Building a TIMEX3 Tag Identifier Model for Hindi

FACULTY: DR. MANISH SHRIVASTAVA

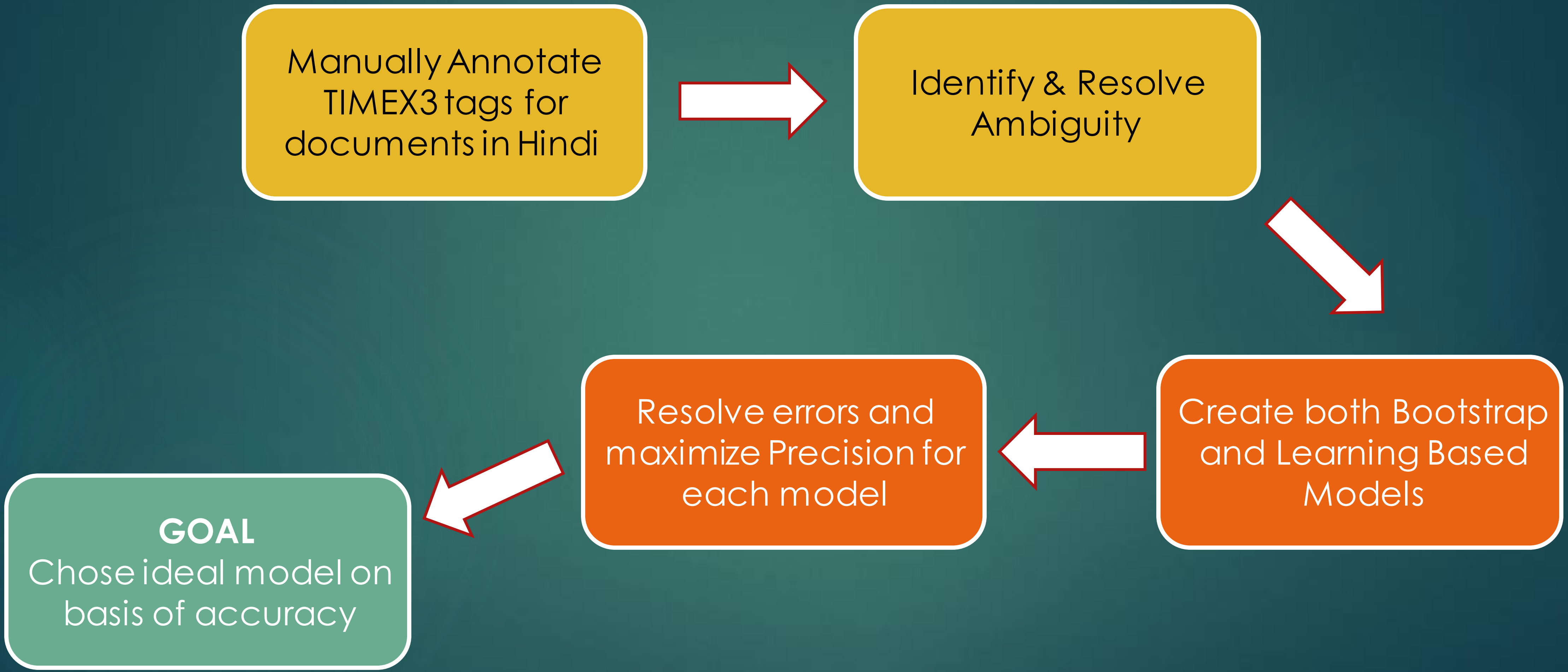
MENTOR: JAIPAL SINGH GOUD, PRANAV GOEL, SUHAN PRABHU, ALOK DEBNATH

TimeML is a markup language that was conceptualized in 2002 for recognition of temporal events in a document.

It essentially introduced a bunch of **tagsets such as EVENT, TLINK, TIMEX3 etc.**

The has been a **whole body of work** dealing with TIMEX3 and temporal event recognition in English

Our goal for IASNLP is to simply develop a **model that can identify TIMEX3 tags** from a document **in Hindi**



Manually Annotate TIMEX3 tags for documents in Hindi

Identify & Resolve Ambiguity

Create both Bootstrap and Learning Based Models

Resolve errors and maximize Precision for each model

GOAL

Chose ideal model on basis of accuracy

TIMEX3 Annotations for Hindi

Attributes:

