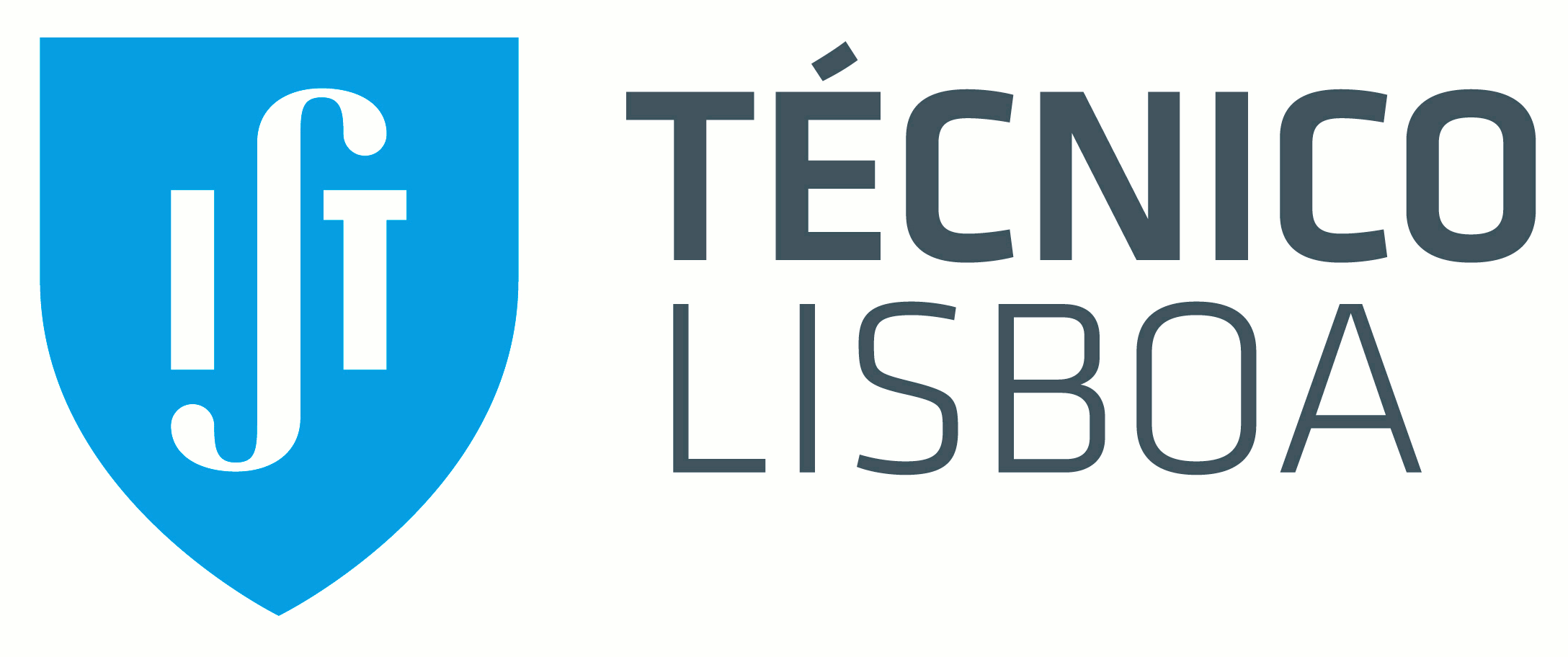
**Natural Language**

****

Scientific Report

First Project – A Retrieval-based chatbot

Group 7

Ricardo Ventura – 79031; Maxwell Junior – 79457; Isabel Raposo - 81524

**Introduction**

On this project we were proposed to build a retrieval-based chatbot with FAQs from “Balcão do Empreendedor”. These FAQs constitute the chatbot’s knowledge base (KB). By being given a user request, the chatbot should find the most similar question from the KB and return an ID corresponding to its answer. Therefore, this is a retrieval-based conversational agent as it doesn’t really reply an answer, nor does it perform any kind of additional interaction.

**Proposal**

We considered 4 preprocessing techniques (all use lowercasing, removal of punctuation and tokenization):

P1 - standard procedure with no additional preprocessing

P2 – with stemming

P3 – with stop words removal

P4 – P2 and P3 combined

and combined them with 3 metrics used, the Cosine similarity, the Jaccard distance and the Dice distance. The accuracies obtained are shown below on table 1, on Experimental Results section.

**Used Corpora**

The corpora we used was based on a FAQ, on which we added more questions by paraphrasing the originals, and thus having a wider KB.

We found out that here are indistinguishable questions with the same answers, but these answers have different IDs. When searching the best answer Id for a question, we are probably getting some Ids correctly but the result is incorrect, which might lead to a decrease on accuracy (we’ll get into more detail on Experimental Results).  
Also, some FAQs only have 1 question what might have decrease accuracy

**Experimental Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics / Pre-processing** | P1 | P2 | P3 | P4 |
| Cosine + Tf-Idf | 0.6833 | 0.7077 | 0.6782 | 0.7067 |
| Jaccard | 0.6142 | 0.6317 | 0.6666 | 0.6930 |
| Dice | 0.5958 | 0.6215 | 0.6347 | 0.6793 |

Table 1. Values on the table correspond to accuracy, resulting from an average of 10 runs of each combination of metrics and preprocessing techniques.

**Conclusion and future work**

**Bibliography**