

---

# Computational Linguistics II

## Rule-based Chatbot

**Jashn Arora - 2018114006**

**Tanish Lad - 2018114005**

### Final Report

#### Problem Statement

To Design a Rule-based Chatbot which interacts with the user about any topic in Hindi.

#### Our Approach

After looking through some of the Previous Works on the topic, we decided to follow this procedure:

- The user is first asked about the domain he wants to converse about with the Chatbot.
- Data related to the domain is scraped from Wikipedia. This forms the dataset for the current run of the Chatbot. The screenshot on the next page shows data being scraped from the wikipedia page of Cricket.

मुख्य रूप से इसका प्रशासन दुबई में स्थित अंतर्राष्ट्रीय क्रिकेट परिषद ( आईसीसी ) के द्वारा किया जाता है , जो इसके सदस्य राष्ट्रों के घरेलू नियंत्रित निकायों के माध्यम से विश्व भर में खेल का आयोजन करती है .  
आईसीसी अंतर्राष्ट्रीय स्तर पर खेले जाने वाले पुरुष और महिला क्रिकेट दोनों का नियंत्रण करती है .  
हालांकि पुरुष , महिला क्रिकेट नहीं खेल सकते हैं पर नियमों के अनुसार महिलाएं पुरुषों की टीम में खेल सकती हैं .  
मुख्य रूप से भारतीय उपमहाद्वीप , आस्ट्रेलिया , यूनाइटेड किंगडम , आयरलैंड , दक्षिणी अफ्रीका और वेस्टइंडीज में क्रिकेट का पालन किया जाता है .  
नियम संहिता के रूप में होते हैं जो , क्रिकेट के कानून कहलाते हैं और इनका अनुरक्षण लंदन में स्थित मेरीलेबोन क्रिकेट क्लब ( एम सी सी ) के द्वारा किया जाता है .  
इसमें आई सी सी और अन्य घरेलू बोर्डों का परामर्श भी शामिल होता है .  
क्रिकेट मुकाबला दो दलों ( टीमों ) या पक्षों के बीच खेला जाता है .  
हर टीम में ग्यारह खिलाड़ी होते हैं .  
इसका मैदान कई आकार और आकृतियों का हो सकता है .  
मैदान घास का होता है और इसे ग्राउंड्समैन के द्वारा तैयार किया जाता है , जिसके कार्य में उर्वरण , कटाई , रोलिंग और सतह को समतल करना शामिल होता है .  
मैदान का व्यास 140–160 गज ( 130–150 मी ) सामान्य होता है .  
मैदान की परिधि को सीमा कहा जाता है और इसे कभी कभी रंग दिया जाता है या कभी कभी एक रस्सी के द्वारा मैदान की बाहरी सीमा को चिह्नित किया जाता है .  
मैदान गोल , चौकोर या अंडाकार हो सकता है , क्रिकेट का सबसे प्रसिद्ध मैदान है ओवल .  
प्रत्येक टीम का उद्देश्य होता है दूसरी टीम से अधिक रन बनाना और दूसरी टीम के सभी खिलाड़ियों को आउट करना .  
क्रिकेट में खेल को ज्यादा रन बना कर भी जीता जा सकता है , चाहे दूसरी टीम को पूरी तरह से आउट न किया गया हो .  
दूसरे रूप में खेल को जीतने के लिए अधिक रन बनाना और दूसरी टीम को आउट करना जरूरी होता है , अन्यथा मुकाबला बिना किसी नतीजे के समाप्त हो जाता है .  
खेल शुरू होने से पहले दोनों टीमों के कप्तान एक सिक्के को उछाल करके निर्धारित करते हैं कि कौन सी टीम पहले बल्लेबाजी या गेंदबाजी करेगी .  
टॉस जीतने वाला कप्तान पिच औसत की वर्तमान और प्रत्याशित स्थिति के अनुसार अपना फैसला लेता है .  
मुख्य आकर्षण मैदान के विशेष रूप से तैयार किए गए क्षेत्र में होता है ( आमतौर पर केन्द्र में ) जो \$ पिच \$ कहलाता है .  
पिच के दोनों ओर 22 गज ( 20 मी ) विकेट लगाए जाते हैं .  
ये गेंदबाजी उर्फ क्षेत्ररक्षण पक्ष के लिए लक्ष्य होते हैं और बल्लेबाजी पक्ष के द्वारा इनका बचाव किया जाता है जो रन बनाने की कोशिश में होते हैं .  
मूलतः एक रन तब बनता है जब एक बल्लेबाज गेंद को अपने बल्ले से मारने के बाद पिच के बीच भागता है , हालांकि नीचे बताये गए विवरण के अनुसार रन बनाने के कई और तरीके हैं .  
यदि बल्लेबाज और रन बनाने का प्रयास नहीं करता है तो गेंद \$ देह \$ हो जाती है और गेंदबाज के पास बकिंग गेंदबाजी के लिए आ

- Upon running, the Chatbot first greets the user.