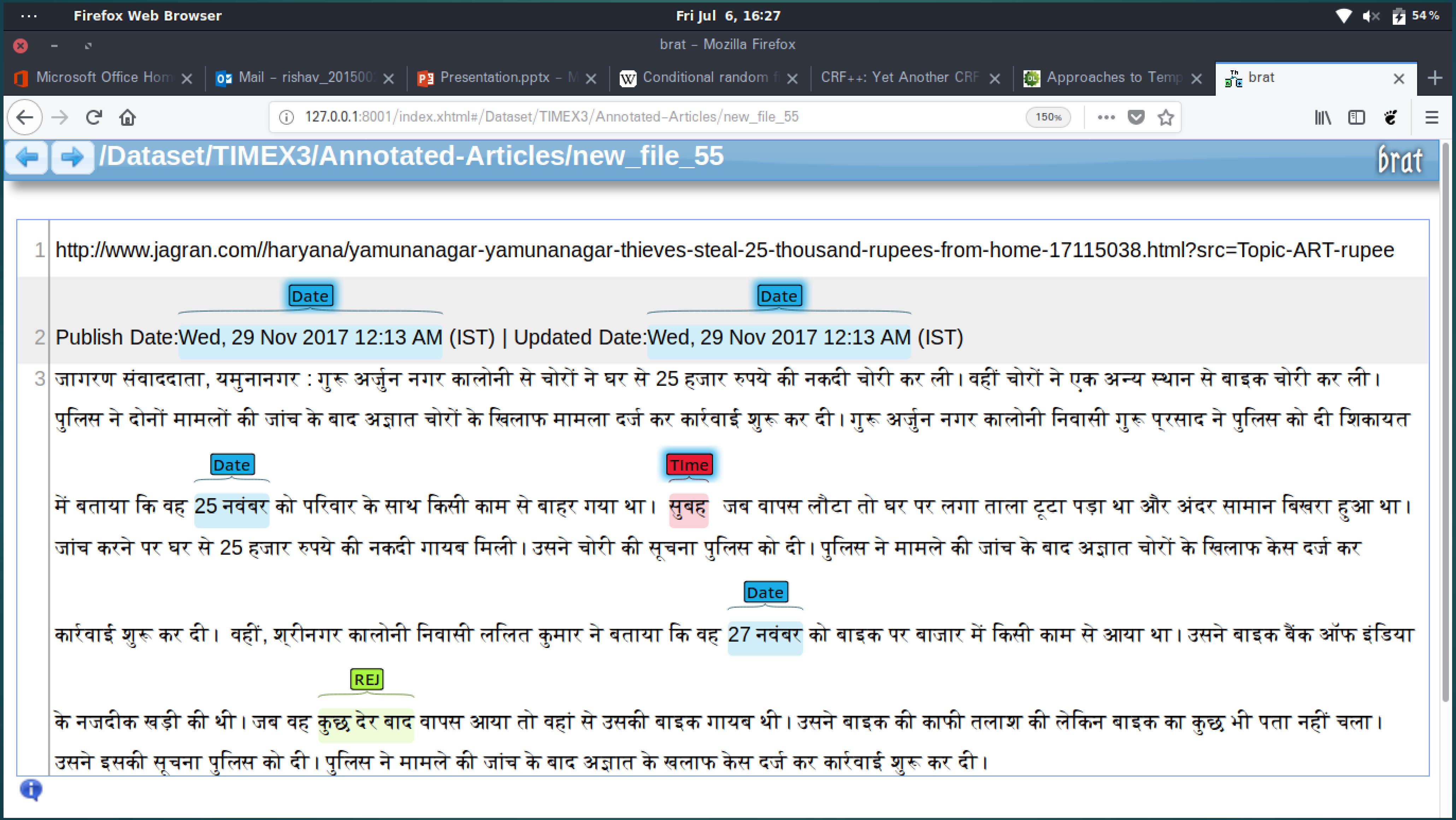
- Date-Time : Describes specific calendar time.
- Period : Describes a duration
- Frequency: Describes a set of times.

ILTIMEX corpus

For tagging we have used ILTIMEX corpus. It was published in the paper

"**Approaches to Temporal Expression Recognition in Hindi**".

