Project Report

Chatbot

Team: Houdini III

[1]Siddharth Bulia - 130050012

[2]Animesh Baranawal - 130050013

[3]Rawal Khirodkar - 130050014

27th October, 2015

Outline

Outline

Problem Statement

Description

- We are trying to implement a **Domain specific knowledge System** to deliver answer to frequently asked questions related to CSE IITB.
- The implementation of this project on a University environment is particularly useful for students looking for information regarding CSE IITB, and its course curriculum. Even though most of the information is available on the web, students often like to have personal interaction.
- The main goal of such a system is to conveniently retrieve information without having to look or browse several web pages to fetch answers to frequently asked questions.

Outline

Problem Statement

User Inputs

- Greetings or FAQ queries in Natural Language (English).
- The FAQ queries are restricted to be related to CSE IITB Courses (The current database only handles few courses, although it can be easily expanded to support on any course).
- For instance, the queries can be regarding Venue, Instructor Name, Grading Statistics,etc. of a particular course.

Outline

Problem Statement

Expected Output

System responses can be majorly categorized into three types:

- **Niceties**: Polite social responses with respect to greetings from the user.
- **Domain Specific responses**: Responses closely satisfying the user queries.
- **Apologetic responses**: Responses meant to convey inability to retrieve requested information.

Outline

Approach

Design and Code Overview

We are using Python as programming language along with AIML (Artificial Intelligence Markup Language) to do pattern matching for response selection.

Following are the **two phases** in execution of our Enquiry System:

- **Training Phase**
  - We have used natural language processing library NLTK to process raw queries (training set) and convert them into a set of synonymous words (reduced query). This in short describes the context/concept of the query. Of course the golden rule applies, more data, better results.
  - The expected answer to such reduced query is known and we use this knowledge to generate aiml/xml files for pattern matching purposes.

Approach

Design and Code Overview

- **Pattern Matching Phase**
  - User's input is again broken down to a reduced query using NLP and we use AIML files generated from phase 1 to find the closest possible pattern existing in our database for which the answer is known..
  - If such pattern exists we output the information retrieved from the database or otherwise we just try to stall the conversation to keep the user interested or apologize for the inability to answer.
- We have tried to automatize the generation of aiml and reduction of queries as much as possible to ensure scalability of the database.

Outline

Approach

References

- **A.L.I.C.E** (Artificial Linguistic Internet Computer Entity) which is an award winning open source natural language artificial intelligence chat robot which utilizes AIML (Artificial Intelligence Markup Language) to form responses to queries.
- **Natural Language Toolkit** (NLTK Python) is used mainly due to its vastness of corpora and lexical resources. Specifically majorly we are using Parts of Speech Tagger, Tokenize, Wordnet, Morphy, Synset.

Contribution of Team Members

- It was a collaborative effort by all the team members and thus equal contribution from each of us.
- Work splitoff:-
  - Siddharth Bulia (130050012):- 33.33%
  - Animesh Baranawal (130050013):- 33.33%
  - Rawal Khirodkar (130050014):- 33.33%

Learning from the Project

This was a very interesting project with a lot of scope to be of real benefits to students. We learnt about the following techniques, concepts, tools:-

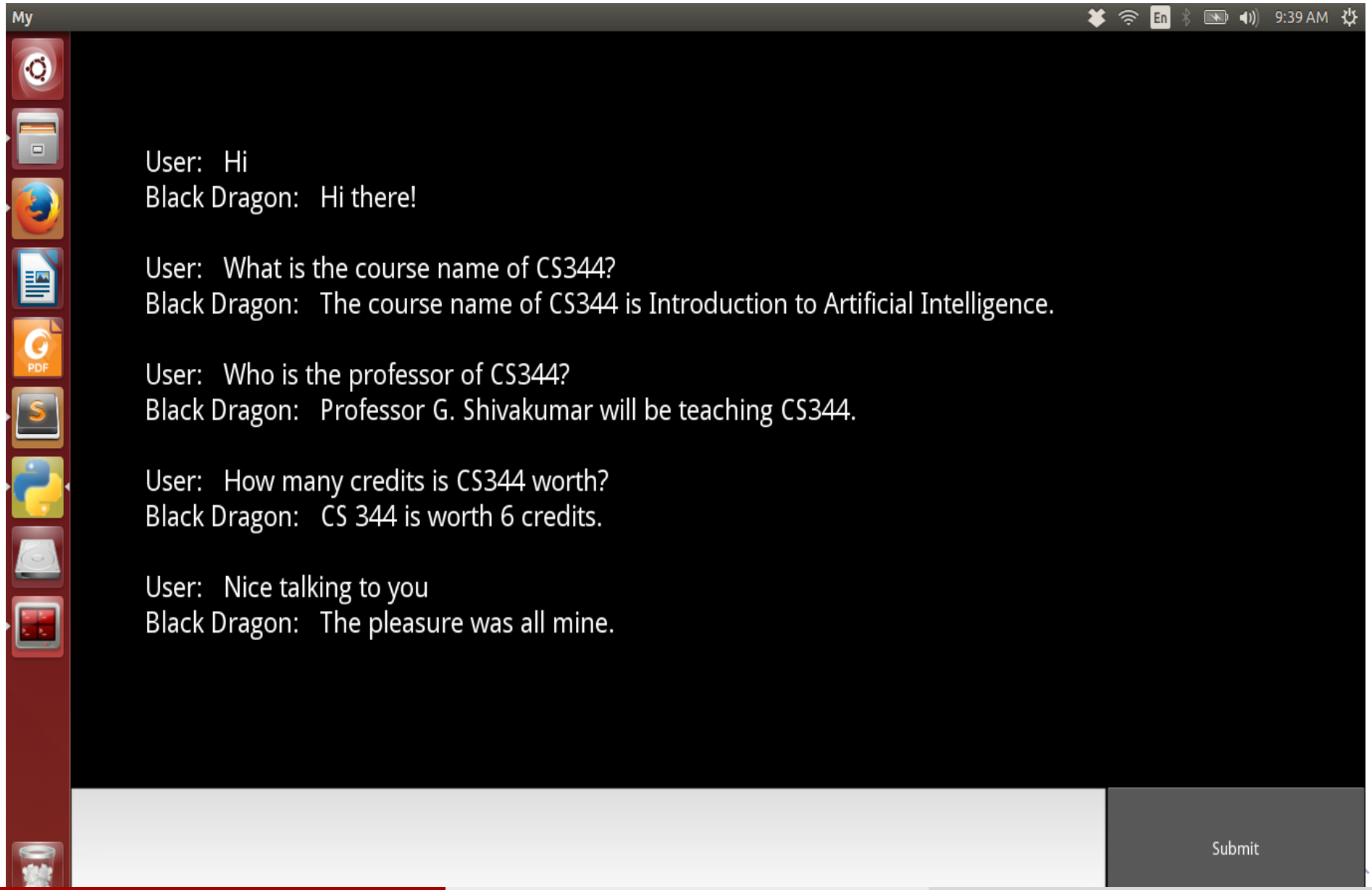
- Natural Language Processing (NLTK).
- Python Artificial Intelligence Markup Language (PyAIML)
- Pattern Matching
- Regular Expressions
- Kivy Python Graphics Library

Further Improvements

Following improvements are possible:-

- Increase the scope of queries.
- Ask user for clarifications on a poorly pattern matched query.
- Web crawling for hunting information requested in query in real time.
- Natural Language processing part can be improved by incorporating relevant corpus of words tagged with its parts of speech.
- Implementing Hidden Markov Model for better confidence on correctness of output.

Screenshots



Screenshots