SPELL CORRECTION FOR ROMAN URDU

(NATURAL LANGUAGE PROCESSING)

NAME: Abdullah Naveed

WEBSITE: CLICK HERE

GITHUB: CLICK HERE

DATED: 9th May, 2021

SPELL CORRECTION FOR ROMAN URDU - DOCUMENTATION

PROBLEM STATEMENT

The goal of this assignment was to implement spell corrector for roman Urdu, using Noisy Channel based on Bayes Theorem.

PROBLEM ANALYSIS

A wrong word will be fed to the algorithm and it should pick the best possible correct word out of some possible candidate words, which will be generated using the error model and n-gram language model.

$$\hat{w} = \texttt{argmax}_{w \in candidates} P(x|w) * P(w)$$

Here:

- P(x|w) is the error model
- P(w) is the language model
- “x” is the possible wrong word
- “w” belongs to candidate words

SOLUTION DESIGN

My technique to solve this problem in a step-wise manner is as following:

- First objective was to read the corpus, and tokenize the corpus into words using spaCy library, and saving the results to lists.
- Second step was to read the misspellings.txt, and make a list of correct words and their possible misspelled versions, this was again done using the spaCy.
- The next objective was to obtain the language models P(w):
  - ✓ UNI-GRAM Model was generated which will be later used in Noisy Channel.
  - ✓ Character Level BI-GRAM Model was generated which was later used for confusion matrices.
- After that, it was time to generate the Confusion Matrices:
  - ✓ Four matrices: insertion, deletion, substitution, transposition matrices were generated with initial values: zeros (later to be updated)
  - ✓ A Function “find_edit_type” was implemented to find the type of edit required to convert a wrong word into correct word.
  - ✓ Using the above function and lists generated from misspellings.txt, all the confusion matrices were updated. (See RESULTS section for outputs)

- The next objective was to obtain the Error Model P(x|w), It was generated using the following equation:

$$P(x|w) = \begin{cases} \frac{\text{del}[x_{i-1}, w_i]}{\text{count}[x_{i-1} w_i]}, \text{if deletion} \\ \frac{\text{ins}[x_{i-1}, w_i]}{\text{count}[w_{i-1}]}, \text{if insertion} \\ \frac{\text{sub}[x_i, w_i]}{\text{count}[w_i]}, \text{if substitution} \\ \frac{\text{trans}[w_i, w_{i+1}]}{\text{count}[w_i w_{i+1}]}, \text{if transposition} \end{cases}$$

Here:

- w belongs to candidate words
- x is the possible wrong word

- Next step was to generate Candidate Words list, for that all the possible words in vocabulary that were 1 edit-distance away from the wrong word were picked.

- Now, the final objective was to implement the Selection Model:

    - This model parsed the candidate words list, sent each candidate word to both error and language models to get the probabilities.

    - Then Multiplied those probabilities with power of 10 (to get rid of very small numbers)

    - Then applied argmax to choose the word with highest probability and returned it back as the output.

EVALUATION

TIME COMPLEXITIES:

For Time Complexity of the Algorithms:

Let N be the number of words in corpus

Let M be the number of misspelled words

Let C be the number of candidate words

Let U be count of unigrams

Let B be count of bigrams

1. **Language Models:**

   Unigram Model: O(N)

   Bigram Model: O(N)

2. **Error Model:**

   O(C)

