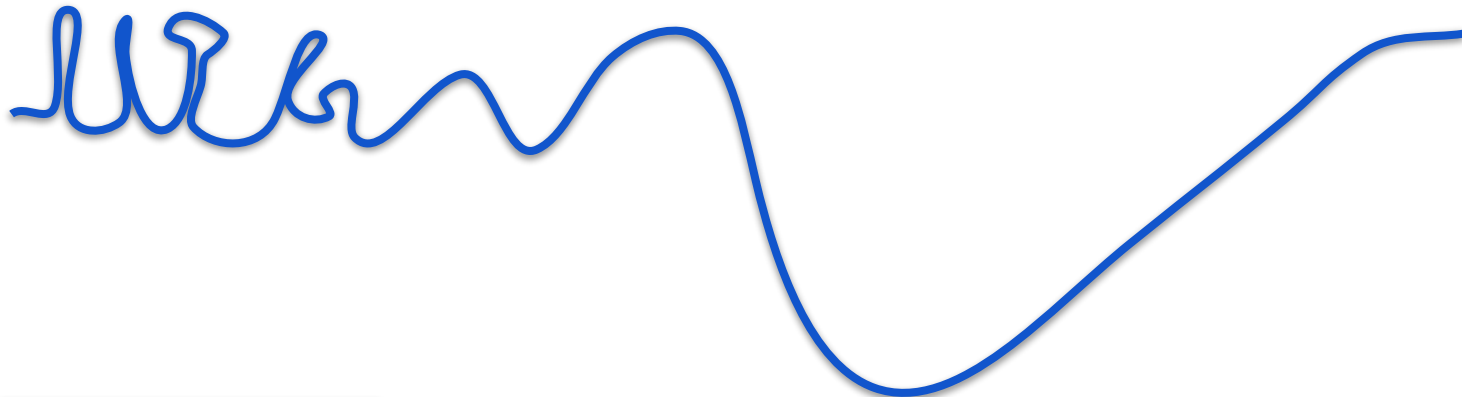


Computing with Signals



DA 623

Jan - May 2024

IIT Guwahati

Instructors: Neeraj Sharma

Lecture-11_misc_app

Sentiment classifier using NLP

Understanding Human Emotions:

- People express their thoughts and feelings through words.
- We can capture these expressions in text, and just like humans can recognize sentiments in a conversation, a sentiment classifier can be trained to do the same using the power of computers.

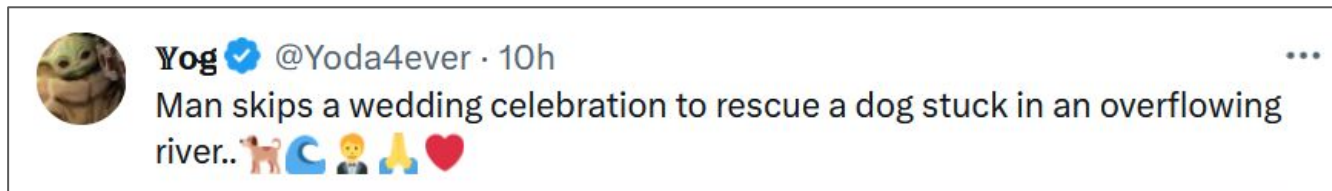
Sentiment classifier using NLP

Use case:

- **Social Media and Reviews:** Surrounded by vast amounts of text every day, especially on social media and review platforms.
- **Helpful for Businesses:** Understand customer feedback. By analyzing reviews and comments, companies can quickly grasp what customers like or dislike about their products or services.
- **Personalization in Services:** Streaming platform use it to recommend movies or shows based on what users enjoy.
- **Efficiency in Customer Support:** Businesses can prioritize and address negative feedback more promptly, leading to better customer satisfaction.
- **Filtering Out Hate Speech:** Filter out hate speech or harmful content. It helps create a safer online environment by identifying and taking action against harmful language.

Sentiment classifier using NLP

- **Focus: Tweets sentiment classification**



TSATC: Twitter Sentiment Analysis Training Corpus

TSATC: Twitter Sentiment Analysis Training Corpus

Contains: The Twitter Sentiment Analysis Dataset contains 1,578,627 classified tweets, each row is marked as 1 for positive sentiment and 0 for negative sentiment (inside a csv file).

