Natural Language

First Project

A retrieval-based chatbot[1]

Técnico, Alameda and Tagus,
2019

This project should be done in group (the same group as for the second homework).

Questions: meic-ln@disciplinas.tecnico.ulisboa.pt (subject: Project 1)

Goal:
To build a retrieval-based chatbot with FAQs from "Balcão do Empreendedor". These FAQs constitute the chatbot's knowledge base (KB), which was built by the different groups for HW2.

In detail:
Being given a user request, the chatbot should find the most "similar question" from the KB (no information should be added to this KB!) and return the ID of its answer. As an example, being given a list of user requests, in which the user request in line 56 is:

Demora quanto tempo para emitir um certificado de admissibilidade?

and considering that in the KB.xml you have:

```
<perguntas>
        <pergunta>Em que prazo é emitido um certificado de admissibilidade?</pergunta>
        <pergunta>Quanto demora a ser emitido um certificado de admissibilidade?</pergunta>
        <pergunta>O certificado de admissibilidade é emitido em quanto tempo?</pergunta>
        <pergunta>Em quantos dias é emitido um certificado de admissibilidade?</pergunta>
</perguntas>
<resposta id = "145">
    O prazo previsto na Lei é de 10 dias, mas, em regra, os certificados são emitidos entre três a cinco dias úteis.
 </resposta>
```

Your system should return ***in line 56***:

145[2]

If you decide that no question is similar enough to the user request (*Viste o Endgame? Qual é a tua flor favorita?*), then return 0 in line 56.

When building the chatbot, you should compare different pre-processing techniques and/or distance/similarity metrics. At least three experiments should be reported. As an example, a group could report results by considering:

[1] It should be clear that in this project we are building a retrieval-based conversational agent, but not a "chatbot" as an agent that performs small talk.

[2] A bot would return the answer and not the ID.

a) data in lowercase + Jaccard;
b) data in lowercase and stemmed + Jaccard;
c) data in lowercase + Dice.

You should implement your chatbot in Python3. You can use measures already available (including the source code), as long as you identify the source. Nevertheless, you are also free to create your own measures.

You should implement **chatbot.py**, which receives as first argument the FAQs file (**KB.xml**) and, as second argument, a file with test questions (**test.txt**, just a list of questions). Results (a list of ID's) should be written in the file **resultados.txt** (same directory). So, to run your program you should do:

```
python3 chatbot.py KB.xml test.txt
```

Scientific Report:

- NUM.pdf[3] with a maximum of 4 pages, containing the following parts:
  1. Group identification (group number +_names)
  2. Introduction (short description of the problem)
  3. Proposal (clearly explain your approach)
  4. Used corpora (briefly explain how you have done HW2, if you have questions, found errors, etc.)
  5. Experimental results (besides results, include error analysis and a brief discussion of results)
  6. Conclusions and future work (main conclusions and what would you do if you add extra time)
  7. Bibliography

Evaluation:

A*ccuracy* will be the evaluation measure. No project should take more than 3 minutes to return the answers to 50 user requests. On the 25/10, around 8 AM, students will receive a test file (around 50 user requests). Each group should run their system with that file and return the obtained results (in resultados.txt, as previously explained), which will give the grade for the automatic evaluation. We will randomly select a set of projects that will be run manually with the same test set. If any difference in results is found, the group will have a 0 in the project.

- Scientific report: 14/20;
- Accuracy: 6/20.

Submit your work (before 25/10/2019, 23h 59!!!)

- Via Fénix.
- zip (and NOT rar) of the project with the group number (ex: 3.zip). (-2 if rar)
- zip should contain:
  - File NUM.pdf (ex: 3.pdf) with the scientific report;
  - File chatbot.py with the project code (you can have extra python files);
  - File resultados.txt with the results from the given test set.

[3] From now on NUM is the number of the group.