This is a research heavy course.
Now let's begin with today's class.
—

We talk about NLP(Natural Language Processing):

Grammar is the most basic thing you can do in the language.
(noun, adjective…..)
This is called part of speech tagging.
This is classification task in SML(supervised ML)
One difference between this and ordinary classification tasks is that each word of a sentence works in collaboration with each other.

BHAINSAA == Buffalo
Buffalo means city and an animal and an verb
Buffalo Buffalo Buffalo. (this is an legitimate english sentence)
Similar poem in chinese.

Therefore words can have different taggings. We have algorithms which work with 95% accuracy.

Breaking the sentences and words into mind is also important. There is “ “ (space) between two words but not all. Like Shouldn't and wouldn't.
Segmentation and Breaking.

Named entity is anything that has a name in real life. (This could be a phrase).(Multiple names in a sentence)

Unlike a part of speech tagging, each name in the sentence is ‘Named Entity Tagging’.

Relation Detection: Detects semantic relations b/w entities (best till date is 70 - 80 % accuracy)
Task is a sub phrase in the text referring to the phrase itself.

Scikit learn is generic library
But to do all of the above we need to use CoreNLP.

In earlier times the newspaper was DATA.
We need better quality data.

Finding Complex piece of language is good data for ML to figure out how humans talk to each other.

Link to Coding in class: Click below

Week4_Day1_Coding_Part.ipynb