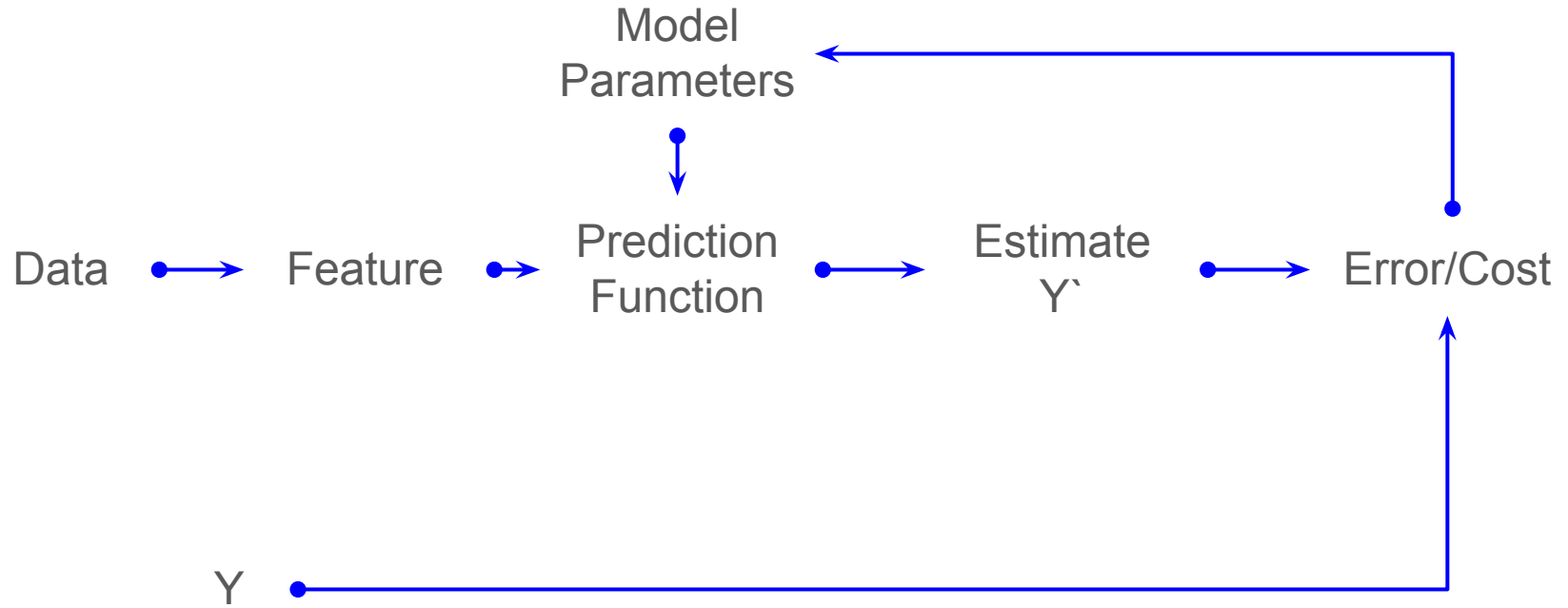
Accessible:

- It can be downloaded from:
<http://thinknook.com/wp-content/uploads/2012/09/Sentiment-Analysis-Dataset.zip>
- Apache License