3. **Selection Model:**

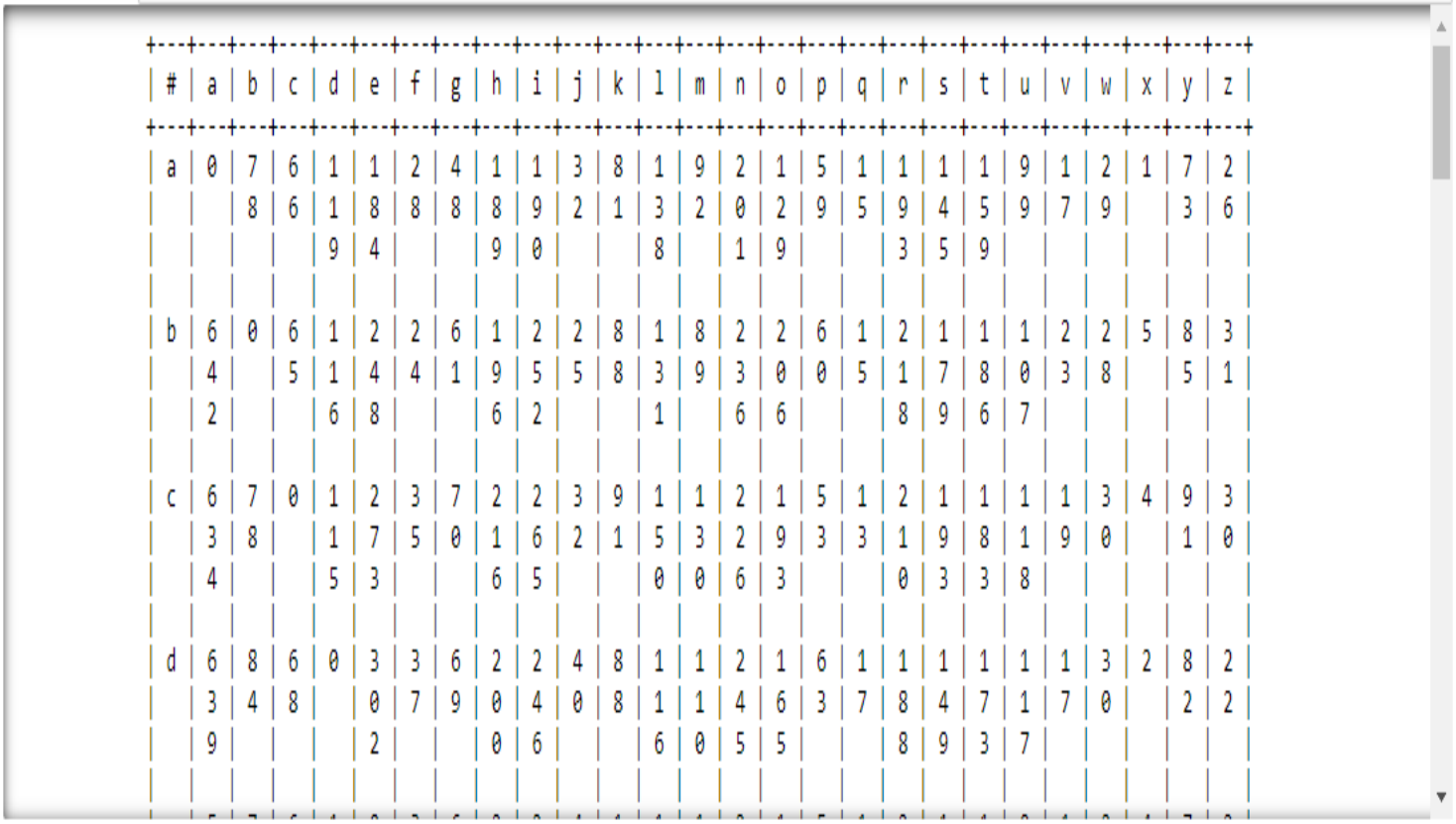
   O(C)

RESULTS (PRINTED USING PRETTY TABLES)

#	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
u	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
v	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Confusion Matrix (Before)

In [16]:
```
#Testing a confusion Matrix
print_confusionMatrix(substitutionMatrix)
```



#	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	7	6	1	1	2	4	1	1	3	8	1	9	2	1	5	1	1	1	1	9	1	2	1	7	2
		8	6	1	8	8	8	8	9	2	1	3	2	0	2	9	5	9	4	5	9	7	9		3	6
				9	4			9	0			8		1	9			3	5	9						
b	6	0	6	1	2	2	6	1	2	2	8	1	8	2	2	6	1	2	1	1	1	2	2	5	8	3
	4		5	1	4	4	1	9	5	5	8	3	9	3	0	0	5	1	7	8	0	3	8		5	1
	2			6	8			6	2			1		6	6			8	9	6	7					
c	6	7	0	1	2	3	7	2	2	3	9	1	1	2	1	5	1	2	1	1	1	1	3	4	9	3
	3	8		1	7	5	0	1	6	2	1	5	3	2	9	3	3	1	9	8	1	9	0		1	0
	4			5	3			6	5			0	0	6	3			0	3	3	8					
d	6	8	6	0	3	3	6	2	2	4	8	1	1	2	1	6	1	1	1	1	1	1	3	2	8	2
	3	4	8		0	7	9	0	4	0	8	1	1	4	6	3	7	8	4	7	1	7	0		2	2
	9				2			0	6			6	0	5	5			8	9	3	7					

