ISCTE IUL Instituto Universitário de Lisboa Escola de Tecnologias e Arquitectura Dep. de Ciências e Tecnologias da Informação	PCL Ano letivo 2017/2018 1º Semestre **Trabalho 2** Versão 1.0

O trabalho deverá ser realizado em **grupo** (com 2 alunos) ou **individualmente**.
No caso de grupos com mais do que 1 elemento, o relatório deve indicar uma estimativa da percentagem de contribuição de cada elemento para o trabalho, por exemplo:
manuel: 60%, pedro: 40%, juntamente com uma justificação.
A submissão deverá ser feita até ao fim do dia **5 de Novembro de 2017.**

Objetivo: Construção de modelos de língua estatísticos e sua utilização prática

Pretende-se realizar a desambiguação de formas superficiais verbais associadas a um verbo (e.g., *foi* é uma forma verbal do verbo *ir* e do verbo *ser*). Para cada grupo são disponibilizados ficheiros associados a duas formas superficiais. Os grupos com um elemento devem anotar apenas uma delas.

Tarefa 1

Para cada uma das palavras/corpus a desambiguar:

1. Anote o lema em todas as ocorrências das palavras a desambiguar usando a aplicação Java `CorpusAnnotator.class`
   - por exemplo, para a palavra *foi* o corpus a anotar é o ficheiro `foiIrSer.txt` e o ficheiro de parametrização a usar na aplicação é `foiParametrizacao.txt`;
   - para iniciar o processo de anotação deve executar o comando `java CorpusAnnotator`. Abre-se uma janela onde deve selecionar os ficheiros com o *corpus* e a respetiva parametrização;
   - o ficheiro anotado terá o mesmo nome do ficheiro a anotar, mas com a extensão `.out`;
2. Converta o ficheiro com a extensão `.out` para outro com a extensão `.final`, em que
   - são eliminadas as linhas em que a palavra a anotar não é um verbo;
   - são eliminadas as linhas em que a palavra a anotar ocorre duas (ou mais) vezes na mesma frase;
   - são eliminadas as linhas em que a frase tem erros (que dificultem a geração dos modelos);
   - são eliminadas as linhas em que o anotador teve dúvidas;
   - a palavra a desambiguar é substituída pelo lema que aparece no início de cada linha;
   - é removido o lema que aparece no início de cada linha;
3. Calcule os unigramas e bigramas (sem e com alisamento) presentes no ficheiro `.final`. Pode usar qualquer ferramenta para calcular os ficheiros de unigramas e bigramas. Para facilitar a tarefa de avaliação por parte do docente, 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).

Terefa 2

Escreva um programa que deve indicar o lema mais provável para cada frase a processar, de acordo com os modelos de língua carregados. O programa tem como entradas:

- um ficheiro com unigramas (por exemplo, `unigramasDEMO.txt` ou `gramasDEMO.arpa`);
- um ficheiro com bigramas (por exemplo, `bigramasDEMO.txt` ou `gramasDEMO.arpa`);
- um ficheiro contendo a forma superficial ambígua e os respectivos lemas (por exemplo, `foiParametrizacao.txt`);
- um ficheiro com frases de teste a processar (por exemplo, `frasesDEMO.txt`).

O programa deve listar o valor calculado para cada uma das opções avaliadas.

Tarefa 3

Teste o programa desenvolvido com 5 novas frases para cada uma das palavras que lhe estão atribuídas. Comente os resultados obtidos.

Tarefa 4

Comente a viabilidade de desenvolver sistemas que seleccionem o lema correcto.

Submissão

Cada um dos grupos deverá começar por se inscrever usando o formulário seguinte. Note que, apesar do formulário permitir grupos de 3 elementos, o terceiro elemento não pode ser utilizado.

http://pcl.dcti.iscte.pt/www/myscripts/Grupos/inscrever_grupos.php

A submissão do trabalho deverá ser feita através do formulário seguinte. Será necessário introduzir a palavra-passe que foi estabelecida quando o grupo foi inscrito no passo anterior.

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

Deverá submeter um ficheiro *Zip*, contendo:

- um ficheiro de texto (com o nome `opcoes.txt`) com a descrição das opções tomadas, não podendo exceder 1 página A4;
- os 2 ficheiros anotados (`palavra1Anotado.out`, `palavra2Anotado.out`) [tarefa 1.1];
- os 2 ficheiros anotados finais (`palavra1Anotado.final`, `palavra2Anotado.final`) [tarefa 1.2];
- os ficheiros com os unigramas e os bigramas com alisamento (`palavra1Unigramas.txt`, `palavra2Unigramas.txt`, `palavra1Bigramas.txt`, `palavra2Bigramas.txt` ou `palavra1.arpa` e `palavra2.arpa`) [tarefa 1.3];
- os ficheiros com as frases usadas para teste (`palavra1Frases.txt` e `palavra2Frases.txt`) [tarefa 3];
- o ficheiro com os resultados obtidos (`palavra1Resultado.txt` e `palavra2Resultado.txt`) [tarefa 3];
- todo o código necessário à obtenção dos resultados apresentados [tarefa 2];
- o ficheiro de texto (`viabilidade.txt`) com a análise à viabilidade, não podendo exceder 1 página A4 [tarefa 4];
- um ficheiro de texto `run.sh` com os comandos usados para obter todos os resultados reportados.

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 (máximo = 4 valores):

- Correcção na construção dos corpora anotados (1,0 valor);
- Correcção no cálculo dos *n*-gramas sem e com alisamento (0,5 valores);
- Frases de teste (0,2 valores);
- Programa apresenta os valores calculados e usados para escolher o lema mais provável (0,3 valores);
- *Script* `run.sh` (0,2 valores);
- Correção dos valores calculados para os exemplos apresentados (1,3 valores);
- Descrição das opções tomadas e viabilidade (0,3 valores);
- Correção ortográfica e sintáctica (0,2 valores).

O não cumprimento de qualquer regra implica um desconto mínimo de 2 valores.

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 deteção de algum tipo de fraude, os trabalhos em questão não serão avaliados, sendo enviados à Comissão Pedagógica ou ao Conselho Pedagógico, consoante a gravidade da situação, que decidirão a sanção a aplicar aos alunos envolvidos. Serão utilizadas as ferramentas *Moss* e *SafeAssign* para detecção automática de cópias.
Recorda-se ainda que o Anexo I do Código de Conduta Académica, publicado a a 25 de Janeiro de 2016 em Diário da Republica, 2ª Série, nº 16, indica no seu ponto 2 que:

> Quando um trabalho ou outro elemento de avaliação apresentar um nível de coincidência elevado com outros trabalhos (percentagem de coincidência com outras fontes reportada no relatório que o referido software produz), cabe ao docente da UC, orientador ou a qualquer elemento do júri, após a análise qualitativa desse relatório, e em caso de se confirmar a suspeita de plágio, desencadear o respetivo procedimento disciplinar, de acordo com o Regulamento Disciplinar de Discentes do ISCTE-Instituto Universitário de Lisboa, aprovado pela deliberação n.o 2246/2010, de 6 de dezembro.

O ponto 2.1 desse mesmo anexo indica ainda que:

> No âmbito do Regulamento Disciplinar de Discentes do ISCTE- -IUL, são definidas as sanções disciplinares aplicáveis e os seus efeitos, podendo estas variar entre a advertência e a interdição da frequência de atividades escolares no ISCTE-IUL até cinco anos.