Named Entity Recognition

Subtitle

Ramaseshan Ramachandran

NAMED ENTITY RECOGNITION

- Sequence Labelling
  - Parts of Speech – Noun, proper noun, verb, pronoun, preposition, adverb, conjunction, participle, and article
  - Proper names – names of person, city, country, state, organisation – or anything that is always spelt in the Capitalised form
  - The task of assigning a tag to the the spans of text that constitute proper names

APPLICATIONS

- Chats
- Tagging the customer's sentiments towards the objects/products
- Linking knowledge sources - Given the term "Himalaya", link the knowledge sources related to this

PROMPT - HIMALAYA

Geographical Grandeur (from bard.google.com)

- The Himalayas, nicknamed "the roof of the world," stretch for over 2,400 km across eight countries: Pakistan,India, Nepal, Bhutan, China, Myanmar, Bangladesh, and Afghanistan.
- Comprising the world's highest peaks, including Mount Everest (8,848.86 m), K2 (8,611 m), Kangchenjunga (8,586 m), Lhotse (8,516 m), and Makalu (8,463 m), these majestic mountains form a formidable barrier between the Indian subcontinent and the Tibetan Plateau.
- The region is further enriched by numerous glaciers, valleys, rivers, and lush forests, creating a breathtaking landscape of diverse ecosystems.

PROMPT - HIMALAYA +1

- The Himalayas cradle a rich tapestry of cultures and traditions, shaped by ancient civilizations and diverse ethnicities.
- From the Buddhist stupas and prayer flags of Nepal to the vibrant Hindu temples of India and the unique customs of Bhutan, the region pulsates with spiritual energy and historical significance.
- Local communities have adapted their lives to the challenging mountain environment, developing unique farming practices, traditional architecture, and vibrant festivals
- Additional information on Cultural Tapestry, tourism, adventure and environmental challenges are also provided

CHALLENGES

- Ambiguity
  - Chidambaram is a person or the name of a town?
  - Cauvery – name of a person or river
  - Apple – name of the company or fruit?
  - JFK – Airport or John F Kennedy

POS

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative

LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer

POS

POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol

TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present

POS

VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

SIMPLIFIED VERSION OF POS

Part-of-speech tag	Description
UKW	Unknown word
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
IN	Preposition or subordinating conjunction
JJ	Adjective
MD	Modal
NN	Noun
NNP	Proper noun
PRP	Pronoun
QT	Quantifier
RB	Adverb
SYM	Symbol, including all types of punctuation
UH	Interjection
VB	Verb
WH	Wh-word, such as the equivalent of what

POS USING NLTK

```
mport nltk
text = '''Infosys is expected to announce its
        third-quarter financial results today.
        Analysts are predicting a drop in revenue and
        margins amid the ongoing IT slowdown'''
sentence = nltk.sent_tokenize(text)
for sent in sentence:
    print(nltk.pos_tag(nltk.word_tokenize(sent))
```

```
[('Infosys', 'NNP'), ('is', 'VBZ'), ('expected', 'VBN'), ('to', 'TO'),
('announce', 'VB'), ('its', 'PRP$'), ('third-quarter', 'JJ'), ('financial', 'JJ'),
('results', 'NNS'), ('today', 'NN'), ('.', '.')]
[('Analysts', 'NNS'), ('are', 'VBP'), ('predicting', 'VBG'), ('a', 'DT'), ('drop',
'NN'), ('in', 'IN'), ('revenue', 'NN'), ('and', 'CC'), ('margins', 'NNS'), ('amid',
'IN'), ('the', 'DT'), ('ongoing', 'VBG'), ('IT', 'NNP'), ('slowdown', 'NN')]
```

MODELS

- HMM
- CRF

What do we need?

- Training Samples – tagged corpus
- Test Corpus

MAXIMUM LIKELIHOOD APPROACH

- Make simplifying assumptions using Markov assumptions
  - The Probability of a state depends only on the previous states
  - The probability of an output observation o_i depends only on the state that produced the observation
  - Example – I will wake up early tomorrow
    - **Will** is likely to be followed by a **verb**
    - Compute the maximum likelihood estimate of this transition probability

$$P(t_i \mid t_{i-1}) = \frac{Count(t_{i-1}, t_i)}{Count(t_{i-1})}$$

$$P(MD \mid VB) = \frac{Count(MD, VB)}{count(MD)}$$

We will also estimate what is most likely word as MD