Innoplexus Hackathon

# Problem Statement:

* In this competition we had to extract entities, from the tweets posted by various users, which were an indication towards a probable disease.
* We were given a labelled dataset in which each individual from a tweet was BIO tagged.

# Result:

* I finished top of the leaderboard out of 1998 participants.
* Click [here](https://datahack.analyticsvidhya.com/contest/innoplexus-online-hiring-hackathon-saving-lives-wi/#lb) for the link to the competition’s leaderboard.

# Approach:

1. I used **Conditional Random Field Classifier (CRF)** to solve this challenge.
2. The library which I used for this purpose is **Stranford CoreNLP**.
3. This library is written in java and has excellent implementation of all the relevant NLP tasks such as tokenization, lemmatization, POS tagging, NER etc.

# Preprocessing:

1. For data preprocessing, I removed all the documents which didn’t contain any INDICATION tags.
2. This step largely reduced the error and gave me satisfactory results.
3. All the final parameters which I used to train the model are mentioned in Properties.txt file.
4. While training my model I didn’t include BIO annotations as these made my model to produce inaccurate results like generating ‘B’ annotation in between two ‘I’ annotations.
5. As a result, the final model only has two classes to classify ‘INDICATIONS’ and ‘O’.
6. Once generating my predictions from the model, I later converted them into BIO tags.