3.8.3 Eliminação de stopwords

Alguns termos tem pouco valor semântico, como artigos, preposições e pronomes, esses termos são conhecidos como *stopwords* (BAEZA-YATES; RIBEIRO-NETO, 2013). A eliminação desses termos diminui a quantidade de memória utilizada, o tempo de processamento em operações e ainda pode evitar que documentos não relevantes sejam recuperados. Uma das formas de eliminar esses termos é excluir todas as palavras com poucos caracteres, outra é armazenar uma coleção de palavras pouco relevantes e usá-las para identificar *stopwords*.

No diretório "*ArquivosGerais*" encontram se dois documentos, "*stopwords.txt*" e "*!stopwords.txt*", o primeiro armazena uma coleção de *stopwords* e o segundo uma coleção de não *stopwords*. O sistema considera que palavras pequenas, de quatro caracteres, não são significativas. A coleção de termos "*!stopwords.txt"* é usada para exceções a essa regra, ou seja, para termos pequenos que são relevantes.

3.9 Experimentos

Com o propósito de conseguir informações qualitativas e quantitativas, experimentos foram relevantes. A Tabela 1 apresenta os resultados obtidos por intermédio da métrica MRR.

Nesta aplicação da métrica foi considerada que a resposta correta seria aquela contida no documento de onde foi retirada a pergunta. Os passos para aplicação da métrica MRR foram os seguintes:

1. Abertura de um documento da base de documentos e identificação da pergunta e da resposta contidas nele;
2. Digitação da pergunta identificada (sequência exata);
3. Identificação da posição do documento na lista de documento relevantes criada pelo modelo probabilístico;
4. Aplicação da métrica.