Brat Annotation Tool

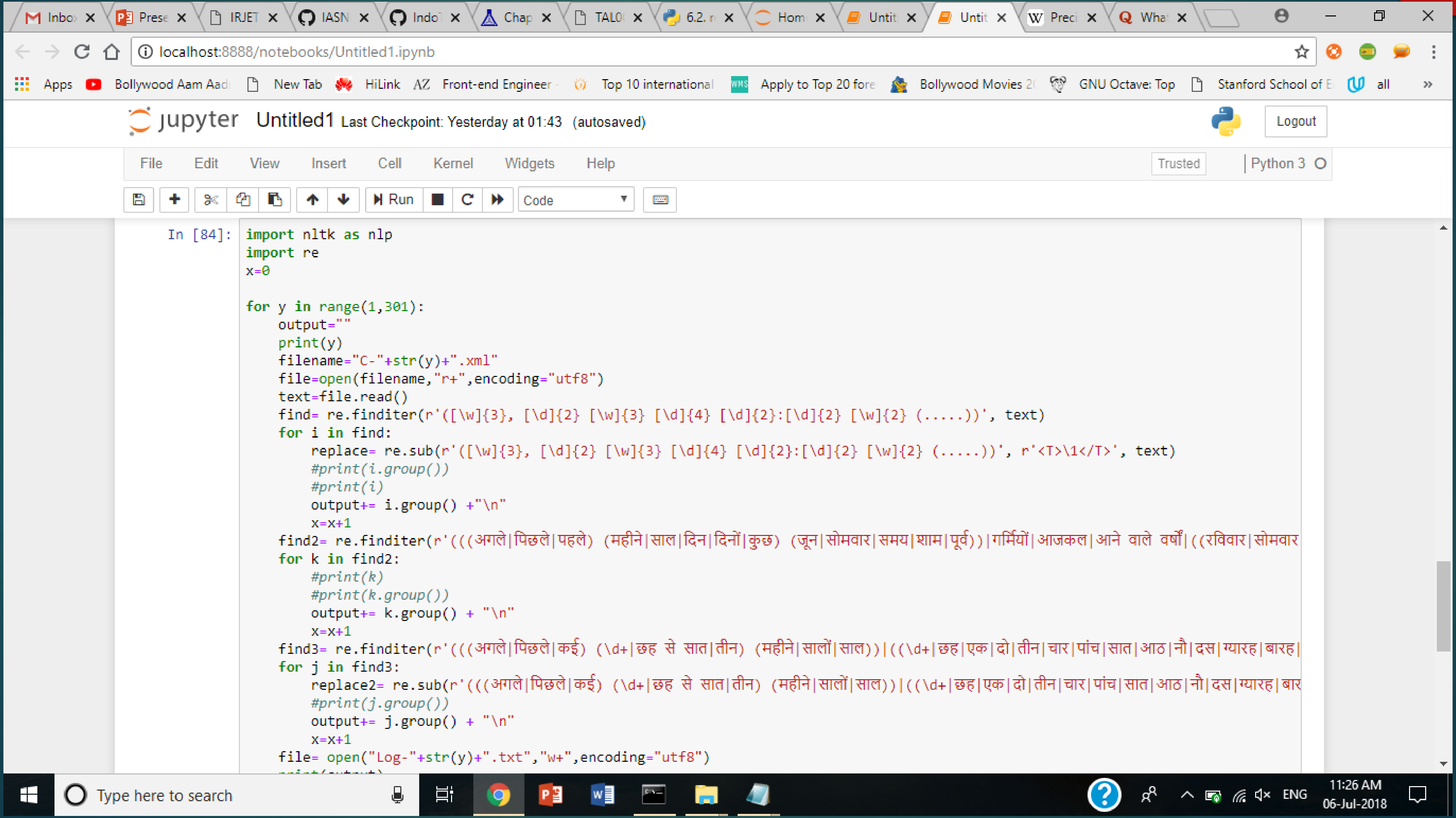


1 http://www.jagran.com//haryana/yamunanagar-yamunanagar-thieves-steal-25-thousand-rupees-from-home-17115038.html?src=Topic-ART-rupee

2 Publish Date:Wed, 29 Nov 2017 12:13 AM (IST) | Updated Date:Wed, 29 Nov 2017 12:13 AM (IST)

3 जागरण संवाददाता, यमुनानगर : गुरू अर्जुन नगर कालोनी से चोरों ने घर से 25 हजार रुपये की नकदी चोरी कर ली। वहीं चोरों ने एक अन्य स्थान से बाइक चोरी कर ली। पुलिस ने दोनों मामलों की जांच के बाद अज्ञात चोरों के खिलाफ मामला दर्ज कर कार्रवाई शुरू कर दी। गुरू अर्जुन नगर कालोनी निवासी गुरू प्रसाद ने पुलिस को दी शिकायत में बताया कि वह 25 नवंबर को परिवार के साथ किसी काम से बाहर गया था। सुबह जब वापस लौटा तो घर पर लगा ताला टूटा पड़ा था और अंदर सामान बिखरा हुआ था। जांच करने पर घर से 25 हजार रुपये की नकदी गायब मिली। उसने चोरी की सूचना पुलिस को दी। पुलिस ने मामले की जांच के बाद अज्ञात चोरों के खिलाफ केस दर्ज कर कार्रवाई शुरू कर दी। वहीं, श्रीनगर कालोनी निवासी ललित कुमार ने बताया कि वह 27 नवंबर को बाइक पर बाजार में किसी काम से आया था। उसने बाइक बैंक ऑफ इंडिया के नजदीक खड़ी की थी। जब वह कुछ देर बाद वापस आया तो वहां से उसकी बाइक गायब थी। उसने बाइक की काफी तलाश की लेकिन बाइक का कुछ भी पता नहीं चला। उसने इसकी सूचना पुलिस को दी। पुलिस ने मामले की जांच के बाद अज्ञात के खलाफ केस दर्ज कर कार्रवाई शुरू कर दी।

Python Notebook with Regex for detecting.