Confusion Matrix (After)

```
Enter Wrong Word:hax
```

Candidates	x|w	P(x|w)	P(w)	P(x|w)*P(w) *10^9
hnx	a|n	2.138072237483023e-05	1.2636470325508209e-06	0.027017686382747156
hab	x|b	1.756632948422524e-05	6.950058679029515e-07	0.01220870206905317
haz	x|z	1.794181581268296e-05	4.2648087348590207e-07	0.007651841279716198
hanx	#|n	5.610979564812425e-06	6.63414692089181e-07	0.003722406280308722
had	x|d	3.0866644327210414e-05	0.00032210362859720425	9.942258140413784
oax	h|o	1.750414464967514e-05	3.159117581377053e-08	0.000552976511097558
haq	x|q	1.462189010674792e-05	9.737979944594764e-05	1.4238767261157985
haj	x|j	1.5339576540985516e-05	6.950058679029515e-07	0.010661095707131394
hay	x|y	1.993297205928863e-05	7.809338661164073e-05	1.5566332933450595
hai	x|i	2.1761604853661295e-05	0.02151063695423914	468.1059815487165
tax	h|t	2.1355815287257937e-05	1.320511149015608e-05	0.28200592183142065
hx	a|#	1.2049763273028463e-05	3.4750293395147577e-07	0.004187328090798129
hau	x|u	1.8287710705297605e-05	2.3061558344052482e-06	0.04217431074093739
hah	x|h	1.0764455190572093e-05	6.202927371033842e-05	0.6677113373586695
hao	x|o	2.2194797791091352e-05	2.0850176037088547e-06	0.04627654410518388
hag	x|g	3.070355532373755e-05	2.353542598125904e-06	0.0722621253683317
hat	x|t	2.3577807629284774e-05	1.7548898164549525e-05	0.4137645450296573
haw	x|w	1.4283237164297506e-05	5.528455767409842e-07	0.007896424487824315
hak	x|k	8.203699063099587e-06	3.03275287812197e-06	0.024879791944861784
han	x|n	2.3933644449436825e-05	0.00012938166054529717	3.096574661768871
fax	h|f	2.70166741123334e-05	1.5479676148747556e-06	0.04182093658751729
hae	x|e	2.1990373552959813e-05	1.1088502710633454e-05	0.2438403167498371
max	h|m	1.3452323619858858e-05	6.460395453916072e-06	0.08690733035834397
bax	h|b	1.7761510922938856e-05	3.4086878703058394e-05	0.6054344684132635
hoax	#|o	5.734711116794847e-07	1.263647032550821e-07	7.246650685274012e-05
ham	x|m	1.659119913115926e-05	3.4055287527244627e-05	0.5650180568333998
ax	h|#	0.00015828628713132487	2.0534264278950842e-06	0.32502924516885207
hav	x|v	8.525455214437634e-05	2.2113823069639366e-07	0.01885304082002082
has	x|s	2.4028608010901176e-05	2.9790478792385606e-05	0.7158237373592983
hap	x|p	1.4150007934889066e-05	1.1293845353422962e-05	0.15980800136634493
haa	x|a	1.2762582336843791e-05	6.2076660474059076e-06	0.07922584904964754
hal	x|l	2.4532033024912186e-05	0.00011213287855097847	2.750847479791071
har	x|r	2.6011222885552197e-05	0.0013881636520207976	36.10783415433509
ha	x|#	1.6005192832849708e-05	0.0002843521734997485	4.5511113693034115

```
Best Correction for: hax is: hai
```

Result#1

INPUT: hax

OUTPUT: hai

Enter Wrong Word:humar

Candidates	x|w	P(x|w)	P(w)	P(x|w)*P(w) *10^9
umar	h|#	2.8826290297408496e-05	3.768827274582824e-05	1.0864130909791536
humay	r|y	1.6610810049407192e-05	5.844367525547547e-07	0.009707967882579424
kumar	h|k	7.596017651018136e-06	3.854123449280004e-06	0.02927598974993381
shumar	#|s	0.0	1.1688735051095094e-06	0.0
homar	u|o	2.0249892830016335e-05	1.579558790688526 3e-06	0.03198589623015286
hamar	u|a	1.3531663031409298e-05	2.3693381860327894e-07	0.003206108594084626
humor	a|o	1.475839646933394e-05	1.453194087433444e-06	0.021446814489234696
humari	#|i	1.1720622552587503e-05	6.618351332984925e-06	0.07757119789433066
humary	#|y	5.0937893972773694e-05	8.371661590649189e-07	0.04264348104804304
hukar	m|k	7.21621676846723e-06	2.7484322957980357e-06	0.01983328321993467
khumar	#|k	2.4614657536269697e-05	2.353542598125904e-06	0.057931645049891546
human	r|n	2.1167978868613013e-05	5.8285719376406616e-06	0.12337908761016832
humai	r|i	1.6247684704929547e-05	1.579558790688526e-07	0.002566417320400698
hunar	m|n	2.3827272696328217e-05	1.750151140082887e-05	0.41701328474544674
humare	#|e	3.5307042261171626e-05	8.087341008325255e-06	0.28554009076144615
humra	ar|ra	1.7601773764710605e-05	4.896632251134431e-07	0.008618941309345386
humara	#|a	1.0838481160281084e-05	7.313357200887876e-06	0.07926568424022926
huma	r|#	1.4841178808642456e-05	6.160279283685252e-07	0.009142580636034868

