Grupo 19

Gonçalo Bustorff Silva 82050

Guilherme Gomes 89378

**MP2 – Segundo Miniprojecto de Língua Natural**

**Descrição do problema**

É necessário desenvolver uma métrica de similaridade que permite identificar o tipo de uma questão de cinema. Dado o ficheiro **QuestoesConhecidas.txt**, onde se encontram questões pré-anotadas, e outros recursos como lista de filmes ou pessoas consideradas na solução, deverá ser possível classificar as novas perguntas com um dos 16 tipos possíveis:

*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*

Exemplificando, diga-se que nas **QuestoesConhecidas.txt** encontram-se as perguntas:

*budget How much budget did Valkyrie have?*

*budget What was the budget of Furious 7?*

Dada a nova questão:

*What was American Gangster's budget?*

A métrica implementada deverá classificar a pergunta do tipo:

*budget*

**Proposta de solução**

A métrica desenvolvida é uma junção de várias, devolvendo a moda de voto de 3 classificadores já implementados.

Primeiramente, esta carrega os ficheiros de recurso **list\_movies.txt** e **list\_people.txt**, linha a linha para cada lista correspondente. Da mesma forma é guardada a lista de perguntas pré-anotadas do ARG1. As novas perguntas, presentes no ARG2, são lidas linha a linha, tokenizadas por frase, removidas as *stops words*, e caso se encontrem ocorrências de *substrings* presentes de forma exata nas 2 listas de recurso processadas, ocorre uma substituição por *movie\_title* e *person\_name.* Assim evita-se que títulos de filmes ou nomes de atores/personagens influenciem o peso da decisão, substituindo pelas palavras referidas, fazendo com que estas se tornem muito comuns, diminuindo a informação, e peso, que estas carregam. Exemplos deste processo:

*Which are the most relevant actors in Bad Boys? → Which relevant actors movie\_title?*

*What movies did Michael Bay direct? → What movies person\_name direct?*

Apresenta alguns erros devido ao *matching* exato na lista, mas parcial na questão:

*Which characters were there on Goldfinger? → Which characters movie\_titleldfinger? –* Existe “Gold” na list\_movies.

*Which are the most relevant actors in Great Gatsby? → Which relevant actors Great Gatsby? –* Não existe “Great Gatsby”, mas sim “The Great Gatsby”.

Visto que estes pequenos defeitos não têm um efeito notável na aprendizagem efetuada (demonstrado na secção de resultados), mantivemos a substituição. Pela mesma razão, e pelo estado da **list\_**