



# **Spelling Corrector**

# Types of Spelling Corrector:-

---

- Unigram Language Model –

Finds the best candidate based on term frequency.

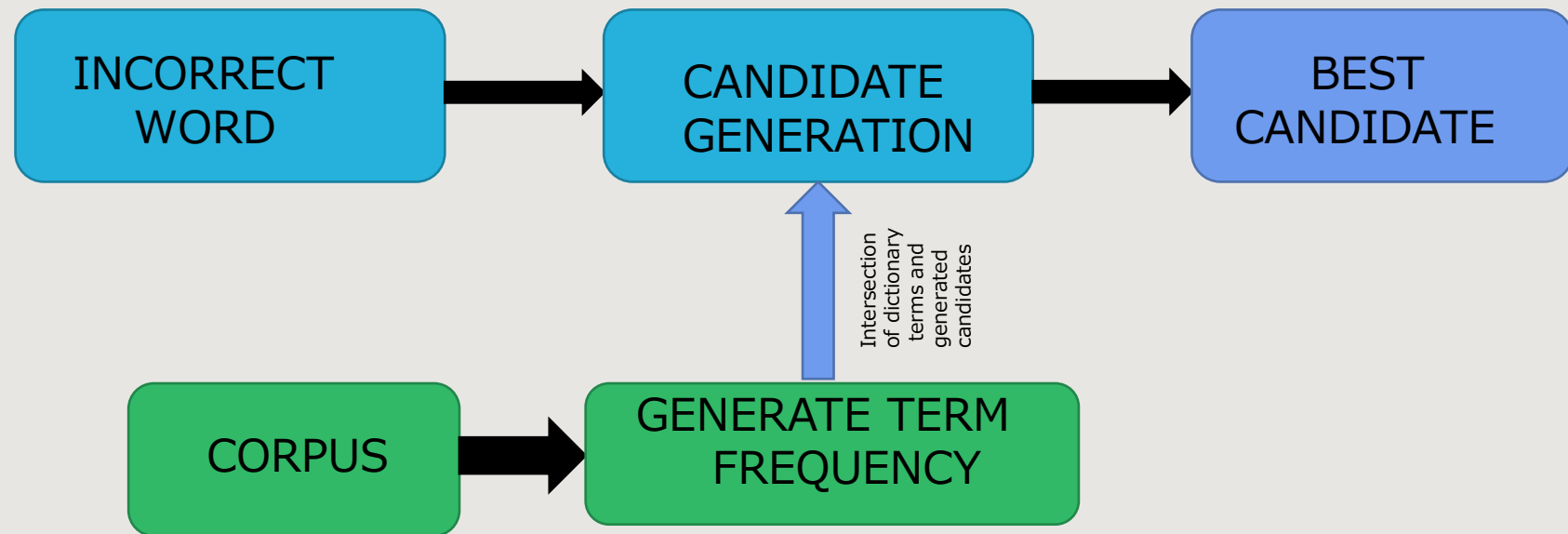
- Unigram Noisy Channel Model –

Uses Bayes theorem to filter the best candidate.

# Language Model :-

---

Workflow –



# Language Model :-

---

- Takes incorrect word as an input.
- Generates all possible candidates within one and two edit distance.
- Since there can be many possible candidates, we consider only those candidates that are present in the corpus.
- Filters the best candidate that have maximum term frequency.

# How it works :-

---

- Give a command line input as shown below .

```
PS D:\Information Retrieval> python .\spelling_corrector.py "language" "thew"
```

Incorrect word

"language" denotes  
the model name

- It gives the best candidate as output.