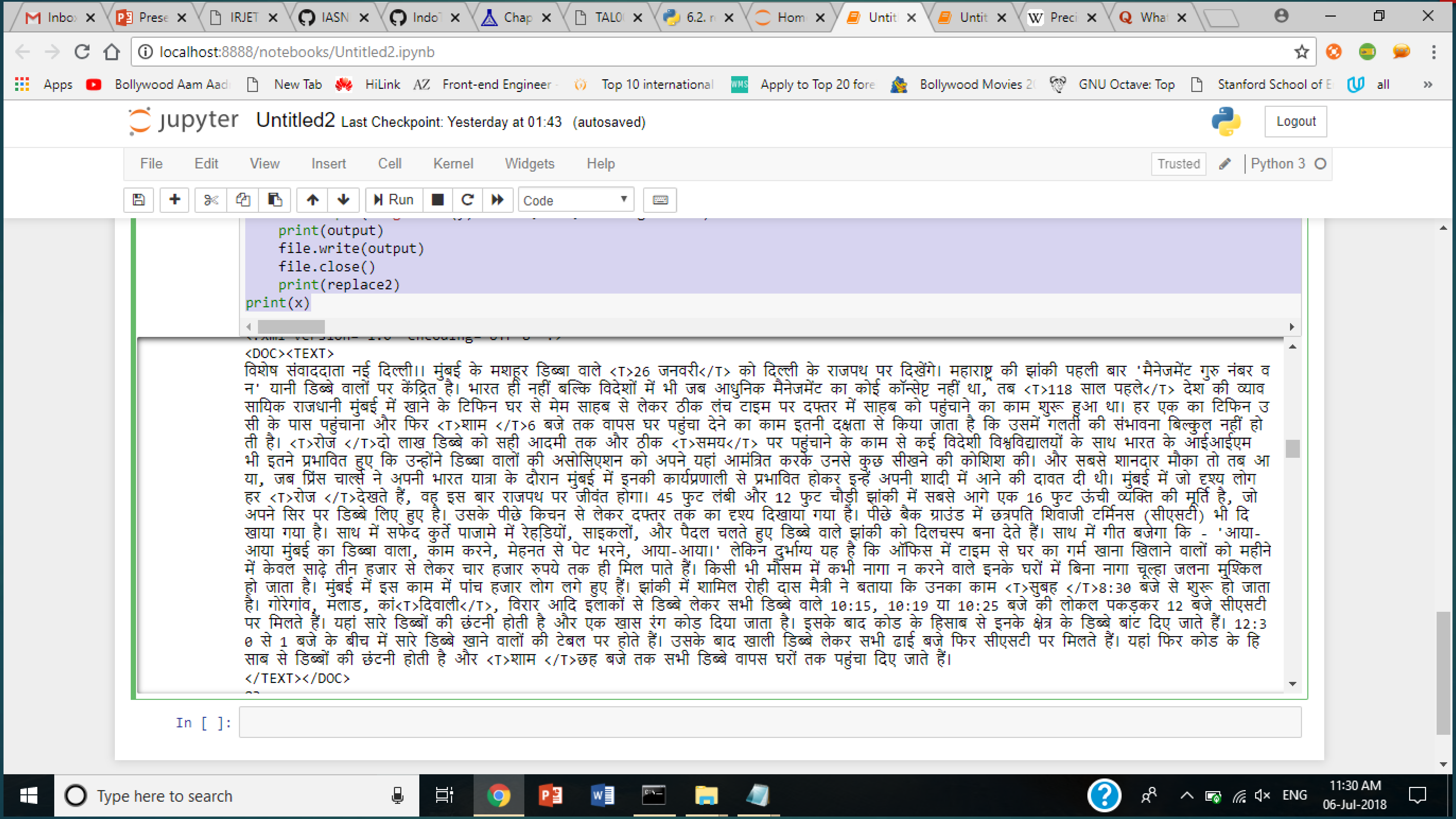


localhost:8888/notebooks/Untitled1.ipynb
jupyter Untitled1 Last Checkpoint: Yesterday at 01:43 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
11:26 AM
06-Jul-2018

```
In [84]: import nltk as nlp
         import re
         x=0

         for y in range(1,301):
             output=""
             print(y)
             filename="C-"+str(y)+".xml"
             file=open(filename,"r+",encoding="utf8")
             text=file.read()
             find= re.finditer(r'([\w]{3}, [\d]{2} [\w]{3} [\d]{4} [\d]{2}:[\d]{2} [\w]{2} (.....))', text)
             for i in find:
                 replace= re.sub(r'([\w]{3}, [\d]{2} [\w]{3} [\d]{4} [\d]{2}:[\d]{2} [\w]{2} (.....))', r'<T>\1</T>', text)
                 #print(i.group())
                 #print(i)
                 output+= i.group() +"\n"
                 x=x+1
             find2= re.finditer(r'(((अगले|पिछले|पहले) (महीने|साल|दिन|दिनों|कुछ) (जून|सोमवार|समय|शाम|पूर्व))|गर्मियों|आजकल|आने वाले वर्षों|((रविवार|सोमवार
             for k in find2:
                 #print(k)
                 #print(k.group())
                 output+= k.group() + "\n"
                 x=x+1
             find3= re.finditer(r'(((अगले|पिछले|कई) (\d+|छह से सात|तीन) (महीने|सालों|साल))|((\d+|छह|एक|दो|तीन|चार|पांच|सात|आठ|नौ|दस|ग्यारह|बारह|
             for j in find3:
                 replace2= re.sub(r'(((अगले|पिछले|कई) (\d+|छह से सात|तीन) (महीने|सालों|साल))|((\d+|छह|एक|दो|तीन|चार|पांच|सात|आठ|नौ|दस|ग्यारह|बार
                 #print(j.group())
                 output+= j.group() + "\n"
                 x=x+1
             file= open("Log-"+str(y)+".txt","w+",encoding="utf8")
```

