Language Modeling

CMSC 473/673 Spring 2017
Bryan Wilkinson

Discussion of Kilgarriff Paper

- Does anyone disagree with Kilgarriff?
- What are the alternatives?
  - UkWaC and WaCkypedia
  - UMBC WebBase

Language Modeling

- A Language Model is a collection of probabilities of some unit of language
    - Word and character level models are the most common.
- Why do we need to know the probability of language?
    - Which spelling is more likely?
    - Which verb is more likely?
- What is the probability of language?
    - Of a word?
    - Of a sentence?

Counting Words

- Before we can calculate probability, we need to count things.
  - Seems like a simple task but,
- What is a word?
  - Still an open question, but mostly agreed upon
  - Is *a lot* one word or two?
  - How about *post office*?
- English is easy to count
  - What about a language with more complex morphology (We will talk more about this later)
  - Iyewičhamačheča (Lakota “I resemble you” or “you resemble me”)
- Other minor decisions
  - Does capitalization matter?
  - How about plurals? Or other endings?

Probability of a Word

- Counting is the hard part

$$P(w) = \frac{C(w)}{N}$$

In Class Activity

Markov Assumption

- The future doesn't depend on the past
  - But the future **does** depend on the past, so we will just limit the amount of past we look at
- In NLP this means when calculating the probability of a word given a context, we often just look at the preceding one or two words. (or none!)
- In terms of probability this looks like
  - P(*homework* | *the dog ate my*) ≈ P(*homework*)
  - P(*homework* | *the dog ate my*) ≈ P(*homework* | *my*)
  - P(*homework* | *the dog ate my*) ≈ P(*homework* | *ate my*)
- We call these small chunks of words N-grams
  - Unigram = homework
  - Bigram = my homework
  - Trigram = ate my homework
- What are all the bigrams and trigrams in the sentence "the dog ate my homework"

Probability of an N-Gram

- Calculating probability of an N-Gram is almost as easy as a word
- Technically we are estimating using Maximum Likelihood Estimation

$$P(w|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)}$$

$$P(w|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

Probability of a Sentence

- To calculate the probability of a sentence, we use the chain rule

$$P(w_1 w_2 w_3 ... w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)...P(w_n|w_1, w_2, w_3 ... w_{n-1})$$

- P(Today is the second day of class) = P(Today | <start>) * P(is | Today <start>) * P(the | is Today <start>) * P(second | the is Today <start>) * P(day | second the is Today <start>) * P(of | day second the is Today <start>) * P(class | of day second the is Today <start>) * P(<end> | class of day second the is Today <start>)

Probability of a Sentence

- To calculate the probability of a sentence, we use the chain rule

$$P(w_1 w_2 w_3 ... w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2)...P(w_n|w_{n-1})$$

- P(Today is the second day of class) ≈ P(Today | <start>) * P(is | Today) * P(the | is) * P(second | the) * P(day | second) * P(of | day) * P(class | of) * P(<end> | class)

Probability of a Sentence

- The following counts are from Google Books N-Gram Corpus v2
    - Via corpus.byu.edu

C(<start>,Today)	
C(Today, is)	49215
C(is, the)	101249451
C(the, second)	17216997
C(second, day)	396983
C(day, of)	5639617
C(of, class)	672495
C(class,<end>)	2703083

C(<start>)	
C(Today)	3251311
C(is)	1368855691
C(the)	7726878625
C(second)	42738935
C(day)	78211120
C(of)	5035745089
C(class)	27260152
C(<end>)	7805955213

Probability of a Sentence

P(Today | <start>)	
P(is | Today)	
P(the | is)	
P(second | the)	
P(day | second)	
P(of | day)	
P(class | of)	
P(<end> | class)	

Probability of a Sentence

- What additional data would we need to calculate the probability of :
    - Today is the eighth day of class ?
    - Today is the 57th day of class ?
    - Today is the somethingth day of class?

Evaluating Language Models

- LMs give us probabilities of a sentence
  - How do we know if they are working well
  - What is the “correct” probability of a given sentence
- We could evaluate them using some task that has LMs as part of it
  - Test a machine translation, speech recognition, etc system and see which LM does better
  - Can conflates any errors
- So we try to evaluate LMs as a standalone system
  - This is hard
  - We do this using perplexity, a measure from information theory.