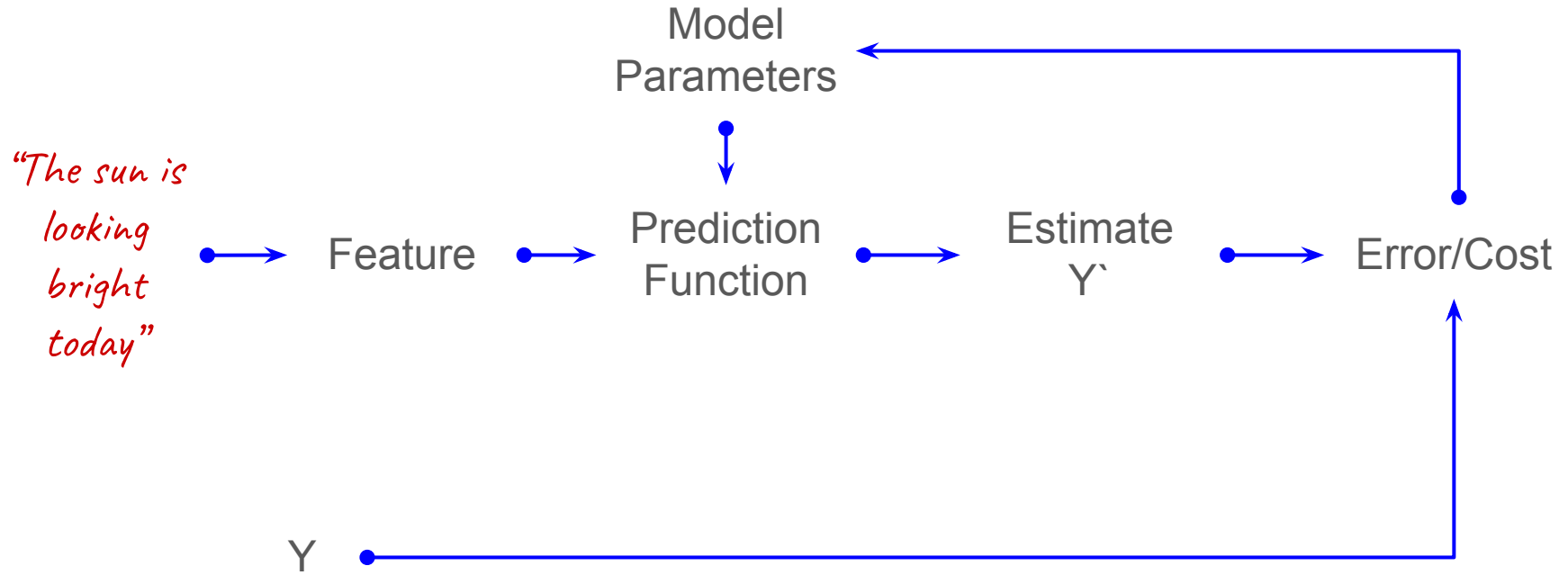
Data Source: The dataset is based on data from the following two sources:

- University of Michigan Sentiment Analysis competition on Kaggle
- Twitter Sentiment Corpus by Niek Sanders

