Caleb L’Italien

01/20/2023

Project 1 CSC-483 Written Report

Table of Contents

Abstract
Example Outputs Based on Training Corpus
Perplexity Studies
Lambda Studies
Language Identification

I. ABSTRACT

This project focused on creating different models for n-gram language modeling. Two models were created, one uses basic n-gram modeling and the other that extends that basic model to incorporate interpolation. The project aims to show the difference in accuracy between different language models, with additional differences of lambda values and n-gram orders. The original training corpus used on the project was 167,204 lines of writings by William Shakespeare. The project then assessed the perplexity of comparing Shakespeare's sonnets, a New York Times article, and "The Communist Manifesto" (Karl Marx and Friedrich Engels) against the aforementioned training corpus. Next, different lambda values were tested in the interpolation model to notice the effects of weighting different n-grams. Finally, various language models were trained on text files of different languages, and then used to assess the accuracy of the models.

II. EXAMPLE OUTPUTS BASED ON TRAINING CORPUS

The following tests were completed on the basic n-gram model, without usage of add-k smoothing. There are two tests of 1000 characters per order of n-gram, each blocked off by a header. Headers have been bolded for ease of readability.

I. TRIGRAM

```
~~Firtur andly
Welf-Judrow:
To bath sarriefur mosee much nedgmer wity, hearbidle,
HAMIA:
Why, I cou ste, ithat I shemat my whim the andy likits not hour pord your of all him artice,
MARBUS:
LUCKLY:
And te, to con;
Birre by weressin; for thre pat whey therelves ithishermonot, aplyme itia?
BER:
You AURY Vaund face dut ond I;
RICHIO:
GLOCLENION:
On wing th pren.
FALINCES:
```

Frampat is mostir to desen,

Thou, fame, Clift shat nour th, thattersece sh ye:

Whease ext try itly so; sell.

And und war'd;

I shat th put but, sway.

BOTTONA:

JULT:

Goo.

OCTORIUS:

My har't, th as good oul thall whould lempar

And ence

To poinfernew ong men and do lady, sue contind, thinetich yould plem anse oureards,

What we's ther: nes re,

Tue! I'll'n!

In hat an ardow som ante you hures.

But behe ont surefus you ray.

Leould lableard.

Can to himmen.

GUS:

Yeave noth th I so evire?

Eve?

Sube an th far,

And was owly un low he Frands, dier lion canny, thow not ith an at a pured th all worl marm

Atterve faver?

In hits, st

II. TRIGRAM 2

~~Firef.

To will can he sh

Thisenot din do,

And that wilie?

The lours;

Of to Laden;

And thick, I'll mought:

Soople dem de of hered of thono he's wed.

QUELCONS:

To al hune kispen, thou

CALIO:

Here wor tor fortice

The a faready; suck, shas mor your yould cong sell songur Rosionow, muckiled I dest no ne the of to patrese pland tholy kingentalies.

OTSPUBET:

Thiss onowereed.

RANA:

Now weave book

an deme, bid my onglee shappip hand emblethearturn, the hise do so:--O duke.

CING HELLIFF:

Go, swee, and 'movoly dot my Cith ell boreds, of dinds, to' tor lesto is thalf,

Secomenned, willoor bre gre thee-seet mors ready, it he what ab sh st?

Aresto o'tionspectior conot upon and answer so mur befoebbuy,

Hery more

The sink sle ther quelve I us:

And dond mothe did end nare'em youdy: wither but ack:

If foreigh empanderead.

I muck.

MARD:

'Thisheepar,

Theet all the i' an metill cul ch?

Sebly ser forrommose libles, ch thich

have aver-gliatield's ates a postat res tood.

Lafold millight wingi ifew tin

III. 4-GRAM 1

~~~First hat's be thy scall your ple, that was ther a law
~~~

Ther him he no much

Ha, and widown with parall;

If younder with a many me to be cout haster far wits,

They place, you art o'er-fleep this

the standed may appart wipes for not?

MECATESBY:

I make agaits sight conjurable:

You in care andle

A ves; werefuse to breat wretly with be here and my lood glord, sir: do Rome ther whom our let'st of along eye of it shinger that

commort make a lean eye, of the

mar guilt thoughts. Ther by the whilesh or blow dothe he ways his murder idly as that me word,

Let and righboung,

Who, it to make it lords with,

That imployal that, no reconvocater I hat some, and I know nine ful double.

Firstake too;

Now sworld

The han thee odder frien with the day.

WARWICK:

My nation, ince confest thathe call for AEnear

Crievell come not in majest vains lunto gent you.

TROILUS:

I go.

POLONIO:

Cleoplear my petitualittle vior, fore words a many be is,

I commore mine so why

leanst Cities?

