REPORT- BADMINTON CHATBOT

-R.Guru Ravi Shanker(2018114011)
-Anubhav Sharma(2018114007)

Prerequisites :

->Python packages flask,nltk,re,beautiful soup,random,sklearn and string.

How To Run:

->To view in a local host
Run python3 app.py
After that open the local host link that comes after the query
->For terminal based output
Run python3 -W ignore main.py
It would be better if you redirect it to output.txt for better reading of Hindi alphabets
->Copy and paste the queries from the queries.txt file to get the functionality of the bot.
->You can also ask questions regarding the game of badminton it will give better output if the keyword is present in the data.txt.
->data.txt file was generated by data.py if you
Run python3 data.py
prerequisites - urllib,beautiful soup.
->It will write the rendered text from wikipedia page to the data.txt file

Algorithm:

->Used cosine similarity to check the similarities between two statements and considered the best two similar statements from the data with the human query as the result.
->Hard coded the greetings and exit statements.

Output:

For Input Queries

1.प्रकाश पादुकोण कौन है (describing functionality of proper noun)
2.बैडमिंटन खेल क्या है
3.बैडमिंटन क्या खेल है(Changing order doesn.t change the answer)
4.बैडमिंटन कोर्ट के आकार क्या होते है (One with the error as we have to ensure we have to have some matching with the relevant text).
5.बैडमिंटन 'कोर्ट' के आकार क्या होते है
6.बैडमिंटन की शुरुआत कब हुई थी
7.बैडमिंटन खेल की शुरुआत कब हुई थी
8.भारत में बैडमिंटन खेल की शुरुआत कब हुई थी
9.भारत में badminton खेल की शुरुआत कब हुई थी (ensured keyword is in hindi).
10.शटलकॉक के बारे में कुछ बताओ
11.शटलकॉक के बारे में tell me something
12.स्मैश क्या होता है
13.बैडमिंटन सेना नेहवाल सिंधु
14. What is नेटशॉट
15.शटलकॉक कितने किस्मों की होती है
16.शटलकॉक कितने types की होती है (since keyword is now in the foriegn language it wouldn’t recognise it as the keyword).
******17.आज की जानकारी के लिए धन्यवाद,अलविदा
18.अलविदा
19.धन्यवाद