Perplexity

- Wikipedia (and probably statistics books) define perplexity as 2 raised to the entropy of the distribution
    - Technically correct, but not very intuitive (to me at least)
- Jurafsky gives the definition as the probability an LM gives to a test set, normalized by the number of words in that test set.
    - The idea is that words and sequences of words that actually exist in text should have a relatively high probability
- The assumption is that perplexity should correlate to performance on the end task
    - Often is the case!
    - But not always!

$$\sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$$

Perplexity Examples

- This data comes from a system I am working on that has a perplexity of about 3

Input	Output
Who are you ?	Who are you ?
This caused problem like the formation and growth of slums which most of the time is not safe due to their unhealthy environment .	This caused problem like the both and growth of of which most of the time not not not be to their . .
Tokyo is often referred to as a city, but is officially known and governed as a "metropolitan prefecture" .	New is often referred to as a city , but is officially known and workforce as a " the prefecture " .
The Tokyo metropolitan government administers the 23 Special Wards of Tokyo	The financially of government classified the 00 Special Wards of New

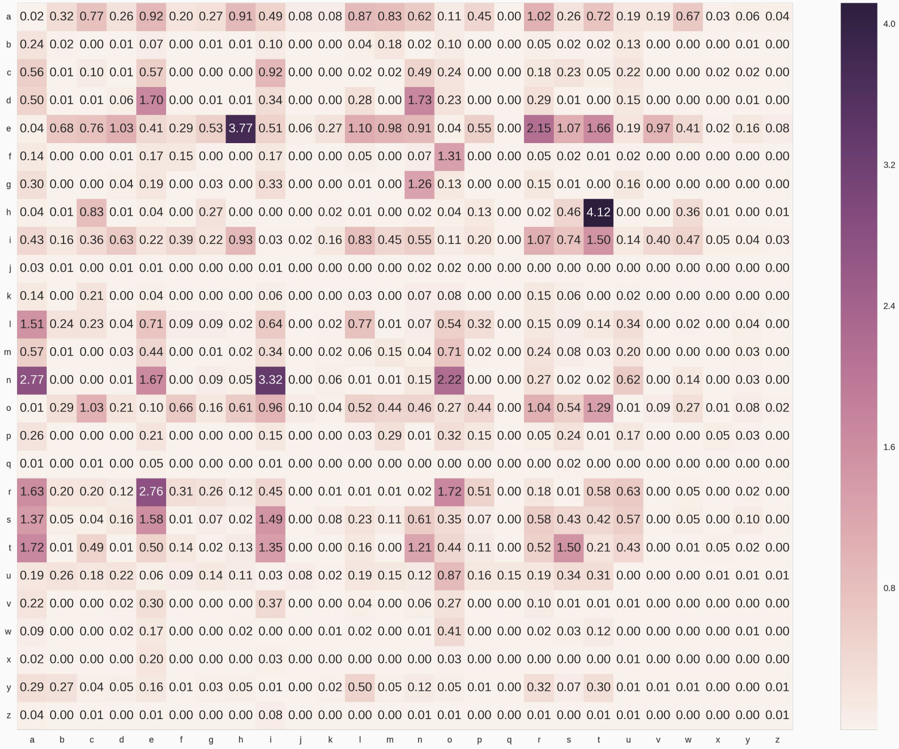
Character Level Language Models

- What are they useful for?
  - Language ID
  - Spell Check
- They are smaller
  - Just need to count letters and a few other symbols
- Becoming more popular with Neural Networks
  - The Unreasonable Effectiveness of Recurrent Neural Networks

- Do they represent actual linguistic units?

