iscte TECNOLOGIAS E ARQUITETURA Dep. de Ciências e Tecnologias da Informação	PCL Ano letivo 2021/2022 1º Semestre **Trabalho 2** Versão 1.1

O trabalho deverá ser realizado em **grupo (2 estudantes)** ou individualmente. No caso de grupos com dois elementos, o relatório deve indicar uma estimativa da contribuição de cada elemento para o trabalho. Por exemplo: manuel: 60%, pedro: 40%, juntamente com uma pequena justificação.
A submissão deverá ser feita até ao fim do dia **20 de Novembro de 2021.**

Construção de modelos de língua estatísticos e sua utilização prática

Neste projeto iremos simular um fórum de avaliação. Os fóruns internacionais de avaliação são concursos nos quais os participantes testam os seus sistemas em tarefas específicas e nas mesmas condições. Os conjuntos de treino/validação são dados com antecedência e mais tarde é lançado um conjunto de testes. Nessa altura, os participantes têm um curto período de tempo para produzir os resultados dos seus sistemas e submetê-los na plataforma, por forma a serem avaliados centralmente e permitindo identificar os mais avançados. O vosso objetivo será construir um modelo que identifica a *Categoria* de um par questão+resposta, sendo essa *Categoria* uma das seguintes: GEOGRAPHY, MUSIC, LITERATURE, HISTORY e SCIENCE.
Tal como num fórum de avaliação, encontram-se disponíveis um conjunto para treino e outro de validação (`train.txt` e `eval.txt`). O conjunto final de testes será divulgado um dia antes do prazo de submissão do trabalho.

Objetivo

Os ficheiros `train.txt` e `eval.txt` contêm exemplos de questões e respectivas respostas (uma versão simplificada do conjunto de dados Jeopardy! disponível no Kaggle), previamente classificadas manualmente usando as etiquetas anteriores. Exemplo:

Tipo	Questão	Resposta
GEOGRAPHY	"The Rhine Valley occupies one-third of this 62-square-mile country; the Alps cover the rest"	Liechtenstein
MUSIC	"PBS fans know that ""Evening at Pops"" refers to this city's Pops"	Boston
LITERATURE	"In 1996 he simultaneously published ""The Regulators"" as Richard Bachman & ""Desperation"" under this name"	Stephen King
HISTORY	"In 1843 Congress allocated $30,000 to string one between Baltimore & Washington; it was completed in 1844"	a telegraph wire
SCIENCE	According to Chuck Jones, whenever possible, this force of nature was to be Wile E. Coyote's greatest enemy	gravity

Note que ambos os ficheiros seguem o mesmo formato e não são balanceados. Dado um novo conjunto de questões e respectivas respostas, a ferramenta desenvolvida deverá identificar a categoria destes pares, tendo como base os pares conhecidos. Por exemplo, sendo dado um ficheiro com (novas) questões e respostas:

Questão	Resposta
"Any device that turns one kind of energy into another; a microphone is an example of one"	Transducer
Phoenix lies on a river named for this substance found in the name of another state capital	Salt
"Many consider his 1952 book ""Invisible Man"" the greatest post-war novel about black life in the U.S."	Ralph Ellison

A ferramenta a implementar deverá ser capaz de identificar as respetivas categorias:

```
SCIENCE
GEOGRAPHY
LITERATURE
```

Tarefa 1 — Unigramas e bigramas

Considerando apenas o conjunto de treino, calcule os unigramas (`unigrams_ETIQUETA.txt`) e bigramas (`bigrams_ETIQUETA.txt`) para cada etiqueta. Faça as respetivas contagens e coloque os resultados na diretoria "*counts*". Pode usar qualquer ferramenta para calcular os ficheiros de unigramas e bigramas. O formato dos ficheiros é o seguinte:

Unigramas

```
palavra frequência
palavra frequência
...
```

Bigramas

```
palavra palavra frequência
palavra palavra frequência
...
```

Tarefa 2 — Aplicação de Modelos de Língua

Considerando que o conjunto de validação (`eval.txt`) contém as etiquetas juntamente com as questões, deverá começar por criar os seguintes dois ficheiros.

- `eval-questions.txt`, que apenas inclui as questões (removidas as etiquetas), seguindo exatamente a mesma ordem do ficheiro original
- `eval-labels.txt`, que inclui apenas as etiquetas, seguindo exatamente a mesma ordem do ficheiro original

Escreva um programa que dado um ficheiro com questões (uma questão por linha), escreva no ecrã a etiqueta de cada uma das questões, pela mesma ordem que aparecem no ficheiro (uma etiqueta por linha), de acordo com os modelos de língua carregados.
O programa deve ter como parâmetros:

- Tipo de modelo: *unigramas*, *bigramas* ou *bigramas com alisamento*
- A diretoria onde se encontram os ficheiros com os unigramas e bigramas para cada etiqueta
- Um ficheiro com questões a processar (por exemplo, `eval-questions.txt`).