नमस्कार |
आप कैसे है ?
मैं हूँ बैडमिंटन चैम्प और मैं आपके बैडमिंटन से जुड़े सभी सवालों का उत्तर देने मैं सक्षम हूँ ।
बैडमिंटन चैम्प: भारत में बैडमिंटन के कई महान् एकल खिलाड़ी हुए हैं, लेकिन भारतीय बैडमिंटन को सही मायने में दुनिया के सामने लाने का श्रेय जाता है- प्रकाश पादुकोण को, जिन्होंने 1981 के 'क्वालालांपुर विश्व कप फाइनल' में चीन के सुपरस्टार 'हान जियान' को 15-0 से हराकर चीनियों के सपनो को ध्वस्त कर दिया था। बैडमिंटन खेल एक ऐसा खेल है जो हर उम्र के लोगों को अपनी ओर आकर्षित करता है।
बैडमिंटन चैम्प: बैडमिंटन रैकेट से खेला जाने वाला, एक अंतर्राष्ट्रीय खेल है। प्राकृतिक रूप से घूमने के कारण कुछ स्ट्रोक को प्रभावित होते हैं: अगर स्लाइसिंग एक्शन बाए से दाहिनी ओर के बजाए दाहिने से बायीं ओर हो तो टम्बलिंग नेटशॉट कहीं अधिक प्रभावी होता है।
बैडमिंटन चैम्प: बैडमिंटन रैकेट से खेला जाने वाला, एक अंतर्राष्ट्रीय खेल है। प्राकृतिक रूप से घूमने के कारण कुछ स्ट्रोक को प्रभावित होते हैं: अगर स्लाइसिंग एक्शन बाए से दाहिनी ओर के बजाए दाहिने से बायीं ओर हो तो टम्बलिंग नेटशॉट कहीं अधिक प्रभावी होता है।
बैडमिंटन चैम्प: विशेष रूप से, रैकेट और शटलकॉक के डिजाइन और आकार को लेकर नियम सीमाबद्ध हैं। नियम निर्दिष्ट करता है कि कौन-सा उपकरण इस्तेमाल किया जा सकता है।
बैडमिंटन चैम्प: युगल खेल के लिए 'कोर्ट' का आकार 44 फुट X 20 फुट तथा एकल खेल के लिए 44 फुट X 17 फुट होता है। बैडमिंटन तीन प्रकार से खेला जाता है- इन तीनों खेलों के लिए बैडमिंटन कोर्ट की नाप 11/2 (4 सेंटीमीटर) सफ़ेद रंग या लाल रंग की रेखाओं से स्पष्ट की जाती है।
बैडमिंटन चैम्प: बैडमिंटन खेल की शुरुआत 19वीं सदी में हुई। भारत में प्रतिभावान बैडमिंटन खिलाड़ियों में प्रकाश पादुकोण, पुलेला गोपीचंद, अभिन श्याम गुप्ता, निखिल कानितकर, सचिन राठी, अपर्णा पोपट, साइना नेहवाल और नेहा अटवाल प्रमुख हैं।
बैडमिंटन चैम्प: बैडमिंटन खेल की शुरुआत 19वीं सदी में हुई। भारत में प्रतिभावान बैडमिंटन खिलाड़ियों में प्रकाश पादुकोण, पुलेला गोपीचंद, अभिन श्याम गुप्ता, निखिल कानितकर, सचिन राठी, अपर्णा पोपट, साइना नेहवाल और नेहा अटवाल प्रमुख हैं।
बैडमिंटन चैम्प: बैडमिंटन खेल की शुरुआत 19वीं सदी में हुई। भारत में प्रतिभावान बैडमिंटन खिलाड़ियों में प्रकाश पादुकोण, पुलेला गोपीचंद, अभिन श्याम गुप्ता, निखिल कानितकर, सचिन राठी, अपर्णा पोपट, साइना नेहवाल और नेहा अटवाल प्रमुख हैं।
बैडमिंटन चैम्प: भारत में बैडमिंटन का खेल बहुत लोकप्रिय है। भारत इस संघ से 1936 में जुड़ा।
बैडमिंटन चैम्प: भारत में बैडमिंटन का खेल बहुत लोकप्रिय है। भारत इस संघ से 1936 में जुड़ा।

बैडमिंटन चैम्प: सही गति के लिए शटलकॉक के परीक्षण का भी नियम में प्रावधान हैं: 3। 1 : शटलकॉक की जांच के लिए फुल अंडरहैण्ड स्ट्रोक का उपयोग करें जो शटलकॉक को पिछली बाउंड्री रेखा तक ले जाता है। विशेष रूप से, रैकेट और शटलकॉक के डिजाइन और आकार को लेकर नियम सीमाबद्ध हैं।
बैडमिंटन चैम्प: सही गति के लिए शटलकॉक के परीक्षण का भी नियम में प्रावधान हैं: 3। 1 : शटलकॉक की जांच के लिए फुल अंडरहैण्ड स्ट्रोक का उपयोग करें जो शटलकॉक को पिछली बाउंड्री रेखा तक ले जाता है। विशेष रूप से, रैकेट और शटलकॉक के डिजाइन और आकार को लेकर नियम सीमाबद्ध हैं।
बैडमिंटन चैम्प: युगल की तुलना में एकल में स्मैश कम ही देखने में आता है, क्योंकि खिलाड़ी स्मैश करने की आदर्श स्थिति में कम ही होते हैं और अगर स्मैश वापस लौट कर आता है तो स्मैश करनेवाले को अक्सर चोट लग जाती है। ड्राप्सशॉर्ट और नेटशॉर्ट के साथ लिफ्ट और क्लियर के संयोजन से खिलाड़ी कोर्ट की पूरी लंबाई का फायदा उठाता है।
बैडमिंटन चैम्प: भारत में प्रतिभावान बैडमिंटन खिलाड़ियों में प्रकाश पादुकोण, पुलेला गोपीचंद, अभिन श्याम गुप्ता, निखिल कानितकर, सचिन राठी, अपर्णा पोपट, साइना नेहवाल और नेहा अटवाल प्रमुख हैं। पीछे की 'गैलरी' 21/2 फुट तथा 'साईड गैलरी' 11/2 फुट होती है।
बैडमिंटन चैम्प: प्राकृतिक रूप से घूमने के कारण कुछ स्ट्रोक को प्रभावित होते हैं: अगर स्लाइसिंग एक्शन बाए से दाहिनी ओर के बजाए दाहिने से बायीं ओर हो तो टम्बलिंग नेटशॉट कहीं अधिक प्रभावी होता है। जब शटलकॉक गिरता है तब यह घड़ी की विपरीत दिशा में जैसा कि ऊपर दिखाया गया है घूमता है।
बैडमिंटन चैम्प: इसके अतिरिक्त, नायलॉन शटलकॉक तीन किस्मों के होते हैं, हरेक किस्म अलग तरह के तापमान के लिए होते हैं। ये नायलॉन शटल या तो प्राकृतिक कॉर्क या कृत्रिम फोम बेस और प्लास्टिक के घेरे से बनाये जाते हैं।
बैडमिंटन चैम्प: स्ट्रोक का चुनाव इस बात पर निर्भर करता है कि शटलकॉक नेट के कितने नजदीक है, कहीं यह नेट की ऊंचाई से ऊपर तो नहीं है और विरोधी की वर्तमान स्थिति कहां है: अगर वे नेट की ऊंचाई के ऊपर शटलकॉक तक पहुंच सकते हैं तो खिलाड़ी बेहतर हमले की स्थिति में होते हैं। उसी कारण से, बैकहैंड स्मैश कमजोर हो जाते हैं।
बैडमिंटन चैम्प: अलविदा