Character LM Example

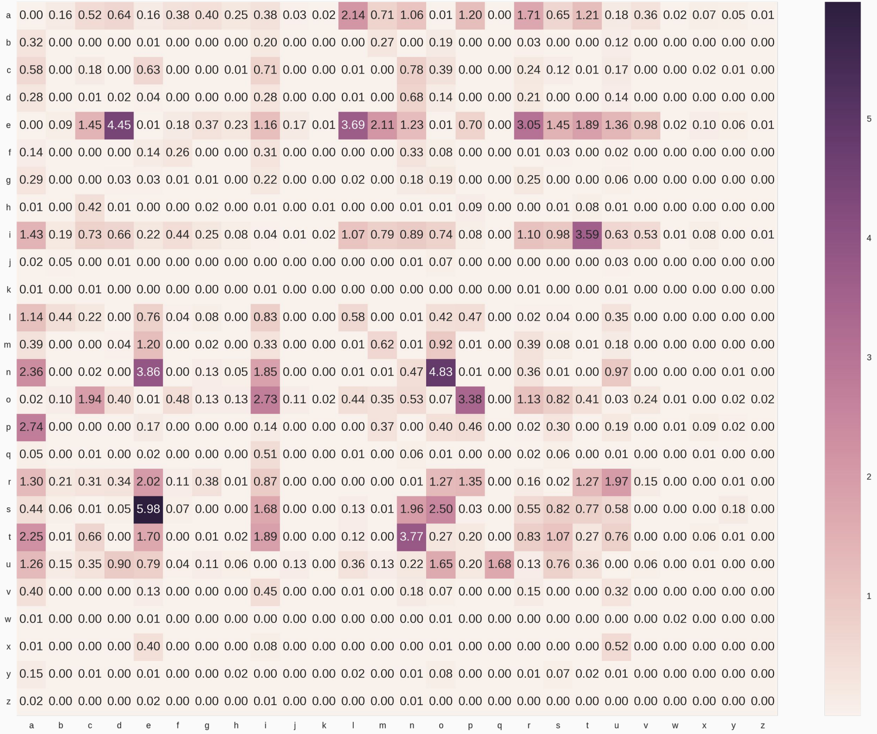
- Let’s say we want to build a simple language ID system
- Bigram model over 2GB of wackypeadia produces the following counts



