INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Processamento Computacional da Língua

2º semestre, 2024/2025

Trabalho 1

Recolha, Análise e *Tokenização* de um *Corpus*

versão 1.0

> Este trabalho foi concebido para ser resolvido em grupos de **2** elementos (aconselhado) ou individualmente. Cada um dos elementos do grupo deverá inscrever-se, através do moodle, num dos grupos disponíveis para realização do Trabalho 1. O trabalho deverá também ser entregue através do moodle, devendo o grupo submeter um ficheiro `.zip` contendo um relatório em formato `.pdf`, assim como os dados recolhidos e o código desenvolvido para a realização das várias tarefas.
>
> O relatório deverá incluir pelo menos os seguintes items: título; resumo; introdução; descrição dos dados recolhidos; descrição das tarefas desenvolvidas e procedimentos efetuados, clarificando as opções tomadas, os recursos utilizados e os resultados obtidos; conclusões e bibliografia. O resumo deve indicar os dados e os métodos utilizados, as opções mais relevantes e as principais conclusões obtidas.
>
> No caso de grupos com mais do que um elemento, no final do relatório deve ser indicada uma estimativa da percentagem de contribuição de cada elemento para o trabalho, juntamente com uma pequena explicação. Por exemplo: manuel: 40%, mariana: 60%. Eventuais atrasos na entrega do trabalho serão penalizados.
>
> **Entrega:** 15 de março de 2025

O objetivo deste trabalho é aplicar os conceitos e as técnicas de Processamento Computacional da Língua (PCL) para recolher, analisar e *tokenizar* um *corpus* composto por textos escritos em português. É permitido usar ou adaptar livremente código ou ferramentas que se encontrem disponíveis *online* — deve ser sempre colocada uma referência para o recurso utilizado. O trabalho a realizar consiste nas etapas correspondentes às secções seguintes.

1 Recolha do *corpus*

(a) Com base num domínio ou tema de interesse, deve ser recolhido um *corpus* com textos em português, usando fontes *online*, tais como *websites*, *blogs*, redes sociais, etc. Alguns domínios de interesse incluem, por exemplo, notícias de jornais; críticas de restaurantes; críticas de hotéis; críticas de produtos; textos literários; conteúdos de páginas da Internet; conteúdos da Wikipedia; *blogs* temáticos; publicações em redes sociais; etc.

(b) O *corpus* deve ser armazenado em ficheiros de texto e deve conter alguns milhares de palavras. Por exemplo, 5000 a 10000 palavras será uma dimensão razoável.

(c) O *corpus* recolhido deve ser dividido em dois conjuntos: um de treino, com cerca de 90% dos textos, e um de teste, com cerca de 10% dos textos.

2 Documentação e caracterização do *corpus*

(a) O *corpus* recolhido deve ser documentado, incluindo informações sobre o domínio, o tema, as fontes utilizadas, datas, idiomas e formato dos textos. Devem ser indicadas as fontes utilizadas e os critérios de seleção dos textos.

(b) Deve ser realizada uma análise estatística e linguística do *corpus*, utilizando ferramentas, tais como o NLTK, o spaCy ou outras. Devem ser calculadas e apresentadas medidas, tais como o número de documentos, o número de palavras, a diversidade de vocabulário (riqueza lexical), as palavras mais comuns e a sua frequência. Devem também ser identificados e comentados alguns aspetos relevantes ou curiosos do *corpus*, tais como o uso de neologismos, estrangeirismos, gírias, etc.

(c) Deve também ser indicado o número de palavras do conjunto de teste que não se encontram no conjunto de treino, considerando uma separação simples de *tokens*, tal como a divisão por espaços em branco, a utilização de expressões regulares, a utilização de uma ferramenta que realize esta operação ou a utilização de um modelo pré-treinado. Apresente exemplos.

3 *Tokenização*

Nesta etapa do trabalho devem ser exploradas abordagens de *tokenização* mais modernas baseadas em sub-palavras.

3.1 Byte Pair Encoding (BPE)

(a) O método de *tokenização* BPE é uma estratégia que divide as palavras em unidades menores, baseadas na frequência e na probabilidade dos pares de carateres. Estude o seu funcionamento e implemente-o numa linguagem de programação à sua escolha.

(b) Usando a sua implementação do BPE, treine um modelo com base no seu conjunto de treino e aplique-o ao conjunto de teste.

(c) Apresente alguns exemplos do conjunto de teste que permitam ilustrar as diferenças entre o BPE e um método de *tokenização* tradicional baseado em palavras.

3.2 Comparação de métodos baseados em sub-palavras

(a) Com base na informação disponível em `https://huggingface.co/learn/nlp-course/en/chapter6/` e `https://huggingface.co/docs/tokenizers/quicktour`, aplique ao seu conjunto de teste um tokenizador pré-treinado baseado em BPE, assim como um baseado num dos outros métodos apresentados (e.g., WordPiece ou Unigram).

(b) Compare os resultados obtidos com os *tokenizadores* pré-treinados entre eles e com os obtidos usando a sua implementação do método BPE. Use alguns exemplos do conjunto de teste para ilustrar as diferenças entre os métodos.

(c) Avalie a eficácia dos métodos de *tokenização* utilizados em termos de preservação de informação semântica e capacidade de lidar com palavras fora do vocabulário.

Política em caso de fraude

Os estudantes 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. As seguintes situações são consideradas fraudes:

- Trabalho parcialmente copiado;
- Facilitar a copia através da partilha de ficheiros;
- Utilizar material alheio sem referir a sua fonte.

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 deteção 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.*