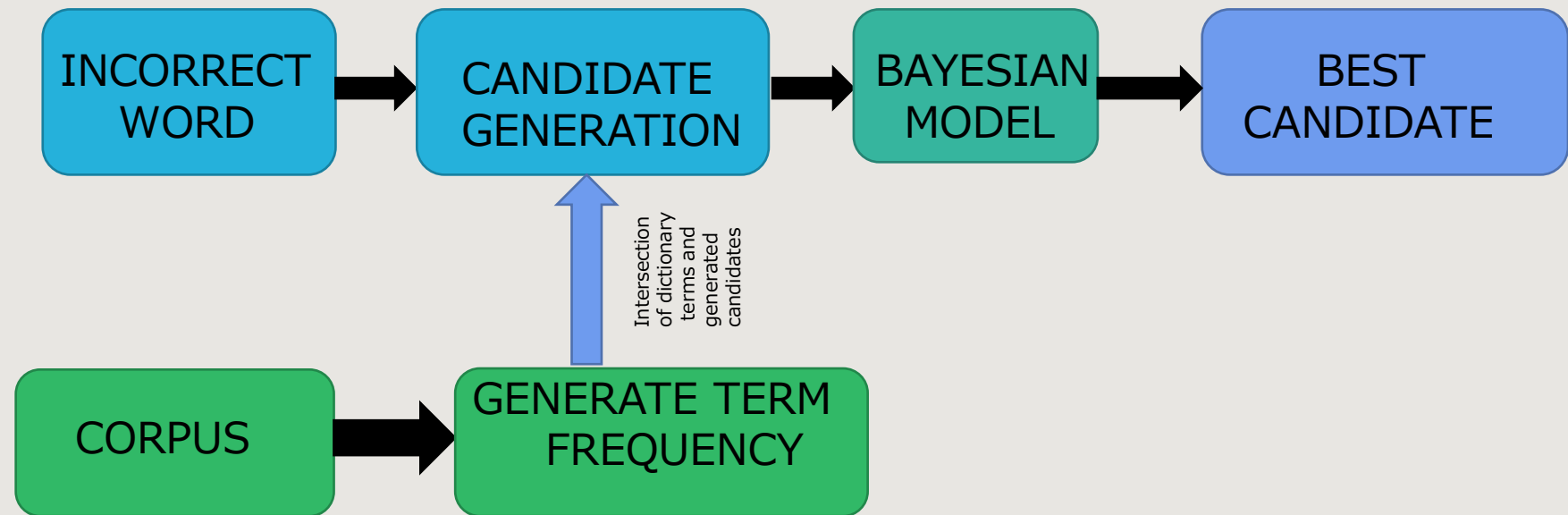
```
PS D:\Information Retrieval> python .\spelling_corrector.py "language" "thew"  
the
```

Here in above snippet, we can see that the correct word for misspelled word "thew" is "the".

# Noisy Channel Model :-

---

Workflow –



# Noisy Channel Model

---

- Takes incorrect word as an input.
- Generates possible candidates within one edit distance.
- Since there can be many possible candidates so, It only considers those candidates that are there in the corpus.
- For each generated candidate, it calculates the likelihood probability based on the number of possible edits between incorrect word and generated candidate.

# Error probability -

---

- $P(c = \text{correct} \mid w = \text{incorrect}) = P(w \mid c) * P(c)$

Where,  $P(w \mid c)$  is likelihood probability and  $P(c)$  is prior probability of each candidate.

- $$P(w \mid c) = \begin{cases} \frac{\text{del}[c_{p-1}, c_p]}{\text{chars}[c_{p-1}, c_p]}, & \text{if deletion} \\ \frac{\text{add}[c_{p-1}, t_p]}{\text{chars}[c_{p-1}]}, & \text{if insertion} \\ \frac{\text{sub}[t_p, c_p]}{\text{chars}[c_p]}, & \text{if substitution} \\ \frac{\text{rev}[c_p, c_{p+1}]}{\text{chars}[c_p, c_{p+1}]}, & \text{if reversal} \end{cases}$$

It calculates the above probability based on the confusion matrix as shown in next slide.