ERROR ANALYSIS AND SOLUTION FOR BETTERMENT:

->The bot doesn't understand the words semantically.

->Certain errors regarding absolute text matching when more that one occurrence is there in different contexts.

->Anaphora resolution cannot be handled.
Like इस खेल के बारेमें बताओ won't give results
*By making an exhaustive list of pronouns and replacing them with badminton

->If the keyword in the query is not present in the data then the bot won't be able to give the desired output and then would return the standard hardcoded output.
Eg: About Lee Chong Wei

*Including more data on badminton may increase the

->A Lemmatizer in place of stemmer would better the results for example
बारे में
Was searched as बार leading to wrong result

REFERENCES:

1. https://stackabuse.com/python-for-nlp-creating-a-rule-based-chatbot/
2. https://blog.nishtahir.com/2015/09/19/fuzzy-string-matching-using-cosine-similarity/##targetText=It's%20a%20pretty%20popular%20way,sequences%20are%20exactly%20the%20same.
3. https://youtu.be/927YDZH_MLo
4.http://research.variancia.com/hindi_stemmer/

RESEARCH PAPERS :

1. Automatic Generation of Jokes in Hindi

By Srishti Aggarwal and Radhika Mamidi
Analysis by: Anubhav Sharma

Aim : Here we wanted to create a model that would act as a joke generator for some basic jokes of type “Dur se Dekha” in Hindi.
Background : There weren’t many attempts prior to this in the Hindi language of joke generation but in English there were quite a few.

- JAPE (Binsted and Ritchie (1997), Ritchie (2003)) consisting of phonologically ambiguous riddles to generate humor .
- HAHAcronym generator (Stock and Strapparava, 2003) using incongruity theories(where the human brain relates to funnines of finding something which is unexpected by itself or contrasts with its expectation) to generate humor.
- Taylor and Mazlack (2004) worked on a specific category of “ knock-knock” jokes .
- Valitutti et al. (2016) used some taboo words and used it in constrained setting to create humour .

Description :
Dur se Dekha is a focused form of poetic three liner jokes in the Hindi language where first two lines create the setting and the third line (punchline) uses the humour lying in its incongruity .

General Structure:

Part1: Dur se dekha to NP1/VP1 tha,
Part2: Dur se dekha to NP1/VP1 tha,
Part3: Paas jaakar dekha to NP2/VP2 tha.

