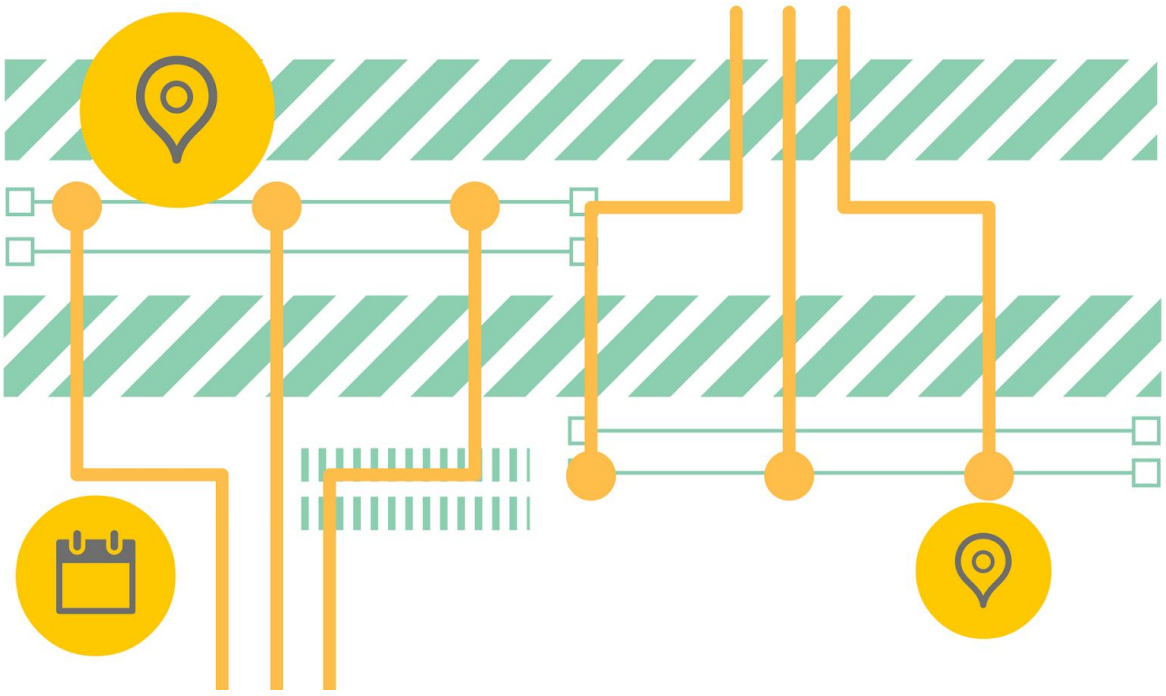

Domain Entity Identification

03.03.2020

Sneks

Jerrin John Thomas - 2019114012 - jerrin.thomas@research.iiit.ac.in

Mayank Goel - 2019114004 - mayank.goel@research.iiit.ac.in

Tanishq Chaudhary - 2019114007 - tanishq.chaudhary@research.iiit.ac.in

Overview

This project is going to be developed as part of the completion of the Computational Linguistics-I course, IIIT-Hyderabad (2020). The intent is to build a Named Entity Recogniser for the domain: FIFA World Cup 2018.

Goals

The objective is to identify and extract frequent Entities for different domains. The project is expected to be completed within one and a half months.

Solution

The current ideas being considered are using a statistical approach, optimally using Machine Learning algorithms from the Scikit-Learn library. More specifically, the idea of working with Neural Networks is being discussed.

No data will need to be manually tagged for testing as the list of players, teams, and locations are already available, which will be helpful for testing the models for accuracy. Also, data will be easier to scrape from Twitter using hashtags.

At the very least, a rule-based Named Entity Recogniser will be developed if things do not go according to plan.

Specifications

Details regarding the organisation of the team, resources, and tools used will be decided upon meeting with the mentor. Change during project execution and risks will be minimised to the extent possible. All solutions will be adequately researched and the best one will be implemented. We are specifically choosing one domain for the benefit of higher accuracy.

Milestones

Phase 1: Scrape data from Twitter, and create a database of sentences using hashtags. Also we plan to compile a list of testing data of players and teams.

Phase 2: We will Part Of Speech tag, and use a logistic regression algorithm for classification. Further details on the specific statistical approach are not confirmed with the possibility of a rule-based approach remains.

Phase 3: ???

Phase 4: Make an accurate classifier for FIFA World Cup 2018 and write a research paper using LaTeX.

Benefits

The project will benefit the team as it is an area which is looked forward to dive into. It will help the team to apply the basic understanding of language analysis, based on linguistic theories. It will familiarise the team with the linguistic challenges in processing natural languages, analysing the structures and dealing with ambiguities in language. It will help to understand and analyse actual language data and develop computational resources for various levels of language structures.

In the real world, such classification tools for specialised domains can help in analysing trends and popularity of players (or products). Also, selecting a domain where we have domain expertise can yield a lot of benefits.