# Confusion Matrix for Insertion and Deletion -

		confusion matrix due to insertion																									
		correct																									
incorrect		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	17	199	30	1	40	57	1	216	0	2	352	32	184	0	159	0	154	114	90	90	2	3	0	18	0
	b	4	0	0	0	44	0	0	0	21	0	0	13	0	0	12	0	0	24	4	2	22	0	0	0	1	0
	c	34	0	0	0	65	0	0	188	88	0	31	11	0	0	72	0	2	32	16	6	22	0	0	0	2	0
	d	32	0	0	0	85	0	11	2	46	0	0	9	0	0	7	0	0	9	15	0	8	0	0	0	8	0
	e	354	0	47	206	0	18	4	2	66	0	0	100	26	150	19	2	0	211	258	67	22	1	8	8	36	2
	f	5	0	0	0	48	0	0	0	81	0	0	13	0	0	14	0	0	36	0	13	2	0	0	0	2	0
	g	12	0	0	0	104	0	0	37	12	0	0	3	0	12	4	0	0	23	10	5	114	0	0	0	4	0
	h	16	0	0	0	129	0	0	0	26	0	0	9	2	6	41	0	0	6	7	18	0	0	0	0	11	0
	i	84	3	65	33	87	16	47	0	0	0	0	61	69	146	64	45	0	35	133	96	5	3	0	0	0	6
	j	0	0	0	0	2	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	4	0	0	0	0	0
	k	0	0	0	0	78	0	1	0	5	0	0	0	0	6	0	0	0	0	14	0	0	0	0	0	0	0
	l	11	0	4	6	263	1	0	0	72	0	0	0	0	1	42	1	0	0	15	7	5	2	0	0	54	0
	m	32	33	0	0	94	0	0	0	48	0	0	0	0	27	24	32	0	0	13	0	4	0	0	0	0	0
	n	49	0	33	68	213	2	83	0	116	2	2	1	1	0	15	1	0	1	142	75	20	1	0	7	20	0
	o	72	7	50	12	33	21	22	1	26	0	2	78	158	81	0	56	0	127	49	39	235	3	83	1	10	0
	p	40	0	0	0	88	0	0	50	33	0	0	38	0	4	21	0	0	59	4	16	3	0	0	0	1	0
	q	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0
	r	94	0	20	15	392	0	4	20	194	0	4	7	5	39	60	1	0	0	48	51	10	0	0	0	35	0
	s	26	0	101	0	128	1	0	62	86	0	4	0	2	1	25	19	0	0	0	120	8	0	20	0	14	0
	t	74	0	4	0	478	0	2	57	98	0	0	19	1	1	63	0	0	74	80	0	42	0	10	0	7	0
	u	61	13	61	8	37	17	14	0	93	0	0	68	16	63	1	34	0	119	24	18	0	0	0	0	0	0
	v	6	0	0	0	83	0	0	0	35	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
	w	4	0	0	2	39	0	0	95	2	0	0	6	0	2	5	0	0	5	3	0	0	0	0	0	1	0
	x	2	0	19	0	2	0	0	13	17	0	0	0	0	0	0	2	1	0	0	4	0	0	0	0	0	0
	y	8	0	2	0	21	0	0	0	1	0	0	2	23	0	2	2	0	0	23	0	0	0	1	0	0	0
	z	1	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

		confusion matrix due to deletion of one character																									
		correct																									
incorrect		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	79	53	9	7	6	1	139	1	3	122	26	108	1	27	0	171	92	60	42	1	3	1	15	0
	b	12	0	0	1	1	0	0	1	5	0	0	10	0	0	12	0	0	5	7	2	20	0	0	0	3	0
	c	19	0	0	0	62	0	0	46	28	0	70	13	0	1	46	2	5	6	20	39	17	1	0	1	4	0
	d	13	0	1	0	115	1	4	3	20	1	0	2	0	5	21	0	0	5	7	1	1	0	0	0	1	0
	e	285	2	99	117	0	46	13	2	81	0	4	82	39	86	20	13	1	162	136	22	10	4	5	11	34	4
	f	9	0	0	0	26	0	0	1	17	0	0	5	0	0	13	0	0	12	0	5	8	0	0	0	0	0
	g	17	0	0	4	47	0	0	13	9	2	1	2	0	3	5	0	3	4	6	5	18	0	0	0	0	0
	h	12	0	0	0	116	0	3	0	21	0	0	3	2	2	26	0	0	11	4	21	12	0	0	0	7	0
	i	71	2	33	22	101	16	6	2	0	0	0	99	17	84	56	7	0	24	122	60	11	7	0	1	4	4
	j	0	0	0	0	1	0	0	0	1	0	0	0	0	0	2	0	0	0	0	2	0	0	0	0	0	0
	k	0	0	1	0	64	0	1	1	2	0	0	0	0	1	0	0	0	6	1	0	0	1	0	0	0	0
	l	20	1	1	2	215	2	1	1	45	0	0	0	3	1	18	2	0	1	6	7	7	0	1	0	22	0
	m	30	4	1	1	79	0	0	1	24	0	0	0	0	32	13	3	0	4	6	4	9	0	0	0	2	0
	n	16	1	17	49	174	2	35	5	61	0	9	5	4	0	11	1	0	3	34	71	16	1	1	0	2	0
	o	53	9	28	9	25	22	2	3	12	0	0	52	42	49	0	25	0	77	22	18	158	0	51	1	3	0
	p	23	0	1	0	76	0	0	37	27	0	0	9	0	1	17	0	0	20	1	17	5	0	0	0	2	0
	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0
	r	29	2	10	18	299	1	2	3	73	0	0	12	3	15	26	1	0	0	24	17	16	1	2	1	19	0
	s	17	0	47	3	158	1	0	31	38	0	1	5	1	3	14	2	0	2	0	38	23	0	1	0	4	2
	t	36	1	2	3	271	1	2	53	67	1	0	8	4	0	34	0	0	31	21	0	18	0	20	0	4	0
	u	46	3	13	8	54	0	5	2	41	0	0	83	8	23	4	4	0	70	38	18	0	1	0	0	0	0
	v	5	1	1	0	34	0	0	0	30	0	0	1	0	1	10	0	0	1	0	0	0	0	0	0	4	0
	w	1	0	0	1	38	0	0	53	1	0	0	2	0	5	3	1	0	1	1	0	0	0	0	0	0	0
	x	2	0	23	0	1	0	1	6	4	0	0	1	0	0	0	1	1	0	37	3	0	0	0	0	0	4
	y	1	0	0	1	27	0	0	0	5	0	0	2	2	1	0	0	0	3	11	2	2	0	0	0	0	0
	z	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