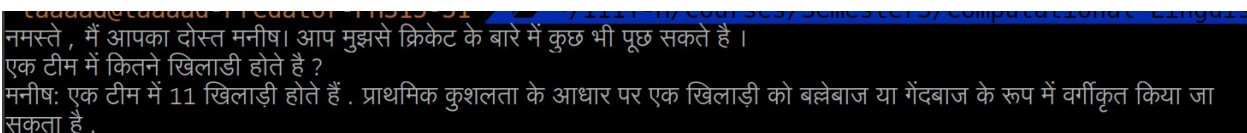
नमस्ते , मैं आपका दोस्त मनीष। आप मुझसे क्रिकेट के बारे में कुछ भी पूछ सकते हैं ।  
धन्यवाद  
मनीष: ये तो मेरा सौभाग्य है।

- The Chatbot asks the user for a query related to the domain the user chose. The user will enter his query.
- The dataset is parsed and each word in every sentence is present in its lemmatized (root) form.
- Each query of the user is parsed and lemmatized in a similar fashion.
- For each query of the user, the lemmatized version of the query is compared with the whole lemmatized dataset iteratively, and the most similar sentence is selected which is found using a variety of Cosine Similarity. The most similar sentence and the next sentence collectively is selected as the response for the query of the user.
- This process will continue until the user signals to exit using the keyword 'नमस्ते'.

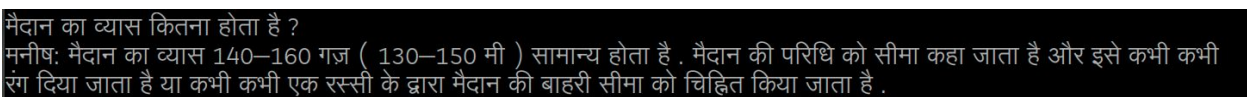
A screenshot of a chatbot interface. The background is black, and the text is in a yellow, slightly pixelated font. The text reads: "नमस्ते" on the first line, and "मनीष: अलविदा दोस्त, आपका दिन शुभ हो।" on the second line. The text is centered and appears to be a response from the chatbot.

## Results

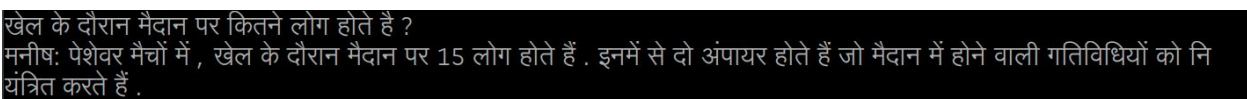
For demonstration, क्रिकेट (Cricket) is selected as the domain.

- A screenshot of a chatbot interface. The user asks: 'नमस्ते, मैं आपका दोस्त मनीष। आप मुझसे क्रिकेट के बारे में कुछ भी पूछ सकते हैं। एक टीम में कितने खिलाड़ी होते हैं?' The chatbot replies: 'मनीष: एक टीम में 11 खिलाड़ी होते हैं। प्राथमिक कुशलता के आधार पर एक खिलाड़ी को बल्लेबाज या गेंदबाज के रूप में वर्गीकृत किया जा सकता है।'

The user asks about the number of players in a team. The bot correctly replies that there are 11 players in a team.

- A screenshot of a chatbot interface. The user asks: 'मैदान का व्यास कितना होता है?' The chatbot replies: 'मनीष: मैदान का व्यास 140–160 गज ( 130–150 मी ) सामान्य होता है . मैदान की परिधि को सीमा कहा जाता है और इसे कभी कभी रंग दिया जाता है या कभी कभी एक रस्सी के द्वारा मैदान की बाहरी सीमा को चिह्नित किया जाता है .'

The user asks about the diameter of the ground. The chatbot correctly replies that its about 130-150 meters.

- A screenshot of a chatbot interface. The user asks: 'खेल के दौरान मैदान पर कितने लोग होते हैं?' The chatbot replies: 'मनीष: पेशेवर मैचों में , खेल के दौरान मैदान पर 15 लोग होते हैं . इनमें से दो अंपायर होते हैं जो मैदान में होने वाली गतिविधियों को नियंत्रित करते हैं .'

The user asks about the number of people present on the ground at an instant. The bot replies 15 (2 batsmen + 2 umpire + 11 players from opposite team)

## Issues

- Chunking and then lemmatizing is a very slow method for finding the root of a word.
- The domain user inputs should be a valid page on Wikipedia.
- The parser can only be run on IIIT Network or on IIIT VPN/Proxy.
- Cosine Similarity is not a very good method of finding the best output for the user query
- Rule-based approach does not always give the best response.

## Future Work

We plan on Running a Language Model to extract the best output.

## GitHub Repository

Checkout our GitHub Repository on the work:

<https://github.com/destinyson7/Chatbot-Hindi>

## Contributions of each member

We both read all the papers together. We both coded all the codes together on a single laptop.

## References

- [1] M.Dahiya. ***A Tool of Conversation: Chatbot (2017)***
- [2] Ananthakrishnan Ramanathan, Durgesh D Rao. ***A Lightweight Stemmer for Hindi (2003)***
- [3] Lovely Sharma, Vijay Dhir and Kamaljeet Kaur. ***A New Model for Question-Answer based Dialogue System for Indian Railways in Hindi Language (2015)***
- [4] Nick Webb. ***Rule-Based Dialogue Management Systems (2001)***