Extracted Text

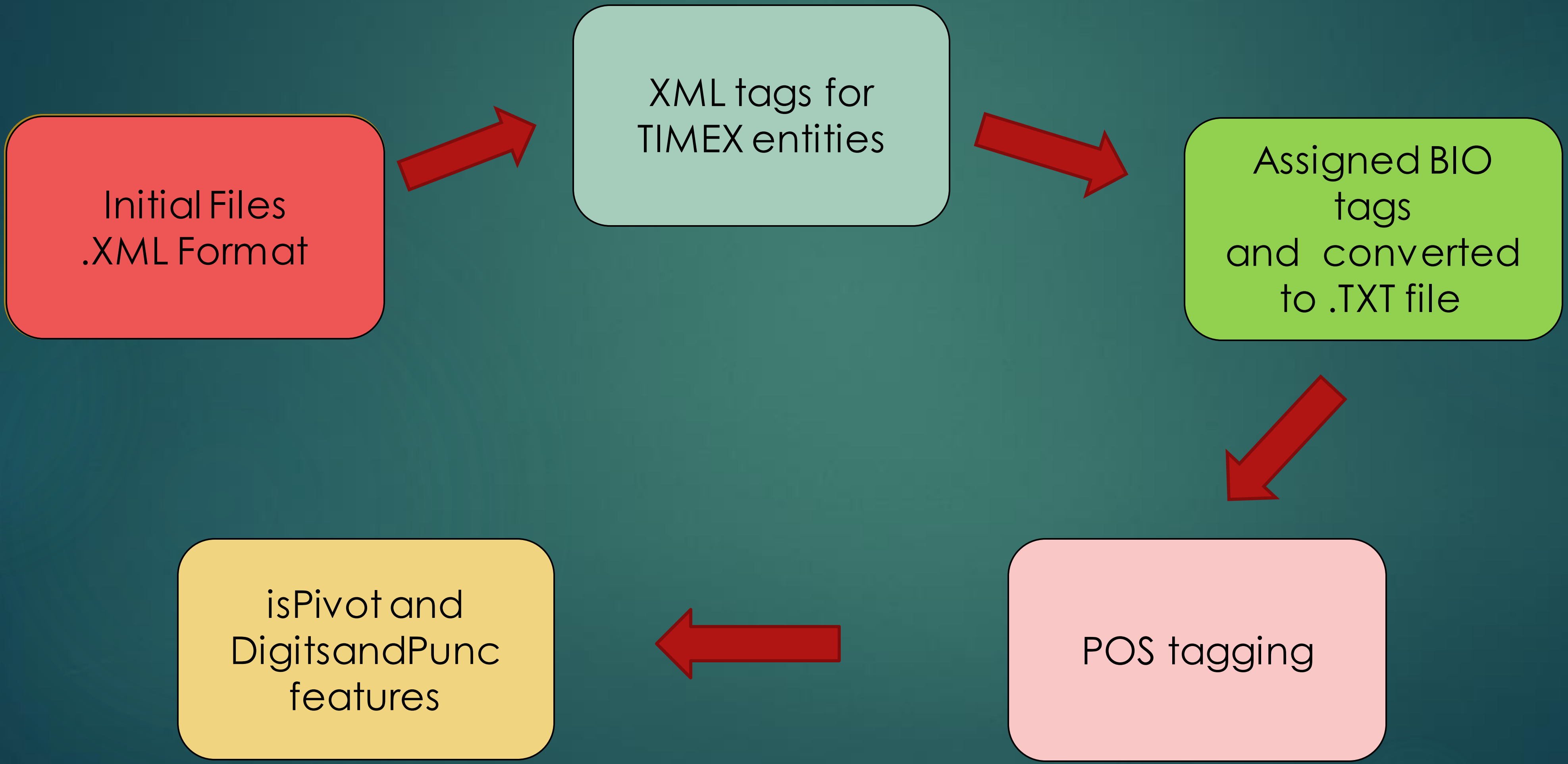

localhost:8888/notebooks/Untitled2.ipynb
jupyter Untitled2 Last Checkpoint: Yesterday at 01:43 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
ENG
11:30 AM
06-Jul-2018

```
    print(output)
    file.write(output)
    file.close()
    print(replace2)
print(x)
```

```
<?xml version="1.0" encoding="utf-8"?>
<DOC><TEXT>
विशेष संवाददाता नई दिल्ली।। मुंबई के मशहूर डिब्बा वाले <T>26 जनवरी</T> को दिल्ली के राजपथ पर दिखेंगे। महाराष्ट्र की झांकी पहली बार 'मैनेजमेंट गुरु नंबर व
न' यानी डिब्बे वालों पर केंद्रित है। भारत ही नहीं बल्कि विदेशों में भी जब आधुनिक मैनेजमेंट का कोई कॉन्सेप्ट नहीं था, तब <T>118 साल पहले</T> देश की व्याव
सायिक राजधानी मुंबई में खाने के टिफिन घर से मेम साहब से लेकर ठीक लंच टाइम पर दफ्तर में साहब को पहुंचाने का काम शुरू हुआ था। हर एक का टिफिन उ
सी के पास पहुंचाना और फिर <T>शाम </T>6 बजे तक वापस घर पहुंचा देने का काम इतनी दक्षता से किया जाता है कि उसमें गलती की संभावना बिल्कुल नहीं हो
ती है। <T>रोज </T>दो लाख डिब्बे को सही आदमी तक और ठीक <T>समय</T> पर पहुंचाने के काम से कई विदेशी विश्वविद्यालयों के साथ भारत के आईआईएम
भी इतने प्रभावित हुए कि उन्होंने डिब्बा वालों की असोसिएशन को अपने यहां आमंत्रित करके उनसे कुछ सीखने की कोशिश की। और सबसे शानदार मौका तो तब आ
या, जब प्रिंस चार्ल्स ने अपनी भारत यात्रा के दौरान मुंबई में इनकी कार्यप्रणाली से प्रभावित होकर इन्हें अपनी शादी में आने की दावत दी थी। मुंबई में जो दृश्य लोग
हर <T>रोज </T>देखते हैं, वह इस बार राजपथ पर जीवंत होगा। 45 फुट लंबी और 12 फुट चौड़ी झांकी में सबसे आगे एक 16 फुट ऊंची व्यक्ति की मूर्ति है, जो
अपने सिर पर डिब्बे लिए हुए है। उसके पीछे किचन से लेकर दफ्तर तक का दृश्य दिखाया गया है। पीछे बैक ग्राउंड में छत्रपति शिवाजी टर्मिनस (सीएसटी) भी दि
खाया गया है। साथ में सफेद कुर्ते पाजामे में रेहड़ियों, साइकलों, और पैदल चलते हुए डिब्बे वाले झांकी को दिलचस्प बना देते हैं। साथ में गीत बजेगा कि - 'आया-
आया मुंबई का डिब्बा वाला, काम करने, मेहनत से पेट भरने, आया-आया।' लेकिन दुर्भाग्य यह है कि ऑफिस में टाइम से घर का गर्म खाना खिलाने वालों को महीने
में केवल साढ़े तीन हजार से लेकर चार हजार रुपये तक ही मिल पाते हैं। किसी भी मौसम में कभी नागा न करने वाले इनके घरों में बिना नागा चूल्हा जलना मुश्किल
हो जाता है। मुंबई में इस काम में पांच हजार लोग लगे हुए हैं। झांकी में शामिल रोही दास मैत्री ने बताया कि उनका काम <T>सुबह </T>8:30 बजे से शुरू हो जाता
है। गोरेगांव, मलाड, कां<T>दिवाली</T>, विरार आदि इलाकों से डिब्बे लेकर सभी डिब्बे वाले 10:15, 10:19 या 10:25 बजे की लोकल पकड़कर 12 बजे सीएसटी
पर मिलते हैं। यहां सारे डिब्बों की छंटनी होती है और एक खास रंग कोड दिया जाता है। इसके बाद कोड के हिसाब से इनके क्षेत्र के डिब्बे बांट दिए जाते हैं। 12:3
0 से 1 बजे के बीच में सारे डिब्बे खाने वालों की टेबल पर होते हैं। उसके बाद खाली डिब्बे लेकर सभी ढाई बजे फिर सीएसटी पर मिलते हैं। यहां फिर कोड के हि
साब से डिब्बों की छंटनी होती है और <T>शाम </T>छह बजे तक सभी डिब्बे वापस घरों तक पहुंचा दिए जाते हैं।
</TEXT></DOC>
```