DUKE VIII:

Some there doning a

IV. 4-GRAM 2

~~~First sping yet that thind
~~~

tear the me are his in you shalth, Put bable
With on
The far mistript have pleave rounto the the Shall your so defend offer world I with of shall place, what thund
dangrave a she me of
Shall upon me.
Gent, fly, mation: and fashe funchee a trace
Comistranacy, made to ball with my for markneed,
Cous to be eage you dothe chich is parate,
Stay, I am not biaded. Come.
Oh, and
Get flown of they woes
Upon caresertall think yountony false make this not a he withou and a king cour?
ALO:
Your come rathe lord; behave.
PERICHARDINE:
Who proven, ta'en tentleman, and ching with to husband which I seem will your do but I thin, and utties musincler give nay, your shall vious lock'd blady?
JOAN LAUDIUS:
As herill tunder pawn they audio:
Let must int our cour me hers absoline expers.
ARVIRGILIP:
Wher.
CONSTER:
As ensure curself,
It him toge might. Bard, dant by Juliame to face
'Twas thy fries ours drawl teemi-world,
The knify ung paster a balm,
To weign.
If touch largar;

V. 5-GRAM 1

~~~~First, it is sister!
Uncle, fathese claim trow, Hal,
To beforester-darts,
Whose behold:
Sir Eglamour
world,
And his bed.

GARDINAL:

ARCHBISHOP OF ORLEANS:
My confess and mine
he back agenerate. Neither white for them I kiss'd: being sound we the get you lies
As if here! Lether.
You are air Tom's loverb when mistreason, 'tis deligion oaths, yours shepherd thy great the primals!
your and at fellow he to his ever bed in Pedro: we married Mariana, this were of a court, made they
this we may busin Brutus Lancastly foot, I have the of rebus' cares forth which then such aim.

MONTJOY:
Thou can; but now a kings in think it wantony, ere mocks so
In the like and the port? Cominish'd,
Swall: now, Sir Valent thou
know; where arm'd his and you laught to my grief hath me?
Try father.

YORK:
Excellusion; and yet 'tis thus,
Into makes
Whitsun evil heaves a crutch, a Christle;
You die.

TALBOT:
~~~~

Ay me:
How sharp as you consign tribute an earth, a mething country leaven, list.

BOYET:
A dog, an ender! i

VI. 5-GRAM 2

~~~~First Outlaws,
Would our bear their flows as I lie.

CARDINAND:
None,
Together: a past not upon thee wondemned these villain expedimentered with my march.

PETER:
Sir, and feard:
Sweet we'll to do.

PETRUCHIO:
Why true.
Sir, if
your taper I am so I be turn he crown!

AGAMEMNON:
Why, my lord your fathere?
O peevised
She content: I that clouds,
Which you make against the what I than of many thy thou yet the on of thy friends. My both you art time sphemous prepairs, he'll loit himself with you have
That was my good as that more; good to thren.

Gentlement;
~~~~

To-day soul
May execution,
You seest makes and is husband picked of thighness the man's bloody knight so I hear be
a discontent man,
Because I do the wolf
With enrich not sing England were is so you. Are their on
Blubber,
Am like!

GUIDERIUS:
Boult once did shall joy,
No more thy prophies heart wives monstanchole she wrap to be
That is to truth.

CYMBELINOR:
I praise?

PARIS:
Ay, marry, how now no other
of virginitent.
I throne, this! a

VII. 8-GRAM 1

~~~~~~~First Citizens
Gave me down.

PANDARUS:
Ay, and rocky-hard,
Who is that stature to the countess Richmond aims
At your honour.

BRUTUS:
~~~~~~~

I hope, lives.
There learn'd his letter in the town
way.

FRIAR LAURENCE:
I'll gilded snakes! Go, gentleman love;
For, till these: that widest that I am all the which not-withstanding on
After him not die by attorney but had he so?

DESDEMONA:
He for his queen. Fair gentle Warwick, hear you now? If
thou darest blood thus;
Whether Brutus hath had not one being answer.'

LENNOX:
May't please your parting. Nay, more: you'll answer'd him with silent.

FALSTAFF:
Let this lady dear!
Are you were villain that duty were favour in't is preferred
me like a turkeys in my face?
Women will I root of fire, and bear the authority.

FERDINAND:
Yes, good man!

ALCIBIADES:
So,
They set me a-weeping my
Lord Stafford fled therein they have him away: come, kissing-conduits of the duke's sure as successors else dead ere thou sayest true; we are great sir, you madded.