** Here we will try to create jokes of type 1 by just using either NP or VP.

There are basically two rough categories of these jokes:

1. Using the same part of speech in all the parts.(here using incongruity of the subject).
2. Using different part of speech in Part3 than in Part 1 and Part 2. (remember Part 1 and Part 2 will have the same part of speech and should have some sort of alliteration to create effect of emphasis) (generates humour due to the unconventionality of punchline).

Joke generation consists of 4 processes:

Step1: Template selection- collection of three templates

Step2: Setup Formation-collect and categorize different lexicons(HUMAN ,NON-HUMAN;) on the basis of sentiments associated with it.

Step3: Punchline Formation-We apply certain constraints

- Category constraint
- Gender constraint (shouldn't be from the same category)
- Form constraint (checks for stylistic effects)

Step4: Compilation- Generation was quite fair and data collected from different native speakers revealed the funniness of jokes was quite similar to that of human generated jokes.

Suggestions: (Trying to point out suggestions other than those as mentioned in the paper)

1. Here we should also create a type of category where there are same NP's and VP's but having a different semantic interpretation in each occurrence.
2. Gender constraints may increase the funniness of type 1 but in certain cases when a human imagines a thing it becomes even more funny in case of ambiguous gender.

2. SOUNDING BOARD: A USER-CENTRIC AND CONTENT-DRIVEN SOCIAL CHATBOT

By :HAO FANG, HAO CHENG, MAARTEN SAP, ELIZABETH CLARK, ARI HOLTZMAN, YEJIN CHOI, NOAH A. SMITH, MARI OSTENDORF
Analysis by : Guru Ravi Shanker

Aim : The team from the University of Washington had developed a socialbot named Sounding Board. This is a chatbot with which you can have a coherent and engaging conversation on sports, politics, entertainment, technology, and other popular topics and events.

Additional information : Sounding Board won the inaugural Amazon Alexa Prize in 2017 with an average score of 3.17 out of 5 and an average conversation duration over 10 minutes

PROCEDURE : The system architecture consisted of several components including spoken language processing, dialogue management, language generation, and content management, with emphasis on user-centric and content-driven design. They also shared insights gained from large-scale online logs based on 160,000 conversations with real-world users.

A system produces the response using three modules:
Natural language understanding (NLU) module analyzes the user's speech to produce a representation of the current event.
Dialogue manager (DM) module executes the dialogue's policy while considering user engagement, maintaining dialogue coherence, and enhancing the user experience. DM also has access to the rich content collection that is updated daily.
Natural language generation (NLG) module builds the response using the content selected by the DM.

In addition, the researchers have found that modeling prosody is important for the chatbot to sound more engaging.

My comments :The bot has the ability to say something interesting to the user thanks to the rich content collection as well as the ability to show interest in the conversation partner by acknowledging user's reactions and requests.

Creating a social bot that can have long and engaging conversations with users on a variety of topics is really amazing.

3. PERSONALIZING DIALOGUE AGENTS: I HAVE A DOG, DO YOU HAVE PETS TOO?

By: SAIZHENG ZHANG, EMILY DINAN, JACK URBANEK, ARTHUR SZLAM, DOUWE KIELA, JASON WESTON

Analysis by : Guru Ravi Shanker

Introduction: Chit-chat models are known to have several problems: they lack specificity, do not display a consistent personality and are often not very captivating.

Aim of the paper-

In this work they presented the task of making chit-chat more engaging by conditioning on profile information i.e speak according to person's background which will is essential for improved chatting.

Procedure-

They collected data and trained models to
(i) condition on their given profile information;
They chat will be according to person's information

(ii) information about the person they are talking to, resulting in improved dialogues, as measured by next utterance prediction. Since (ii) is initially unknown our model is trained to engage its partner with personal topics, and we show the resulting dialogue can be used to predict profile information about the interlocutors.

By asking questions which we will get to know the social condition of the person like
A- My wife is from Austria.
B- I have 4 children.
By asking question asked by A we will get the answer related to his family status.

Dataset-

Facebook AI Research team suggests that assigning a personality to the agent will make chit-chat dialogues much more consistent and engaging. To this end, they introduce a PERSONA-CHAT dataset, where each out of 10K dialogues is conditioned on a particular personality. Testing of various baseline models on this dataset shows that models that have access to their own personas are perceived as more consistent by annotators, but not more engaging. However, PERSONA-CHAT appeared to be a very strong source of training data for the beginning of conversations, when the speakers do not know each other and focus on asking and answering questions.