In [ ]:

Data Pre-Processing



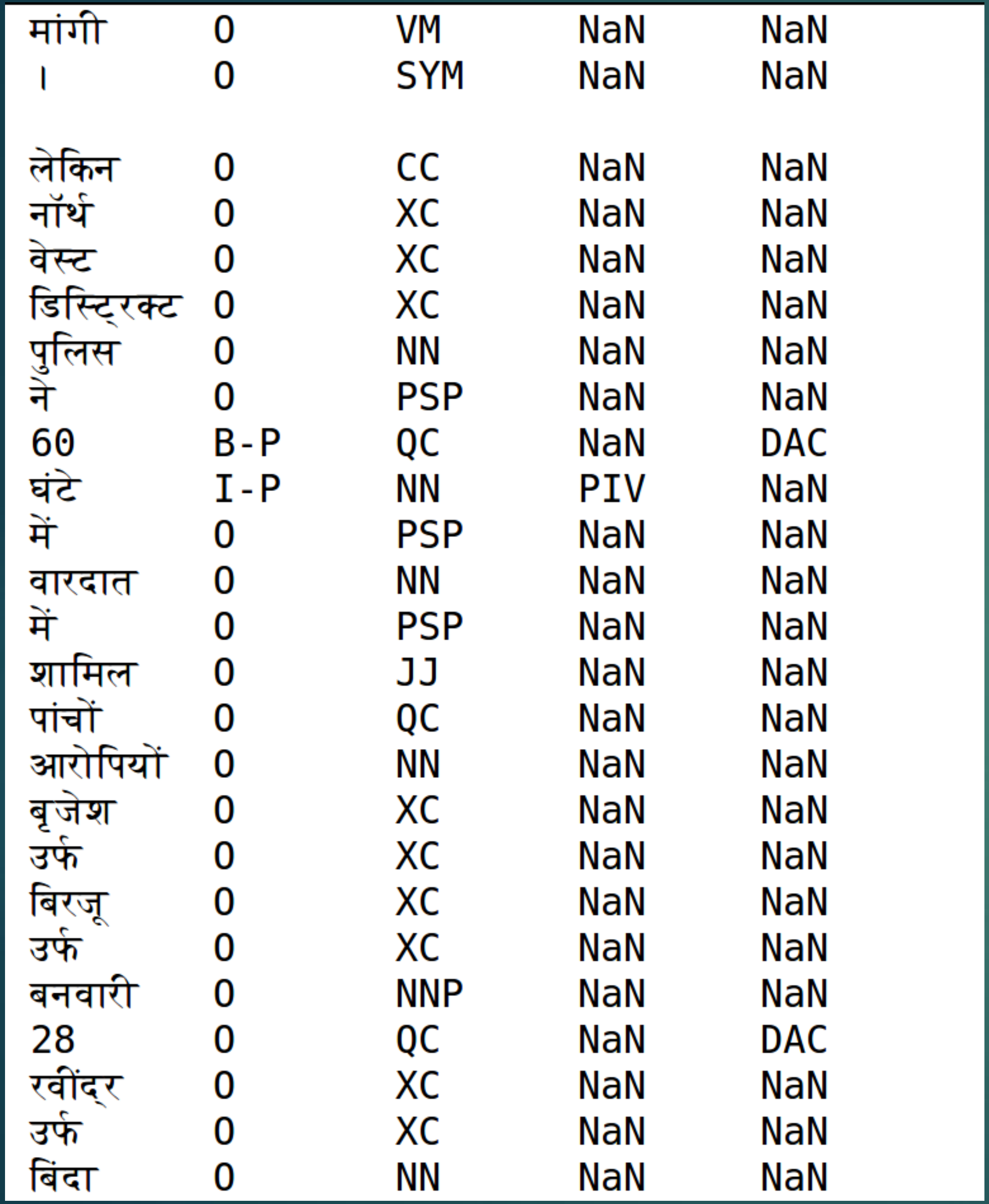
Initial Files
.XML Format

XML tags for
TIMEX entities

Assigned BIO
tags
and converted
to .TXT file

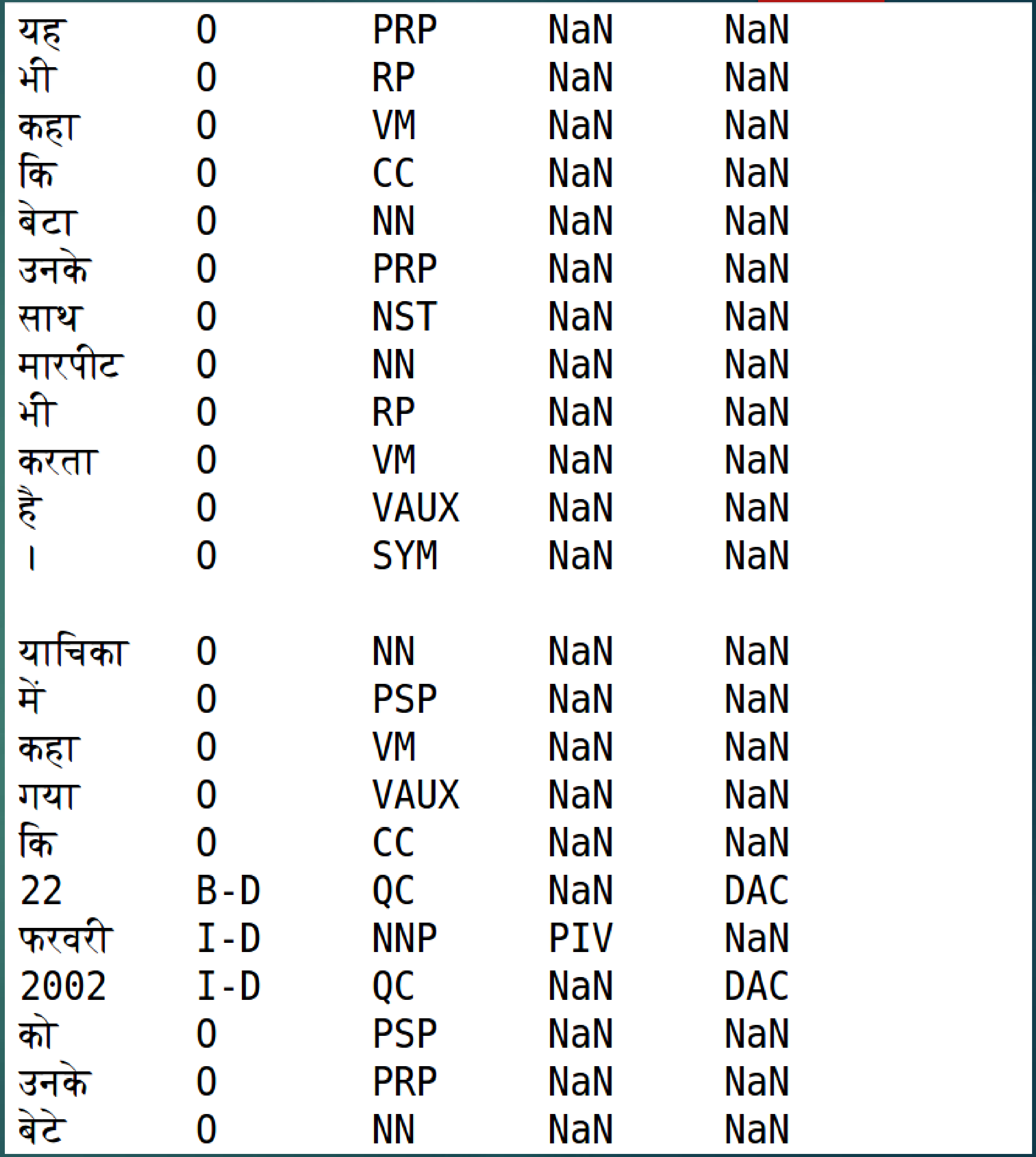
POS tagging

isPivot and
DigitsandPunc
features



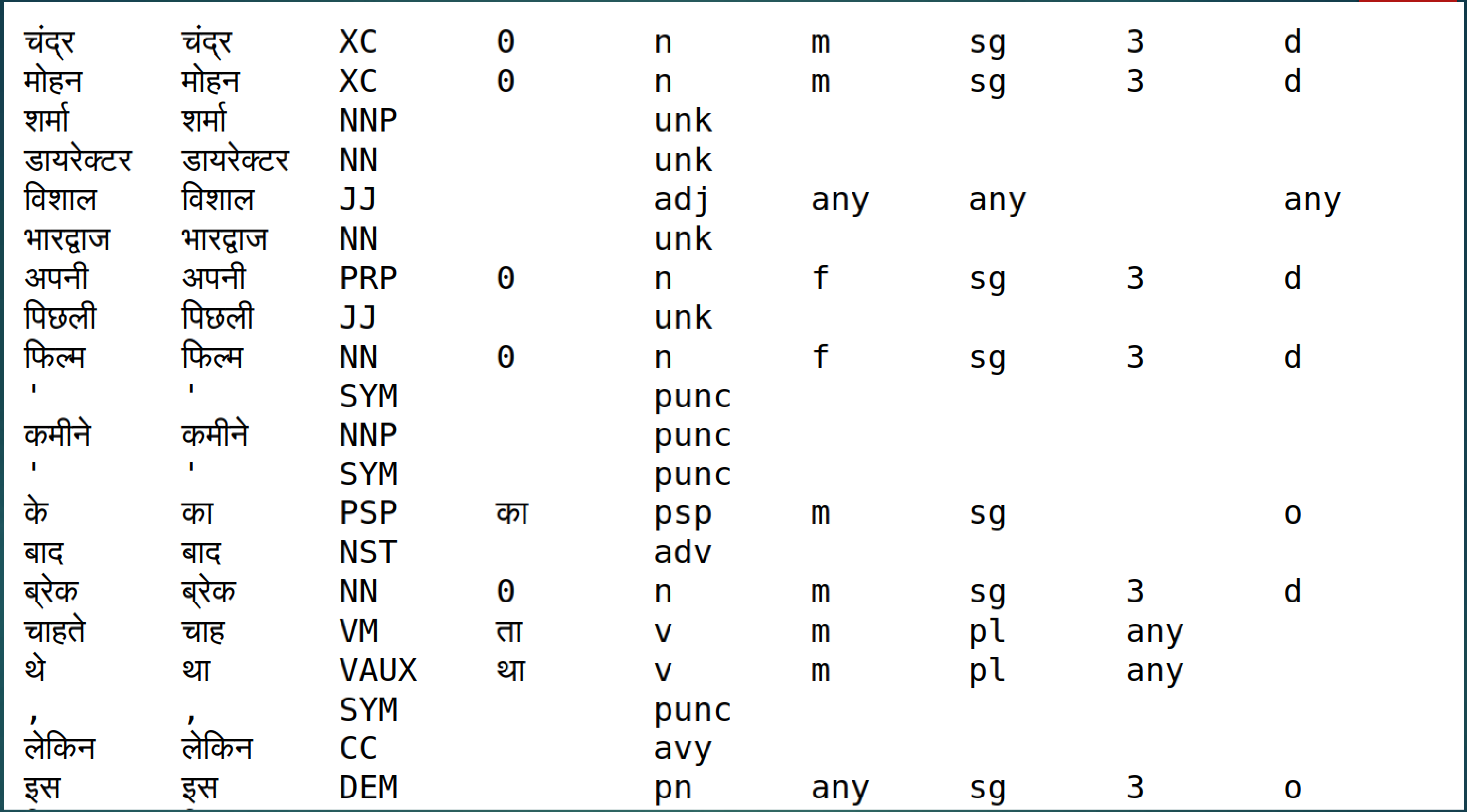
```
मांगी        O      VM     NaN    NaN
।            O      SYM    NaN    NaN

लेकिन        O      CC     NaN    NaN
नॉर्थ          O      XC     NaN    NaN
वेस्ट         O      XC     NaN    NaN
डिस्ट्रिक्ट     O      XC     NaN    NaN
पुलिस        O      NN     NaN    NaN
ने           O      PSP    NaN    NaN
60           B-P    QC     NaN    DAC
घंटे          I-P    NN     PIV    NaN
में           O      PSP    NaN    NaN
वारदात       O      NN     NaN    NaN
में           O      PSP    NaN    NaN
शामिल        O      JJ     NaN    NaN
पांचों         O      QC     NaN    NaN
आरोपियों      O      NN     NaN    NaN
बृजेश         O      XC     NaN    NaN
उर्फ          O      XC     NaN    NaN
बिरजू         O      XC     NaN    NaN
उर्फ          O      XC     NaN    NaN
बनवारी       O      NNP    NaN    NaN
28           O      QC     NaN    DAC
रवींद्र        O      XC     NaN    NaN
उर्फ          O      XC     NaN    NaN
बिंदा         O      NN     NaN    NaN
```



```
यह           O      PRP    NaN    NaN
भी           O      RP     NaN    NaN
कहा          O      VM     NaN    NaN
कि           O      CC     NaN    NaN
बेटा          O      NN     NaN    NaN
उनके         O      PRP    NaN    NaN
साथ          O      NST    NaN    NaN
