO.txt” que contêm o formato desejado);
      - ARPA format (ver secção 4.8 do [Jurafsky & Martin, 2009], ver o ficheiro “gramasDEMO.arpa” que contém o formato desejado).

b. Tente identificar o autor dos textos da coleção "teste.zip", usando os modelos de língua calculados anteriormente.

   1. Faça três experiências, podendo variar:
      - a normalização (maiúsculas/minúsculas, palavras funcionais, ...);
      - a dimensão dos N-gramas (unigramas/bigramas);
      - os N-gramas a usar (todos, os mais frequentes, ...);
      - ...

      Nota: A totalidade das experiências deve poder ser reproduzida através da execução de um shell script "run.sh".

   2. Faça um relatório (não pode exceder 3 páginas A4) com a análise crítica das experiências/resultados obtidos.

Submissão

Submeta no Fenix, no agrupamento *Mini-projeto*, um ficheiro zip (o nome do ficheiro deve ser formado por concatenação de “MP2-TAGUS” ou “MP2-ALAMEDA” (campus onde estão efetivamente a fazer a UC) com o número do grupo e com extensão “.zip”) que deve conter:

- todo o código necessário à obtenção dos resultados apresentados (exceto ferramentas públicas que tenha usado; estas devem ser claramente identificadas no relatório);
- os ficheiros calculados na alínea a (textos normalizados, unigramas e bigramas);
- o ficheiro run.sh com TODOS os comandos usados para gerar os resultados reportados nas 3 experiências (incluindo a geração dos modelos de língua);
- um relatório (ficheiro txt, máximo 3 páginas), contendo a identificação dos elementos do grupo, a descrição das opções tomadas em cada experiência e comentários críticos aos resultados obtidos.

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. correcção no cálculo dos n-gramas sem e com alisamento (1,0 valores);
2. programa apresenta os valores corretos calculados em cada experiência (0,5 valores);
3. correto funcionamento do "run.sh" (0,5 valores);
4. qualidade do relatório (2 valores):
    - descrição das opções tomadas em cada experiência (0,5 valores);
    - comentários críticos aos resultados obtidos (1,2 valores);
    - correção ortográfica e sintáctica do relatório (0,3 valores);
5. cumprimento de todas as regras de submissão. O não cumprimento de qualquer regra implica um desconto mínimo de 1 valor. Se os programas não respeitarem o formato de entrada indicado, ocorrerá uma penalização extra de 2 valores (em 4).

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.