	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	0.02	0.32	0.77	0.26	0.92	0.20	0.27	0.91	0.49	0.08	0.08	0.87	0.83	0.62	0.11	0.45	0.00	1.02	0.26	0.72	0.19	0.19	0.67	0.03	0.06	0.04
b	0.24	0.02	0.00	0.01	0.07	0.00	0.01	0.01	0.10	0.00	0.00	0.04	0.18	0.02	0.10	0.00	0.00	0.05	0.02	0.02	0.13	0.00	0.00	0.00	0.01	0.00
c	0.56	0.01	0.10	0.01	0.57	0.00	0.00	0.00	0.92	0.00	0.00	0.02	0.02	0.49	0.24	0.00	0.00	0.18	0.23	0.05	0.22	0.00	0.00	0.02	0.02	0.00
d	0.50	0.01	0.01	0.06	1.70	0.00	0.01	0.01	0.34	0.00	0.00	0.28	0.00	1.73	0.23	0.00	0.00	0.29	0.01	0.00	0.15	0.00	0.00	0.00	0.01	0.00
e	0.04	0.68	0.76	1.03	0.41	0.29	0.53	3.77	0.51	0.06	0.27	1.10	0.98	0.91	0.04	0.55	0.00	2.15	1.07	1.66	0.19	0.97	0.41	0.02	0.16	0.08
f	0.14	0.00	0.00	0.01	0.17	0.15	0.00	0.00	0.17	0.00	0.00	0.05	0.00	0.07	1.31	0.00	0.00	0.05	0.02	0.01	0.02	0.00	0.00	0.00	0.00	0.00
g	0.30	0.00	0.00	0.04	0.19	0.00	0.03	0.00	0.33	0.00	0.00	0.01	0.00	1.26	0.13	0.00	0.00	0.15	0.01	0.00	0.16	0.00	0.00	0.00	0.00	0.00
h	0.04	0.01	0.83	0.01	0.04	0.00	0.27	0.00	0.00	0.00	0.02	0.01	0.00	0.02	0.04	0.13	0.00	0.02	0.46	4.12	0.00	0.00	0.36	0.01	0.00	0.01
i	0.43	0.16	0.36	0.63	0.22	0.39	0.22	0.93	0.03	0.02	0.16	0.83	0.45	0.55	0.11	0.20	0.00	1.07	0.74	1.50	0.14	0.40	0.47	0.05	0.04	0.03
j	0.03	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
k	0.14	0.00	0.21	0.00	0.04	0.00	0.00	0.00	0.06	0.00	0.00	0.03	0.00	0.07	0.08	0.00	0.00	0.15	0.06	0.00	0.02	0.00	0.00	0.00	0.00	0.00
l	1.51	0.24	0.23	0.04	0.71	0.09	0.09	0.02	0.64	0.00	0.02	0.77	0.01	0.07	0.54	0.32	0.00	0.15	0.09	0.14	0.34	0.00	0.02	0.00	0.04	0.00
m	0.57	0.01	0.00	0.03	0.44	0.00	0.01	0.02	0.34	0.00	0.02	0.06	0.15	0.04	0.71	0.02	0.00	0.24	0.08	0.03	0.20	0.00	0.00	0.00	0.03	0.00
n	2.77	0.00	0.00	0.01	1.67	0.00	0.09	0.05	3.32	0.00	0.06	0.01	0.01	0.15	2.22	0.00	0.00	0.27	0.02	0.02	0.62	0.00	0.14	0.00	0.03	0.00
o	0.01	0.29	1.03	0.21	0.10	0.66	0.16	0.61	0.96	0.10	0.04	0.52	0.44	0.46	0.27	0.44	0.00	1.04	0.54	1.29	0.01	0.09	0.27	0.01	0.08	0.02
p	0.26	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.15	0.00	0.00	0.03	0.29	0.01	0.32	0.15	0.00	0.05	0.24	0.01	0.17	0.00	0.00	0.05	0.03	0.00
q	0.01	0.00	0.01	0.00	0.05	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
r	1.63	0.20	0.20	0.12	2.76	0.31	0.26	0.12	0.45	0.00	0.01	0.01	0.01	0.02	1.72	0.51	0.00	0.18	0.01	0.58	0.63	0.00	0.05	0.00	0.02	0.00
s	1.37	0.05	0.04	0.16	1.58	0.01	0.07	0.02	1.49	0.00	0.08	0.23	0.11	0.61	0.35	0.07	0.00	0.58	0.43	0.42	0.57	0.00	0.05	0.00	0.10	0.00
t	1.72	0.01	0.49	0.01	0.50	0.14	0.02	0.13	1.35	0.00	0.00	0.16	0.00	1.21	0.44	0.11	0.00	0.52	1.50	0.21	0.43	0.00	0.01	0.05	0.02	0.00
u	0.19	0.26	0.18	0.22	0.06	0.09	0.14	0.11	0.03	0.08	0.02	0.19	0.15	0.12	0.87	0.16	0.15	0.19	0.34	0.31	0.00	0.00	0.00	0.01	0.01	0.01
v	0.22	0.00	0.00	0.02	0.30	0.00	0.00	0.00	0.37	0.00	0.00	0.04	0.00	0.06	0.27	0.00	0.00	0.10	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
w	0.09	0.00	0.00	0.02	0.17	0.00	0.00	0.02	0.00	0.00	0.01	0.02	0.00	0.01	0.41	0.00	0.00	0.02	0.03	0.12	0.00	0.00	0.00	0.00	0.01	0.00
x	0.02	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
y	0.29	0.27	0.04	0.05	0.16	0.01	0.03	0.05	0.01	0.00	0.02	0.50	0.05	0.12	0.05	0.01	0.00	0.32	0.07	0.30	0.01	0.01	0.01	0.00	0.00	0.01
z	0.04	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01

Character LM Example

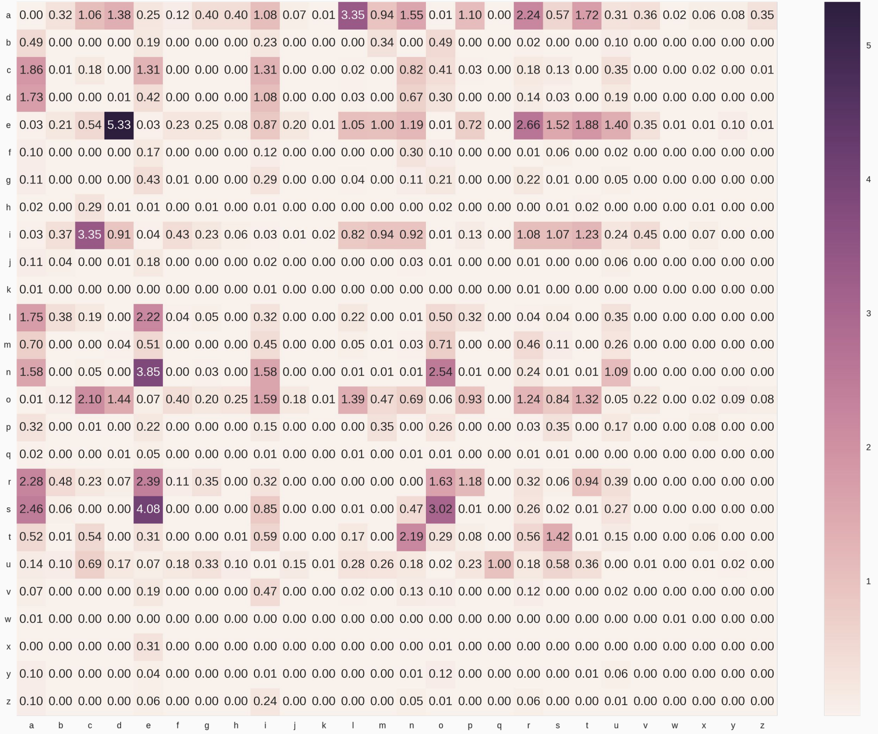
- Let’s say we want to build a simple language ID system
- Bigram model over 2GB of French UN text produces the following counts



