**Will start by 10:10, please complete yesterday’s exercise.  
Today’s Agenda: Hands ON – Quora Review**

**What is NLP? Natural Text – write**

**Eg. Next word prediction, Sentiment,translation,NER,Summarisation, chatbots- Intent detection, spam filters, Text generation**

**NLP vs NLU – Natural Language Understanding**

**NLP Pipeline:**

**Problem:**

Agent: “When will you pay the bill”

# Customer:

“I won’t pay the bill” -> customer\_refused ->

“ i will do the payment in 2-3 days” -> payment\_future 🡪

“I will do it today” -> payment\_today

1. Pre-Processing:
   1. Lowercase
      1. i won’t pay the bill
      2. i will do the payment in 2-3 days
   2. Punctuation
      1. i wont pay the bill
      2. i will do the payment in 2 3 days
   3. Stop words: I, the, in, do
      1. wont pay bill
      2. will payment 2 3 days
   4. Text Standard
      1. wont pay bill
      2. will pay 2 3 day
   5. Tokens
      1. [‘wont’, ‘pay’, ‘bill’] ->
      2. [‘will’ ,’pay’, ‘2’,’3’ ,’day’]
   6. ~~BASELINE APPROACH <OPTIONAL HACK>~~
   7. Text to Computer – Text to Number
      1. The numbers should be representative.
      2. **Bag of Words**
         1. **Text into fixed-length Vector**
            1. [‘wont’, ‘pay’, ‘bill’] ->
            2. [‘will’ ,’pay’, ‘2’,’3’ ,’day’]

Sentence 1 : the cat sat

Sentence 2 : the cat sat in the hat

Sentence 3 : the cat with the hat

**After Cleaning:**

Sentence 1 :

cat sat

Sentence 2 : cat sat hat

Sentence 3 : cat hat

Words:

CAT

SAT

HAT

|  |  |  |  |
| --- | --- | --- | --- |
|  | CAT(f1) | SAT(f2) | HAT(f3) |
| Sent-1 | 1 | 1 | 0 |
| Sent-2 | 1 | 1 | 1 |
| Sent-3 | 1 | 0 | 1 |

[‘wont’, ‘pay’, ‘bill’]

[‘will’ ,’pay’, ‘2’,’3’ ,’day’]

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | wont | Pay | Bill | Will | 2 | 3 | day |
| Sent-1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Sent-2 | 0 | 1 | 0 | 1 | 1 | 1 | 1 |

Drawbacks: IF I consider these Stopword: will, No

>~~Will~~ pay today -> payment\_today[INTENT]

>~~No~~ pay~~ment~~ today ->refused\_pay[INTENT]

Big sentences:

Full paragraph -> 1200-1300 words

Multiple pargraphs ->

2000-3000 words, 10,000 words

ML: More columns as compared to rows

Frequency Based:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pay | Won’t | Bill | Will | 2 | 3 | day |
|  | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| Sent-1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Sent-2 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |

Cutoff:

10,000 words

Word-1 :100

Word2: 89

.

.

. <10 times

Top 100 words

W1,w2,w3…w100

Top-100

Top-150

1. Equal Importance

TF-IDF :

TF: Term Frequency

IDF: Inverse Document Frequency

TF \* IDF

Ptp\_future

Intent:

1. Query\_bill\_payment
2. Rtp

PRODUCTION PIPELINE

* Sentiment Analysis(P/N/NU)

1. Data Ingestion Pipeline
   1. Data Gathering (scrapping tweets, CSAT)
   2. Annotate the data(Training)
      1. Positive Tweets
      2. Negative Tweets
      3. Neutral Tweets
2. Data Preprocessing pipeline
   1. Clean the data.
   2. Removing duplicates
   3. Etc..
3. Defining your evaluation metrics. What is your success criteria(Precision - fraud)
   1. Example:
      1. Fraud/Not
      2. 1/10000
      3. High Accuracy ?
      4. 10001 🡪 Non fraud
      5. 10000/10001~99.9% Accuracy
      6. Precision (Fraud): 0
      7. Precision(Non-Fraud):99.9
4. BASELINE MODEL
5. “the apple phone is awesome”
6. “battery sucks, drain machine”
7. “apple is doing okay”

Positive: any tweet that has a word “awesome”

Negative: any tweet that says “suck” or “sucks”

Neutral: anything that is not positive or negative

TRAINING -ML/DL

EVALUATION(CV)

Deployment of the model: AWS/Heroku/Azure/GCP

Sentiment Model☹☺

1. Your product is working GOOD ☺
2. Your product works Good☺
3. Your product works good☺
4. Your product is working awesome 😐
5. Your product doesn’t work at all 😐
6. This product is shitty 😐
7. The product beats everything else in market. ☺

Baseline(Over simiplification):

“good”,”awesome”,”working” --Positive

“doesn’t work”,”shitty” –negative

Remember

Eating, eating, EATING, eatng, eat9g 🡺 eating

1. lowercasing

2.Removal of noise:

# Sentence-1: I want to go to Paris and visit Eiffel tower.

Sentence-2: want go Paris visit Eiffel tower.

Example:

I was playing in the park

I played in the park

I play at weekend

Playing, played, play 🡪 pla

I was pla in the park

I pla in the park

I pla at weekend

Stemming:

Keep on removing the last character till you find the root word.