# Confusion Matrix for Substitution and Transposition -

		confusion matrix due to substitution of one character																											
		correct																											
		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		
incorrect	a	15	1	1	13	85	6	1	7	3	55	9	0	0	17	2	7	35	3	2	0	15	9	16	12	3	0	21	0
	b	1	0	0	41	1	2	1	0	0	0	0	1	5	2	1	31	0	3	0	2	0	9	1	0	0	0	0	
	c	1	1	0	3	2	3	58	5	5	0	28	8	0	8	1	11	11	3	20	9	65	3	2	2	16	0	2	
	d	7	37	7	8	18	1	1	1	3	7	3	4	3	16	0	3	0	11	7	10	6	5	5	4	0	2	1	
	e	74	9	1	13	11	2	7	12	9	17	0	1	5	4	15	2	9	12	23	35	16	0	6	0	15	4	0	
	f	0	2	4	2	3	2	5	1	1	0	1	1	0	4	0	4	0	7	3	4	0	3	1	0	0	0	0	
	g	1	0	29	24	10	1	0	1	0	30	11	1	3	0	1	1	9	3	4	15	6	0	0	7	2	1	1	
	h	1	1	3	0	15	2	0	0	14	1	6	2	1	17	6	7	0	6	6	14	12	0	1	0	1	0	0	
	i	31	3	0	1	2	7	1	0	15	0	1	0	11	1	4	1	0	0	4	2	16	11	9	0	4	1	14	0
	j	0	0	0	2	0	0	12	2	0	0	0	0	0	0	0	16	0	0	0	1	0	1	0	0	0	0	0	0
	k	9	0	53	4	0	0	13	32	0	0	1	3	0	0	2	3	0	0	2	23	6	0	0	2	2	0	0	
	l	6	16	1	16	12	3	2	1	19	2	3	7	5	9	1	2	0	9	3	34	9	2	11	0	5	0	0	
	m	0	5	0	6	2	2	3	0	0	0	0	0	0	14	0	5	0	3	3	2	1	1	3	4	0	0	0	
	n	8	4	5	26	2	3	9	1	8	0	0	15	2	30	16	2	1	0	23	14	20	13	1	4	1	3	1	
	o	35	2	0	0	1	2	16	0	2	3	1	0	0	4	0	1	9	0	0	3	1	8	1	0	3	0	7	0
	p	2	46	8	1	0	12	1	3	1	2	0	1	8	0	1	2	0	6	3	10	0	1	0	0	1	0	0	
	q	2	0	22	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	
	r	15	5	6	15	27	14	5	4	15	1	1	2	11	25	1	11	0	5	14	13	26	2	9	4	6	0	0	
	s	13	0	38	3	12	7	3	2	6	0	0	7	3	9	1	5	0	13	7	60	5	1	4	14	5	61	0	
	t	8	4	64	87	21	7	6	25	15	2	12	18	5	2	0	15	1	11	38	1	6	0	3	1	6	0	0	
	u	12	6	0	1	0	16	2	1	1	2	13	1	0	10	3	10	13	0	0	12	10	4	0	1	44	0	10	0
	v	0	8	3	1	0	40	2	0	3	0	0	2	1	4	0	1	0	0	0	4	1	0	3	3	1	1	1	
	w	2	1	0	1	8	1	0	9	1	0	2	2	2	0	5	0	1	10	4	1	52	6	0	0	1	0	0	
	x	0	0	15	0	0	0	0	0	0	0	0	1	1	2	0	0	0	0	8	2	0	0	0	0	1	2	0	
	y	2	0	3	1	49	0	2	2	12	1	0	2	1	4	10	0	0	3	7	5	10	2	3	0	10	0	0	
	z	0	0	5	1	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	4	1	0	0	5	0	0	0	