SIR AN

VIII. 8-GRAM 2

~~~~~~~First Citizen:
No, sir; but
the bodies
High on a sure forth and bloodless; and the business: but to making him; and it is,
Because Cassandra.

CASSIUS:
As one the sun?
No; dark spirit of night, what has a strangle with me, Pistol and free, so the army stood by whom our king
To those thorns are
ill-rooted; and you, sir: and
yet he lose.
Fie, fie upon thy breaths: your calls me the word--for 'tis
not worthier.

CASSIUS:
'Tis wonder infamy, but has left unseen grief:
The heaven, fasting-iron,
That know my death.

HECTOR:
I know by his liar, gone well met, my brother.
We have heard, and do my bellyful! Spit, fire! Corrupt bloody axe.
Ah, kill himself shallow be.
Most excellent! You have borne men; but we will be revenge.
Now the matter; I will work any. Will you learn to mock you it? by this deed: now thy traitor to the remembrance over-daring Typhon's breast,
And any thing health.
~~~~~~~

KATHARINA:
That Cassio came hither,
Then never dragons' spleenful mutiny and hose, well draw them;
And he wi

Each passage begins with the character “F”. This is because the input text starts with “F”, so each model's creation of sample text has a 100% probability of beginning with “F”. As the n-gram’s order increases, the first characters then begin with “First”. This is because the input text begins with “First”.

III. Perplexity Studies

Three additional texts were compared with the training corpus on the basis of compared perplexity. Each text was compared against the training corpus with three different n-gram orders and three different add-k smoothings. The calculation is “inf” (positive infinity) if the language model assigns any zero probabilities. Figure 1 (below) shows these results. As can be seen, any model that does not implement k-smoothing assigns at least one zero probability, and thus, outputs positive infinity.

	shakespeare_sonnets.txt	nytimes_article.txt	CommunistManifesto.rtfd

Basic Model (c, k)			
(1, 0)	inf	inf	inf
(1, 0.5)	6.519012549566104	8.229441845201531	6.930521076109183
(1, 1)	7.471945586685537	8.720005077879373	7.845851929425858
(2, 0)	inf	inf	inf
(2, 0.5)	6.00509730895666	8.905732969572123	7.899770257451494
(2, 1)	7.162228272954077	9.237199477324491	8.54537551617449
(4, 0)	inf	inf	inf
(4, 0.5)	6.503146358911397	9.621066645203284	10.263880962570392
(4, 1)	7.544410590991413	9.7389792556104	10.137088562954675
Interpolation (c, k)			
(1, 0)	inf	inf	inf
(1, 0.5)	7.964440932103137	10.37829776796252	9.920202924542274
(1, 1)	8.607152397058284	10.342822784040376	10.216856269850483
(2, 0)	inf	inf	inf
(2, 0.5)	7.100004955169095	9.676523344906748	9.004931786589953
(2, 1)	8.017366103560562	9.862564685816896	9.520741635155217
(4, 0)	inf	inf	inf
(4, 0.5)	6.7201826234176725	9.433081663075273	9.052623653221515
(4, 1)	7.741542952851087	9.687803722166661	9.5080114484722

Figure 1

IV. Lambda Studies

Lambda values are used in this project to weight different n-grams in interpolated language models and calculate accordingly adjusted probabilities. . The language models represented in Figure 2 (below) were given three texts of ‘abab’, ‘abcd’, and ‘abdcabb’. Lambda values are presented as a list, with the first value being the unigram’s weight and the last being the highest order n-gram. The calculated probabilities of ‘c’, given a context of ‘ab’ are shown below.

	[0.5, 0.5]	**[0.4, 0.6]**	**[0.1, 0.9]**	**[0.9, 0.1]**
(c, k)				
(1, 0)	0.191666666666665	0.203333333333334	0.238333333333334	0.145
(1, 0.5)	0.1985294117647059	0.2088235294117647	0.239705882352941	0.1573529411764706
(1, 1)	0.203947368421053	0.2131578947368421	0.240789473684211	0.1671052631578947
	[0.333, 0.333, 0.333]	**[0.15, 0.3, 0.55]**	**[0.1, 0.2, 0.7]**	**[0.7, 0.2, 0.1]**
(2, 0)	0.2109	0.2325	0.238333333333334	0.168333333333333
(2, 0.5)	0.2154705882352941	0.2345588235294118	0.239705882352941	0.1779411764705882
(2, 1)	0.219078947368421	0.2361842105263158	0.240789473684211	0.1855263157894737
	[0.2, 0.2, 0.2, 0.2, 0.2]	**[0.05, 0.1, 0.15, 03225, 0.375]**	**[0.04, 0.06, 0.1, 0.2, 0.6]**	**[0.6, 0.2, 0.1, 0.06, 0.04]**
(4, 0)	0.226666666666667	0.238166666666667	0.245333333333333	0.18
(4, 0.5)	0.2294117647058824	0.2382352941176471	0.2458823529411765	0.1882352941176471
(4, 1)	0.2315789473684211	0.238289473684211	0.2463157894736842	0.1947368421052632

Figure 2

V. Language Identification

One of the goals of language models like n-gram models is the ability to identify the language of a text. Eighteen different language models were trained on eight different languages. The language models varied according to the (c, k) values used in figure 1 and 2. They also varied on using interpolation or not. The models generally reported accurately, especially those with higher n-gram order.