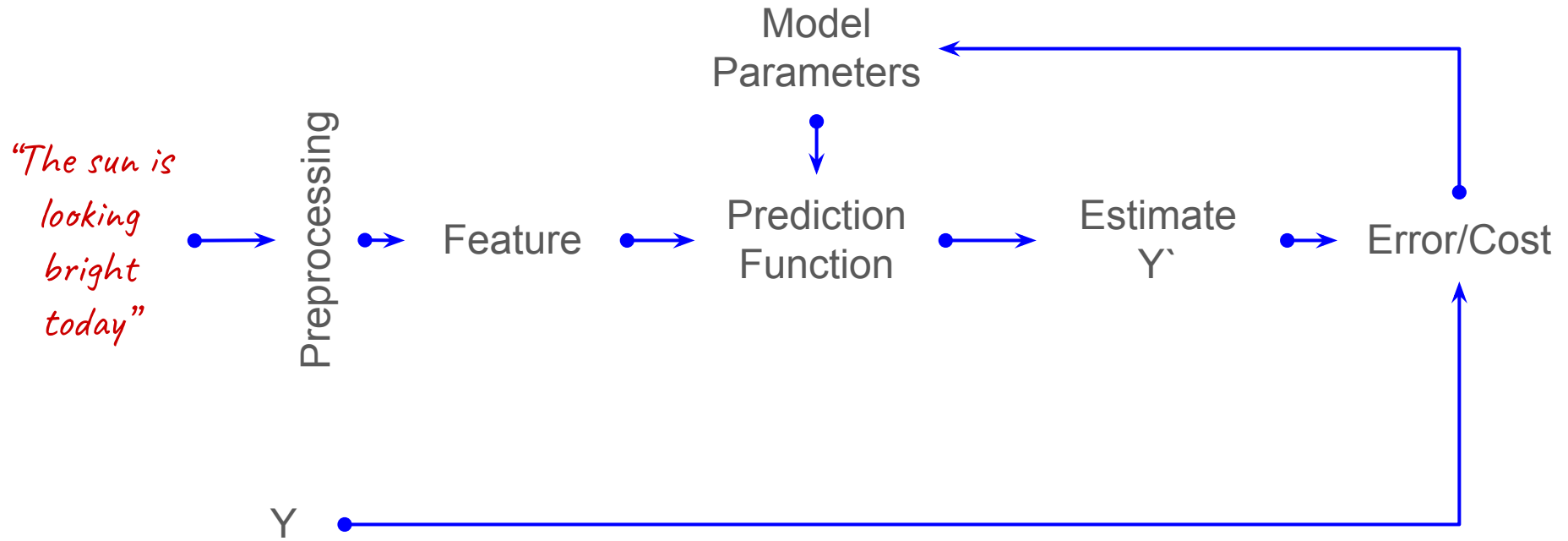
Methodology



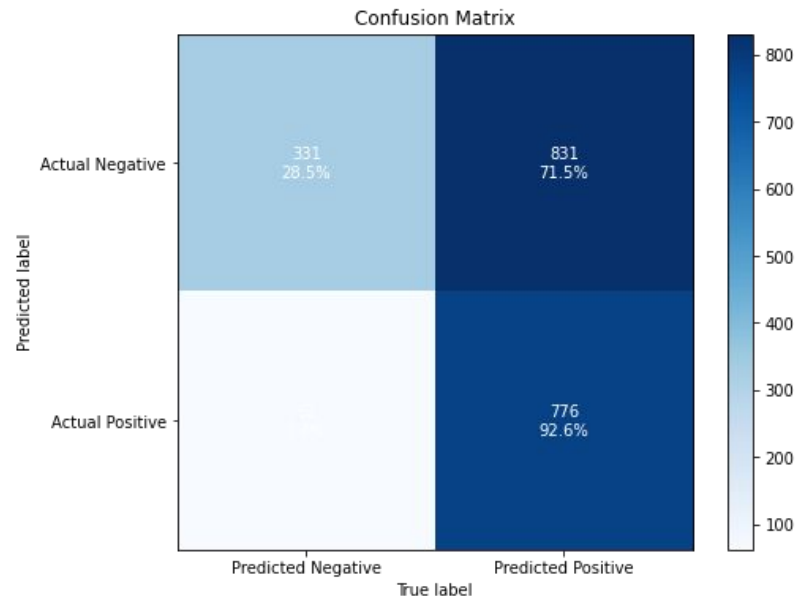
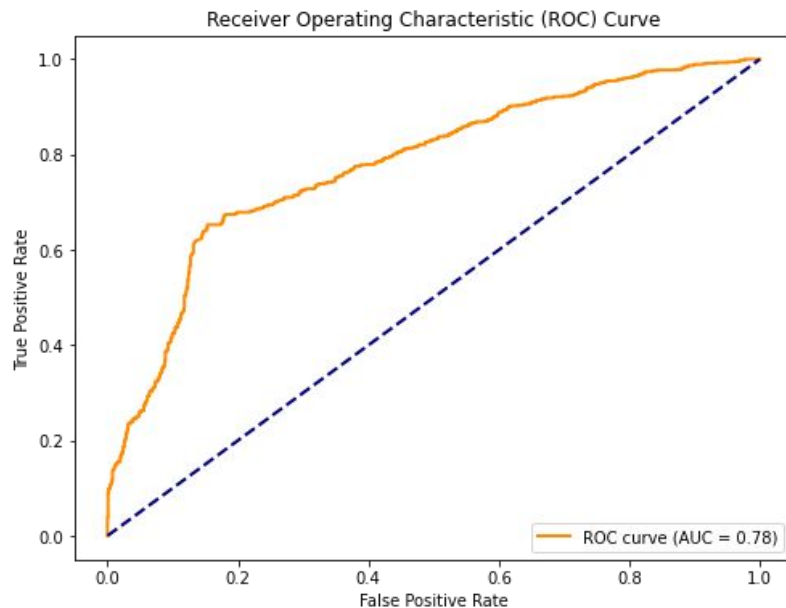
Methodology



Methodology



Results



Input Text: 'Man skips a wedding celebration to rescue a dog stuck in an overflowing river..'
Predicted Sentiment: negative

Takeaways

Data Preprocessing is Crucial

Effective data preprocessing is essential for successful sentiment analysis. Steps such as text cleaning, stop word removal, and stemming contribute to better feature representation.

Feature Engineering Matters

Extracting meaningful features is crucial for sentiment analysis. In this example, we used word frequencies as features.