Best Correction for: humar ïs: umar

Result#2

INPUT: humar

OUTPUT: umar

```
Enter Wrong Word:kaya
```

Candidates	x|w	P(x|w)	P(w)	P(x|w)*P(w) *10^9
aya	k|#	1.585969107982 38e-05	5.514239738293645e-05	0.8745413878942565
kayam	#|m	1.730050490442282e-05	4.422764613927873e-07	0.007651606089436687
kama	y|m	1.4797555981844743e-05	3.190708757190823e-06	0.04721469145629347
kaza	y|z	1.7381134068536616e-05	1.9112661367331165e-06	0.03321997296321233
kaga	y|g	2.1825418844584523e-05	7.89779395344263e-08	0.0017237266098211249
haya	k|h	1.0978673204314823e-05	8.150523359952796e-06	0.08948193241305578
gaya	k|g	2.3305108257776693e-05	0.0028844638988521315	67.22274342839756
aaya	k|a	1.3469305137255337e-05	0.0008346862517635379	11.242643818875024
kaka	y|k	7.671977827528317e-06	2.479907301380986e-06	0.01902579383052051
keya	a|e	1.692982733784354e-05	3.996283740441972e-06	0.06765639371871413
kaay	ya|ay	5.709835667757351e-06	9.477352744131157e-08	0.0005411412673435808
kyaa	ay|ya	1.517156848066723e-05	3.2222999330045936e-06	0.04888734409882862
paya	k|p	1.1973083637213824e-05	3.073821406679872e-05	0.3680312078803636
kasa	y|s	1.751839401779987e-05	1.7059234939436083e-06	0.029885039931125963
khaya	#|h	4.280854903847718e-07	6.659419861542827e-05	0.028508010171066505
kayak	#|k	0.0	1.4373984995265588e-06	0.0
klya	a|l	2.5454289905547983e-05	1.0267132139475421e-06	0.026134255797677648
laya	k|l	2.3609776144276392e-05	4.4085485848116766e-05	1.0408484520857018
karya	#|r	4.7477531258332e-06	4.422764613927873e-07	0.002099819452060 0523
kamya	#|m	1.730050490442282e-05	2.2113823069639366e-07	0.0038258030447183436
naya	k|n	2.3508157437002392e-05	0.00033053847253948097	7.77035045144441
saya	k|s	1.8820436816420132e-05	0.0001071730639482165	2.017043878459563
kayo	a|o	1.4758396469333 94e-05	1.1056911534819683e-07	0.0016318228415722052
kay	a|#	9.722902593303213e-06	8.640186585066238e-05	0.8400769255456416
zaya	k|z	1.2334998371219534e-05	5.828571937640662e-05	0.7189542535733344
kada	y|d	2.020765923579962e-05	2.053426427895084e-07	0.004149494152068912
kaha	y|h	1.0603791680265047e-05	0.002813636482677658	29.835215126307556
kya	a|#	9.722902593303213e-06	0.009296303715542438	90.38705550398187
maya	k|m	1.4947026244287 62e-05	1.8812545197100348e-05	0.28119160678290595
baya	k|b	1.5809696535802716e-05	3.0011617023081997e-07	0.004744745576836572
kana	y|n	2.372090094321961e-05	2.843205823239347e-07	0.006744340369424571
kala	y|l	2.5454289905547983e-05	0.00023709177448234779	6.035002761894484
kiya	a|i	1.3968597710120426e-05	0.0001630736495506 8344	2.2779102076946574
daya	k|d	2.109590799341719e-05	7.629268959025582e-06	0.1609463560166374
kayar	#|r	4.7477531258332e-06	3.9488969767213157e-07	0.001874838796482 19
kraya	#|r	2.7420298332845863e-05	2.843205823239347e-07	0.007796155189490751
kara	y|r	2.3083552777221646e-05	1.8228108444545592e-05	0.42076950330858776
vaya	k|v	6.730622537713922e-05	4.7386763720655784e-08	0.0031894241988757025
kawa	y|w	7.141618582148753e-06	2.2113823069639366e-07	0.0015792848975648627
yaya	k|y	1.8382629788010626e-05	1.2478514446439356e-06	0.022938791137323705
kayi	a|i	1.3968597710120426e-05	0.0005555308266851546	7.759986633535559
kaye	a|e	1.692982733784354e-05	4.106852855790168e-07	0.00695283097504572
kata	y|t	2.0244819116244516e-05	6.397213102288531e-06	0.1295104221039007 6
kaa	y|#	1.504903845582232e-05	1.4879443808285918e-05	0.22392132207214208
kaia	y|i	1.698287405809378e-05	1.690127906036723e-06	0.028703229370291426
jaya	k|j	1.3672231264791436e-05	3.0580258187729865e-05	0.4181003620796746
raya	k|r	2.5110401313758 18e-05	1.8954705488262313e-07	0.004759602615943614
kayl	a|l	2.5454289905547983e-05	7.581882195304925e-07	0.019299142742900413
taya	k|t	2.2466811458271353e-05	1.2873404144111489e-05	0.28922434373188194

```
Best Correction for: kaya is: kya
```

