MP2 – Segundo Miniprojecto de Língua Natural
IST, Alameda e Tagus,
2018

Data de entrega do enunciado:
– 15 de Outubro de 2018

Data de entrega do projecto:
– 5 de Novembro de 2019 (23h 59)

Número de alunos por grupo:
– 2 ou 3 alunos; não têm de ser os mesmos do MP1

Dúvidas:
luisa.coheur@tecnico.ulisboa.pt

Objetivo do projeto:
O projeto tem como objetivo desenvolver uma métrica de similaridade que permita identificar o tipo de uma questão dada sobre cinema. Na base deste processo está um ficheiro com questões sobre cinema (**QuestoesConhecidas.txt**), não balanceado, cujo tipo foi previamente anotado, contendo exemplos para os 16 tipos de questões possíveis[1]. Dado um novo conjunto de questões, a métrica desenvolvida será responsável por devolver o tipo destas questões, tendo como base as questões conhecidas. Por exemplo, considere que em **QuestoesConhecidas.txt** se encontram as frases:

actor_name	*What are the most relevant actors in Bad Boys?*
budget	*How much budget did Valkyrie have?*
release_date	*The debut of The Matrix occurred in what month?*

Sendo dado um ficheiro com (novas) questões:

Who starred in Batman Begins?
Release date of Inside Out?

A métrica implementada deverá ser capaz de identificar as etiquetas:

actor_name
release_date

que correspondem às etiquetas das questões conhecidas pelo sistema "mais próximas" das novas questões. De notar que, para fins de avaliação, as etiquetas calculadas deverão ser enviadas, uma por linha, para o ficheiro **Resultados.txt** e que a sequência em que aparecem no ficheiro corresponde à sequência das questões a que dizem respeito, isto é, a etiqueta devolvida na linha 33 no ficheiro **Resultados.txt** corresponderá à nova questão da linha 33.

[1] actor_name, budget, character_name, genre, keyword, original_language, original_title, overview, person_name, production_company, production_country, release_date, revenue, runtime, spoken_language, vote_avg

As propostas deverão ser implementadas preferencialmente em **Python**. Atenção que os alunos **PODEM** usar implementações disponíveis de funções potencialmente úteis desde que tenham o cuidado de identificar as fontes.
Os alunos deverão entregar um pacote com o código, que deve incluir obrigatoriamente o ficheiro **run.sh**. Este recebe como primeiro argumento o ficheiro das questões conhecidas e como segundo argumento o ficheiro com as questões a anotar, escrevendo em **Resultados.txt** o resultado do programa, na mesma diretoria, tal como especificado anteriormente. A avaliação será feita tendo em conta as etiquetas de **Resultados.txt**, após execução da seguinte linha de comando:

./run.sh ARG1 ARG2 > Resultados.txt

A medida usada será a *accuracy*.

Sugestão: podem usar o ficheiro **NovasQuestoes.txt** para desenvolvimento, dado que a solução é dada (encontra-se no ficheiro **NovasQuestoesResultado.txt**). No entanto, devem ter em conta que não cobre todo os tipos de questões. Será também disponibilizado um conjunto de ficheiros adicionais, contendo nomes de filmes, nomes de atores, etc. podendo os alunos tirar partido desses ficheiros ou não. De notar que não é garantida a correção desses dados (bem-vindos ao mundo real!).

Relatório:

- Pdf com um máximo de 4 páginas, contendo:
    1. Identificação dos alunos (número do grupo e nomes)
    2. Breve introdução (breve descrição do problema)
    3. Proposta de solução, motivando e identificando claramente a técnica desenvolvida
    4. Resultados experimentais e discussão dos resultados. Note que deverão ser reportados os resultados obtidos para o ficheiro **NovasQuestoes.txt**, apesar de ser um corpus de desenvolvimento e não de teste.
    5. Conclusões e trabalho futuro
    6. Bibliografia

Entrega:

- Via fénix até às 23h 59 do dia 5 de Novembro de 2018
- ZIP (e não rar) do projecto com o número do grupo (ex: 3.zip)
- No zip devem constar os seguintes ficheiros:
    - O ficheiro num-grupo.pdf (ex: 3.pdf) contendo o relatório
    - O ficheiro **ResultadosDes.txt** com as etiquetas obtidas através de:
      ./run.sh **QuestoesConhecidas.txt NovasQuestoes.txt** > ResultadosDes.txt
    - O código do projeto, incluindo o ficheiro run.sh

Avaliação do Projecto:

- Relatório e avaliação automática (com novas questões): 10 valores cada;
- O desrespeito por alguma das especificações relativas à entrega do projeto leva a penalizações e, em casos extremos, à sua não avaliação;
- Os alunos poderão ser chamados a defender o projeto.