Exemplo:

```
./lmclassifier -smooth counts eval-questions.txt
```

Teste o desempenho do seu classificador e reporte o resultado obtido no conjunto de validação, usando:

- um modelo de língua baseado em unigramas
- um modelo de língua baseado em bigramas sem alisamento
- um modelo de língua baseado em bigramas com alisamento. Ao aplicar alisamento, o tamanho do vocabulário $|V|$ deve ser calculado usando as palavras de todo o texto do treino e não apenas as palavras de cada uma das categorias isoladamente.

Para avaliar o desempenho dos seus modelos, poderá utilizar e/ou adaptar o programa `evaluate.py`, fornecido juntamente com os dados. Exemplo de utilização:

```
./lmclassifier -smooth counts eval-questions.txt > result.txt
python3 evaluate.py -v data/eval-labels.txt result.txt
```

Tarefa 3 — Pré-processamento dos dados

Repita a Tarefa 1, fazendo algum processamento aos dados. Por exemplo, converta tudo em minúsculas e substitua as palavras que correspondem a quatro dígitos (exemplos: 1991, 1993, 2005) por `_YEAR_`. Recalcule os unigramas e bigramas e aplique os modelos língua novamente às questões do conjunto de validação. Comente os resultados obtidos.
Nota: deverá guardar os novos ficheiros de unigramas e bigramas na diretoria *"counts2"*

Tarefa 4 — Análise dos resultados

Comente a viabilidade de classificar questões usando modelos de língua.

Tarefa 5 — Avaliação final

O ficheiro `test-questions.txt` ficará disponível no e-learning no dia 19 de Novembro (um dia antes do prazo de submissão do trabalho). Recorda-se que o formato do ficheiro é igual ao do ficheiro `eval-questions.txt` e é ilustrado na primeira página deste documento.
Classifique as questões do ficheiro `test-questions.txt` , usando um modelo de língua à sua escolha e o pré-processamento que considerar mais adequado. O resultado obtido deverá ser guardado no ficheiro `results/test-guess.txt`. A avaliação do seu sistema será feita posteriormente pelo "*organizador*" do evento, com base nesse ficheiro e nas anotações manuais que só ele tem. Mais tarde será publicada uma lista de participantes, ordenada de acordo com a pontuação obtida.
O seu objetivo será obter o melhor desempenho possível e poderá usar todo o tipo de processamento ou informação adicional que entender. Por exemplo, poderá juntar o ficheiro de validação ao seu treino, por forma a produzir modelos de língua mais robustos.

Submissão

Cada um dos elementos do grupo deverá inscrever-se através do e-learning, num dos grupos disponíveis para realização do Trabalho 2. Cada um dos grupos deverá submeter um ficheiro *zip*, contendo as seguintes diretorias e ficheiros, dentro da diretoria principal:

- Diretorias "*counts*" e "*counts2*", cada uma contendo 5 pares de ficheiros de unigramas e bigramas, sem e com alisamento [tarefas 1 e 3]
- Diretoria "*data-processed*", com os ficheiros resultantes do pré-processamento [tarefa 3]
- Diretoria "*results*" com o ficheiro `test-guess.txt` [tarefa 5]
- O código correspondente aos programas desenvolvidos nas tarefas 2 e 5, bem como outro código necessário à obtenção dos resultados apresentados
- O ficheiro de texto `run.sh` com os comandos usados para obter todos os resultados reportados
- Um relatório em formato PDF, com um máximo de 2 páginas, contendo:
    1. Identificação dos alunos (número do grupo e nomes)
    2. Breve introdução (breve descrição do problema)
    3. Descrição das opções tomadas
    4. Resultados experimentais e discussão dos resultados (originais, com e sem alisamento; usando pré-processamento)
    5. Resposta ao pedido feito na tarefa 4 e comentários à solução da tarefa 5
    6. Conclusões
    7. Referências Bibliográficas

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

Observação. **Os ficheiros submetidos têm de respeitar exatamente os nomes referidos acima**

Critérios de avaliação

Na avaliação serão tidos em conta os seguintes critérios (máximo = 4 pontos):

- Cálculo dos *n*-gramas (unigramas e bigramas) (0,75 pontos)
- Aplicação dos modelos de língua, unigramas, bigramas sem e com alisamento (1 ponto)
- Resultados com pré-processamento dos dados (0,75 pontos)
- Desempenho na tarefa 5 (0,5 pontos)
- *Script* `run.sh` (0,25 pontos)
- Relatório (0.75 pontos)

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

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, 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 até cinco anos.