Model Evaluation Beyond Accuracy

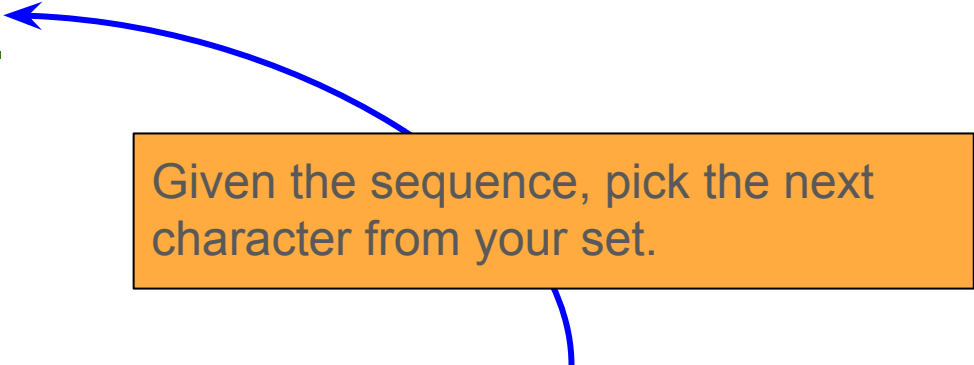
While accuracy is a common metric, it might not be sufficient for imbalanced datasets. Confusion matrices, precision, recall, and ROC curves provide a more comprehensive understanding of a model's performance, especially when dealing with false positives and false negatives.

Class Imbalance Impacts Model Performance

Imbalanced class distribution can lead to challenges, especially when one class is underrepresented. Techniques like stratified sampling, class weighting, or resampling can help address these issues and improve model performance.

Designing and implementing a bi-gram model for predicting next character

shilp_



Given the sequence, pick the next character from your set.

[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 .]

Indian names dataset

indian-names 0.3

```
pip install indian-names
```



Contains:

- first.male.names first.female.names last.names

Accessible:

- This package is released under a BSD 3-Clause License.

Data Source:

- Names are compiled in the following files through various online sources available publically.

Steps

- Obtain all possible bi-grams from the data
- Get the count of each bi-gram
- Re-structure the count as a matrix
- Obtain conditional probabilities
- Sample from a multinomial using the conditional probabilities
- Create new names