		confusion matrix due to transposition																										
		correct																										
		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	
incorrect	a	0	2	10	2	25	0	13	11	32	0	0	14	18	44	7	8	0	40	3	25	39	3	1	0	1	0	
	b	14	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	
	c	12	0	0	0	16	0	0	0	38	0	0	0	0	3	1	0	0	0	3	8	4	0	0	0	1	9	0
	d	2	0	0	0	68	0	0	0	10	0	0	9	0	7	8	0	0	1	0	0	3	0	0	0	0	0	
	e	7	2	27	23	0	8	9	6	12	0	5	12	41	74	12	12	0	18	9	5	23	2	1	8	2	2	
	f	0	0	0	0	21	0	0	0	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
	g	1	0	0	9	8	0	0	0	14	0	0	1	0	6	4	0	0	1	0	0	0	0	0	0	0	0	
	h	0	0	8	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	31	0	0	7	1	0	0	
	i	56	17	94	12	81	17	26	3	0	0	0	41	21	82	14	0	0	43	67	72	17	8	2	1	2	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	6	0	1	0	1	0	0	0	0	0	0	1	0	5	0	0	0	3	0	0	0	0	0	0	0	0	
	l	18	4	4	1	75	0	0	0	21	0	0	0	0	2	19	6	0	0	0	4	11	0	1	0	2	0	
	m	9	0	0	0	49	0	0	0	21	0	0	0	0	1	12	0	0	2	2	0	5	0	0	0	0	0	
	n	33	0	0	0	62	0	10	0	56	0	3	0	2	0	5	0	0	2	0	0	6	0	5	0	0	0	
	o	3	0	7	2	4	0	2	5	19	0	0	26	9	21	0	10	0	30	8	5	4	0	6	0	0	1	
	p	1	0	0	0	11	0	0	0	3	0	0	2	4	0	10	0	0	0	2	0	2	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	47	0	1	2	11	0	0	0	28	0	0	0	0	0	42	2	0	0	0	6	17	0	1	0	0	0	
	s	18	0	1	0	64	0	0	0	57	0	2	4	3	4	3	4	0	5	0	23	8	0	0	0	5	0	
	t	14	0	1	0	37	1	0	32	33	0	0	4	0	5	4	0	0	2	13	0	5	0	0	1	0	0	
	u	13	0	7	12	21	2	7	1	6	0	0	10	0	9	15	1	0	12	14	6	0	0	0	0	0	0	
	v	2	0	0	0	5	0	0	0	7	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
	w	1	0	0	0	4	0	0	0	0	0	0	0	0	0	8	0	0	0	5	1	0	0	0	0	0	0	
	x	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	y	5	0	1	2	5	0	0	0	0	0	0	5	0	1	0	0	0	4	7	1	0	0	0	0	0	0	
	z	0	0	0	0	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

# Calculation of error probability, Confusion Matrix -

---

```
del [x,y] : count(xy typed as x)
ins [x,y] : count(x typed as xy)
sub [x,y] : count(y typed as x)
trans [x,y] : count(xy typed as yx)
```

**Note:-** Insertion and deletion has been performed on previous character.

# Unigram prior probability

---

$$P(c) = T(c)/N$$

Where,  $N$  = Total number of terms in the dictionary,

$T(c)$  = Total count of candidate  $c$ .

# How it works :-

---

- Give a command line input as shown below . Incorrect word

```
PS D:\Information Retrieval> python .\spelling_corrector.py "noisy" "thew"
```

↑  
"noisy" denotes the  
model name

- It gives the best candidate as output.

```
PS D:\Information Retrieval> python .\spelling_corrector.py "noisy" "thew"  
the
```

Here in above snippet, we can see that the correct word for misspelled word "thew" is "the".

# Testing -

---

- Accuracy for Language model is 35.3%.
- Accuracy for noisy channel model is 26.5%.

*Language model outperforms noisy channel model because for noisy channel model, only candidates within one unit distance have been considered. While for language model, candidates within both one and two unit distance have been considered.*

# References -

---

- <https://norvig.com/ngrams/>
- <https://www.2power3.com/rajendra/#slides>
- <https://norvig.com/ngrams/ch14.pdf>
- <https://web.stanford.edu/~jurafsky/slp3/B.pdf>