मारपीट       O      NN     NaN    NaN
भी           O      RP     NaN    NaN
करता         O      VM     NaN    NaN
है            O      VAUX   NaN    NaN
।            O      SYM    NaN    NaN

याचिका       O      NN     NaN    NaN
में           O      PSP    NaN    NaN
कहा          O      VM     NaN    NaN
गया          O      VAUX   NaN    NaN
कि           O      CC     NaN    NaN
22           B-D    QC     NaN    DAC
फरवरी        I-D    NNP    PIV    NaN
2002         I-D    QC     NaN    DAC
को           O      PSP    NaN    NaN
उनके         O      PRP    NaN    NaN
बेटे          O      NN     NaN    NaN
```



```
चंद्र      चंद्र      XC    0    n     m    sg   3    d
मोहन      मोहन      XC    0    n     m    sg   3    d
शर्मा      शर्मा      NNP        unk
डायरेक्टर  डायरेक्टर  NN         unk
विशाल     विशाल     JJ         adj   any  any       any
भारद्वाज   भारद्वाज   NN         unk
अपनी      अपनी      PRP   0    n     f    sg   3    d
पिछली     पिछली     JJ         unk
फिल्म      फिल्म      NN    0    n     f    sg   3    d
'         '         SYM        punc
कमीने     कमीने     NNP        punc
'         '         SYM        punc
के        का        PSP   का   psp   m    sg        o
बाद       बाद       NST        adv
ब्रेक       ब्रेक       NN    0    n     m    sg   3    d
चाहते     चाह       VM    ता   v     m    pl   any
थे        था        VAUX  था   v     m    pl   any
,         ,         SYM        punc
लेकिन     लेकिन     CC         avy
इस        इस        DEM        pn    any  sg   3    o
```