o	a1348	b373	c163	d447	e23	f8	g319	h321	i133	j283	k623	l114	m733	n601	o23	p747	q879	r1543	s190	t179	u546	v10	w0	x0	y86
a3134	aa93	ab235	ac36	ad564	ae0	af8	ag298	ah146	ai178	aj341	ak512	al756	am753	an2966	ao15	ap147	aq1543	ar774	as696	at155	au502	av51	aw3	ax3	ay349
b0	ba145	bb0	bc0	bd0	be11	bf0	bg0	bh568	bi45	bj0	bk0	bl7	bm0	bn0	bo16	bp0	br71	bs0	bt0	bu24	bv0	bw0	bx0	by0	
c0	ca0	cb0	cc6	cd0	ce0	cf0	cg0	ch365	ci0	cj0	ck14	cl0	cm0	cn0	co0	cp0	cr0	cs0	ct0	cu0	cv0	cw0	cx0	cy0	
d255	da331	db0	dc0	dd65	de339	df0	dg0	dh549	di320	dj0	dk9	dl0	dm24	dn0	do15	dp8	dr419	ds0	dt0	du95	dv0	dw11	dx0	dy32	
e74	ea0	eb10	ec0	ed81	ee739	ef7	eg23	eh38	ei0	ej56	ek139	el63	em124	en261	eo0	ep123	er108	es502	et315	eu0	ev171	ew0	ex0	ey58	
f0	fa7	fb0	fc0	fd0	fe0	ff0	fg0	fh0	fi0	fj0	fk0	fl0	fm0	fn0	fo8	fp0	fr0	fs0	ft0	fu8	fv0	fw0	fx0	fy0	
g44	ga279	gb0	gc0	gd42	ge84	gf0	gg10	gh76	gi121	gj0	gk0	gl0	gm0	gn24	go79	gp0	gr12	gs0	gt0	gu50	gv31	gw15	gx0	gy43	
h1110	ha241	hb8	hc11	hd8	he226	hf0	hg0	hh17	hi1166	hj0	hk27	hl7	hm67	hn96	ho98	hp9	hr183	hs0	ht32	hu317	hv28	hw139	hx0	hy94	
ia18	ib52	ic38	id182	ie0	if0	ig44	ih41	ii0	ij93	ik570	il289	im226	in588	io0	ip114	iq261	ir672	is782	it0	iu186	iv8	iw8	ix0	iy136	
j139	ja563	jb9	jc0	jd0	je169	jf0	jj0	jh19	ji79	jk30	jl7	jm10	jn16	jo8	jp22	jr0	js5	jt0	ju74	jv0	jw25	jx0	zy43		
k160	ka1207	kb0	kc0	kd0	ke91	kf0	kg0	kh178	ki142	kj0	kk0	kl18	km11	kn0	ko41	kp13	kq95	ks225	kt44	ku162	kv0	kx0	ky14		
l414	la528	lb11	lc0	ld9	le87	lf0	lg0	lh0	li271	lj13	lk29	ll48	lm0	ln0	lo59	lp30	lr5	ls9	lt6	lu0	lv0	lw13	lx0	ly53	
m222	ma1077	mb59	mc16	md34	me163	mf0	mg0	mh0	mi252	mj9	mk23	ml33	mn0	mo35	mp112	mq52	mr81	ms27	mt7	mu104	mv8	mw0	mx0	my22	
n634	na1017	nb19	nc30	nd664	ne237	nf0	ng124	nh22	ni955	nj237	nk268	nl0	nm35	nn38	no71	np0	nq10	nr207	nt346	nu225	nw44	nx0	ny12		
o71	oa0	ob5	oc11	od54	oe0	of0	og21	oh110	oi0	oj35	ok36	ol45	om87	on135	oo182	op47	or65	os56	ot69	ou5	ov32	ow0	ox0	oy0	
p112	pa429	pb0	pc0	pd0	pe37	pf0	pg0	ph0	pi100	pj0	pk0	pl0	pm9	pn22	po46	pp0	pr607	ps10	pt20	pu108	pv0	pw0	px0	py11	
q571	qa1523	qb9	qc20	qd58	qe323	rf0	rg29	rh0	ri980	rj33	rk13	rl8	rm63	rn42	ro81	rp26	rr0	rs108	rt94	ru296	rv82	rw10	rx0	ry69	
s79	sa554	sb0	sc0	sd0	se37	sf0	sg0	sh3156	si97	sj0	sk15	sl0	sm11	sn24	so113	sp0	sr22	ss0	st20	su557	sv27	sw72	sx0	sy0	
t432	ta860	tb0	tc0	td0	te123	tf0	tg0	th240	ti689	tj0	tk24	tl0	tm0	tn16	to51	tp0	tr102	ts7	tt51	tu74	tv0	tw10	tx0	ty106	
u232	ua4	ub87	uc54	ud245	ue0	uf0	ug42	uh38	ui0	uj88	uk73	ul197	um180	un306	uo0	up162	uq359	ur299	us113	uv14	uw9	ux0	uy6		
v168	va416	vb10	vc0	vd19	ve179	vf0	vg0	vh0	vi723	vj0	vk19	vl0	vm0	vn14	vo0	vp5	vr12	vs9	vt0	vu0	vv0	vw0	vx0	vy92	
w0	wa312	wb0	wc0	wd0	we16	wf0	wg0	wh0	wi25	wj0	wk0	wl0	wn0	wo11	wp0	wr0	ws0	wt0	wu0	wv0	ww0	wx0	wy0		
x0	xa0	xb0	xc0	xd0	xe0	xf0	xg0	xh0	xi0	xj0	xk0	xl0	xm3	xn0	xo0	xp0	xr0	xs0	xt0	xu0	xv0	xw0	xx0	xy0	
y173	ya897	yb0	yc0	yd0	ye8	yf0	yg0	yh0	yi0	yj0	yk0	yl0	ym0	yn0	yo48	yp6	yr10	ys13	yt0	yu71	yv0	yw0	yx0	yy0	

Takeaways

Sequential Dependency:

The bi-gram model emphasizes the importance of sequential dependencies in language. This concept highlights the significance of context in language modeling.

N-gram Modeling:

N-gram models, including bi-gram models, are valuable tools for understanding and predicting patterns in sequential data. They provide a balance between simplicity and capturing local context.

Probability Estimation:

The model estimates conditional probabilities of characters given their preceding characters.

Takeaways

Model Evaluation:

The effectiveness of the bi-gram model is assessed through evaluation metrics, such as accuracy and perplexity.

Challenges and Limitations:

Despite its simplicity and effectiveness, the bi-gram model has limitations, such as the inability to capture long-range dependencies.

Application to Text Generation:

Bi-gram models can be applied to text generation tasks, where the goal is to generate coherent and contextually relevant sequences of characters or words.

Food for thought

- Can you use such an approach for written language identification?
- Can you implement the same using a neural network?
- How do you quantify the quality of the generated data?