CORE IDEA OF THIS PAPER

Making chit-chat models more engaging and consistent via conditioning on persistent and recognizable profile information.
Suggesting a dataset, collected via Amazon Mechanical Turk, where each of the pairs of speakers conditions their dialogue on a given profile.

ACHIEVEMENT

Introducing a PERSONA-CHAT dataset with:
1155 personality profiles each consisting of at least 5 short sentences;
162,064 utterances over 10,907 dialogues conditioned on some personality profiles.
Taking an important step towards modeling dialogue agents that can ask personality-related questions, remember the answers, and use them naturally in conversations.

My comments-

If they have a virtual personality to talk who knows about their interests ,it will be a very interesting conversation.
Asking questions which will tell us about the person's interest helped in enhancing the bot.

4. Addition of Code Mixed Features to Enhance the Sentiment Prediction of Song Lyrics

By : Rama Rohit Reddy Gangula, Radhika Mamidi
Analysed by : Anubhav Sharma

Aim: Using code-mixed lyrics of a song(which itself is a rich source of datasets containing words that are helpful in analysis and classification of sentiments generated from it) and predict the feeling of arousal that they have on the users since the song selected in a particular mood acts as a good parameter to judge our mood.

** DIFFERENCE BETWEEN VALENCE AND AROUSAL-Valence is positive or negative affectivity, whereas arousal measures how calming or exciting the information is.**

Background:
Sentiment analysis In English:

- Using customer reviews.(Hu and Liu, 2004; Liu, 2015; McGuinness and Ferguson, 2004)

But in Telugu there are a lot of complications with it being a morphologically complex language.
Sentiment analysis In Telugu :

- Song Lyrics [Abburi et al., 2016]
- Reviews[Gangula and Mamidi, 2018]

Very little work is done in the field of code-mixed data of Telugu.

Procedure : Firstly created a Telugu songs dataset which contained both Telugu-English code-mixed and pure Telugu songs.

After mining the lyrics, cleaned them.

Now instead of the traditional Russell's model (which has both the dimensions of valence and arousal) here they dropped the valence parameter as it favours the valence alot and in Telugu movies the songs are mostly occurring in the positive valence so is irrelevant .Annotating also included the context of the situation in the movie when the song was occurring .

Now to extract the code mixing feature they used a CRF model on the certain parameters:

- lexical feature:
    - Word

- sub-lexical features:
    - Prefix, suffix character strings
    - Infix character strings
    - Presence of Post positions
    - Prefix, Suffix character strings of neighboring words
- other features:
    - length of the word
    - neighboring words
    - presence in English dictionary

Then extracted certain statistical features like Number of Telugu-non-Telugu words ,their length etc.

Then implemented the CMNN (code mixed neural network) on the data.The CMNN model itself consisted of 4 parts :

1. **Convolution Layer:**
2. **Long Short-Term Memory (LSTM)-** LSTM outputs a hidden vector h that reflects the semantic representations at position t. To select the final representation of the song lyric, a temporal mean pool is applied to all LSTM outputs
3. **Code mixed features**
4. **Fully-connected Hidden Layer**

The first two steps are the standard algorithms in Machine learning but in step three and four they concatenated the results with the extracted information of code-mixed part. The concatenated vector contained the parameters of the hidden layer, variables for code mixed features and final representation obtained from the temporal mean pooling.

For Experimental setup:

Used 5 fold cross validation and tested a certain lot of techniques :

- Naive Bayes
- Support Vector Machine
- CNN
- LSTM
- CMNN Model

Conclusion:
With their CMNN model they achieved the highest accuracy and it resulted in a hike of about 4 to 5 percent of accuracy as compared to CMNN model without code mixed features.

My Take:
This indeed is an exceptional model to analyse the sentiments but instead of just ignoring the valence aspect they should have taken that into account with proper proportion .Also they should regularly update their dataset(can actually make a web miner searching for data all by itself and piping it through algorithms to get it cleaned) since associated sentiments evolve.