Chatbot Project 1: 250 Points

Project Objective: Create a chatbot using NLP techniques learned in class so far. This means that machine learning techniques cannot be used in this project. The chatbot should be able to carry on a limited conversation in a particular domain using a knowledge base or knowledge from the web, and knowledge it learns from the user.

Overview: The project has three parts:

1. 100 points: Create a knowledge base (corpus) by writing a custom web crawler
2. 100 points: Create a chatbot that can discuss a topic and remember things about the user
3. 50 points: Report and Evaluation

Part 1 Web Crawler Details (100 points)

Objective: Create a knowledge base using a custom web crawler

- Start with 1-3 URLs that represent a topic (a sport, a celebrity, a place, etc.) and crawl to find 15 – 25 relevant URLs. Make sure that some of the URLs are outside the original starter URLs.
- Write a function to loop through the URLs, scrape text off each page and write it to a file. Write each URL's text to a separate file.
- Write another function to clean up the text files. Read in each raw file and clean it up as much as possible with NLP techniques. If you have x files in, you will have x files out.
- Write a function to extract at least 25 important terms from the cleaned-up files using an importance measure such as tf-idf. You might want to lower-case everything, remove stopwords and punctuation first. Output the top 25-40 important terms.
- Manually determine the top 10-15 terms based on your domain knowledge.
- Build a searchable knowledge base of facts that a chatbot can share related to your important terms. The "knowledge base" can be a simple as a Python dict which you pickle or a simple sql database.
- See Part 3 Report for further requirements.

Part 2 Chatbot Details (100 points)

Objective: Create a chatbot that can carry on a limited conversation on a narrow subject

- Design and build a chatbot from scratch. Your code should include:
    - Retrieval using similarity measures from the corpus you created in part 1
    - Hard-coded responses with some randomization (perhaps for greetings, etc.)
    - Optional live web look-ups

- Maintain a user model within your chatbot system. You should have a different user model saved for each user who converses with the bot. The user model should store the user's name, personal information it gathers from the dialog, and the user's likes and dislikes. Add personalized remarks from the user model to the dialog engine. The user model can be a simple text or xml file.
- Make sure your project includes NLP techniques learned in class. Examples: parse user response, use term frequency measures, NER, or information retrieval techniques, or any techniques we learned in class.

Part 3 Webcrawler Report, Chatbot Report and Evaluation (50 points)

- For the web crawler, write a page or two:
  - Describe how you created the knowledge base
  - Include screen shots of the knowledge base
  - Share your important terms
- For the chatbot, write up:
  - A system description, including description of specific NLP techniques you used and how you used them
  - A diagram of your dialog tree or logic
  - One or more sample dialog interactions
  - An appendix for the knowledge base (and live lookup) you created with samples
  - An appendix for sample user models that were created
  - Evaluations of the chatbot and analysis of its strengths and weaknesses. In addition to your own evaluation, get survey results from people who are not on the team using Likert-style questions, probably at least 3 questions
- No specific format/font/length is required for the report, but it should look professional

What to upload to eLearning and send to peers (zipped together):

- Webcrawler code and files
- Chatbot code
- Report on webcrawler and chatbot