https://bitbucket.org/sivareddyg/hindi-part-of-speech-tagger

Using a CRF Based Model

- We Use – CRF ++ a CRF Toolkit that is designed specifically for NLP tasks
- Feature Set Used -

  1. Word itself
  2. POS
  3. BIO Information (along with P,D, Tagging)
  4. Binary Label – isPivot
  5. Binary Label – isDigitorChar

Window Size - 4 (Two Tokens before the current token & two tokens after

Results

- **Rule Based Approach:**

  Mean precision (P) : 66.1%
  Mean recall (R) : 64.9%

  For rules based on the First 30 documents

- **CRF Model Based Approach:**

Precision (P) : 71.3%
Recall (R) : 75.5%

F1 : 73.34%

Improvements Needed on the Project

- Reaching our Goal State of choosing the ultimate model – rule based/ CRF/ **SVM / Voting System (PENDING)**
- **Rule Based -**
  1. **Streamlining** existing rules
  2. Solving Issues of **Ghost Tagging & Over-tagging** (affects Precision)
  3. **Handling Caveats** (such as tagging '1612' in 'Prisioner No.1612' as a TIMEX element)
- **CRF Based -**
  1. Increasing **Window Size**
  2. Increasing the **Number of Features**

Improvements Needed on the Project

- **OPTIMIZE** THE MODEL
- **FINISH TIMEX3 TAG ANNOTATION** OF DAINIK BHASKAR CORPUS
- **HANDLE CAVEATS** WHEN DEALING WITH DAINIK BHASKAR CORPUS
- **GENERATE A MODEL TAILORED** TO THIS CORPUS

Future Work

- **LINKING** OF TIME TAGS WITH EVENT TAGS IN **HEED**
- **APPLICATIONS**

CHRONOLOGICAL ORDERING OF EVENTS (Generating Inter & Intra Corpus Timelines)

QUERYING BY TIMESTAMPS ("What Happened on 15th September, 1939?")

QUERYING BY EVENT ("When did Hitler invade Poland?")

REASONING ABOUT THE LENGTH OF EVENTS ("How long did it last/ Period")

REASONING ABOUT THE OUTCOME OF EVENTS ("What caused the Financial Crash of 2008?")

THANK YOU

Our GitHub Repo:

https://github.com/RishavR/IASNLP-2018