Result#3

INPUT: kaya

OUTPUT: kya

```
Enter Wrong Word:asdfasdfas
+------------+-----+--------+------+-------------------+
| Candidates | x|w | P(x|w) | P(w) | P(x|w)*P(w) *10^9 |
+------------+-----+--------+------+-------------------+
+------------+-----+--------+------+-------------------+

Best Correction for: asdfasdfas ïs: None
```

Result#4

INPUT: asdfasdfas

OUTPUT: None

```
Enter Wrong Word:hampe
+------------+-----+-----------------------+-----------------------+-----------------------+
| Candidates | x|w |         P(x|w)        |         P(w)          |   P(x|w)*P(w) *10^9   |
+------------+-----+-----------------------+-----------------------+-----------------------+
|   sampe    | h|s | 1.751839401779987e-05 | 2.3693381860327894e-07 |  0.004150699990434161 |
|   hamle    | p|l | 2.2871970639767755e-05 | 6.160279283685252e-07  |  0.014089772690921862 |
|   hamre    | p|r | 2.8150674118562983e-05 | 6.318235162754104e-07  |  0.017786257907113655 |
|   hamne    | p|n | 2.3401785683893784e-05 | 1.6111499665022968e-06 |   0.0377037862206994  |
|   humpe    | a|u | 1.5474216750636436e-05 | 3.317073460445905e-07  |  0.00513291137047236  |
|    hame    | p|# | 1.8235372018030897e-05 | 1.4610918813868867e-05 |   0.2664355400961455  |
|    ampe    | h|# | 2.9535187910042257e-05 | 1.263647032550821e-07  | 0.0037322052558355786 |
|   hamse    | p|s | 1.9530641979303913e-05 | 7.739838074373778e-07  |  0.015116400640837928 |
|    hamp    | e|# | 1.6980009521866877e-05 | 1.4216029116196736e-07 | 0.0024138830975615733 |
+------------+-----+-----------------------+-----------------------+-----------------------+

Best Correction for: hampe ïs: hame
```

Result#5

INPUT: hampe

OUTPUT: hame

SOME CHALLENGES FACED

- Implementation of Unigram Model was initially taking O(N*N) which was never ending on the huge corpus, it was reduced to O(N) later.
- Finding the type of edit required to make wrong word into correct was challenging, so the technique used solve this was Divide and Conquer.
- Computation of Error Model, where character level bigram model is used, was challenging to implement.

CONCLUSION AND FUTURE WORK

Despite of all the challenges and difficulties faced throughout the implementation, observing the end results was satisfying. As of now this model only suggests corrections for words with edit-distance=1, also it only caters the word that are present in vocabulary and for non-vocabulary words it gives none as output, these parts can be improved in the future.

REFERENCES

https://www.aclweb.org/anthology/C90-2036.pdfss

https://web.stanford.edu/~jurafsky/slp3/B.pdf