	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	0.00	0.16	0.52	0.64	0.16	0.38	0.40	0.25	0.38	0.03	0.02	2.14	0.71	1.06	0.01	1.20	0.00	1.71	0.65	1.21	0.18	0.36	0.02	0.07	0.05	0.01
b	0.32	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.27	0.00	0.19	0.00	0.00	0.03	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00
c	0.58	0.00	0.18	0.00	0.63	0.00	0.00	0.01	0.71	0.00	0.00	0.01	0.00	0.78	0.39	0.00	0.00	0.24	0.12	0.01	0.17	0.00	0.00	0.02	0.01	0.00
d	0.28	0.00	0.01	0.02	0.04	0.00	0.00	0.00	0.28	0.00	0.00	0.01	0.00	0.68	0.14	0.00	0.00	0.21	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00
e	0.00	0.09	1.45	4.45	0.01	0.18	0.37	0.23	1.16	0.17	0.01	3.69	2.11	1.23	0.01	0.70	0.00	3.05	1.45	1.89	1.36	0.98	0.02	0.10	0.06	0.01
f	0.14	0.00	0.00	0.00	0.14	0.26	0.00	0.00	0.31	0.00	0.00	0.00	0.00	0.33	0.08	0.00	0.00	0.01	0.03	0.00	0.02	0.00	0.00	0.00	0.00	0.00
g	0.29	0.00	0.00	0.03	0.03	0.01	0.01	0.00	0.22	0.00	0.00	0.02	0.00	0.18	0.19	0.00	0.00	0.25	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00
h	0.01	0.00	0.42	0.01	0.00	0.00	0.02	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.09	0.00	0.00	0.01	0.08	0.01	0.00	0.00	0.00	0.00	0.00
i	1.43	0.19	0.73	0.66	0.22	0.44	0.25	0.08	0.04	0.01	0.02	1.07	0.79	0.89	0.74	0.08	0.00	1.10	0.98	3.59	0.63	0.53	0.01	0.08	0.00	0.01
j	0.02	0.05	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00
k	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
l	1.14	0.44	0.22	0.00	0.76	0.04	0.08	0.00	0.83	0.00	0.00	0.58	0.00	0.01	0.42	0.47	0.00	0.02	0.04	0.00	0.35	0.00	0.00	0.00	0.00	0.00
m	0.39	0.00	0.00	0.04	1.20	0.00	0.02	0.00	0.33	0.00	0.00	0.01	0.62	0.01	0.92	0.01	0.00	0.39	0.08	0.01	0.18	0.00	0.00	0.00	0.00	0.00
n	2.36	0.00	0.02	0.00	3.86	0.00	0.13	0.05	1.85	0.00	0.00	0.01	0.01	0.47	4.83	0.01	0.00	0.36	0.01	0.00	0.97	0.00	0.00	0.00	0.01	0.00
o	0.02	0.10	1.94	0.40	0.01	0.48	0.13	0.13	2.73	0.11	0.02	0.44	0.35	0.53	0.07	3.38	0.00	1.13	0.82	0.41	0.03	0.24	0.01	0.00	0.02	0.02
p	2.74	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.37	0.00	0.40	0.46	0.00	0.02	0.30	0.00	0.19	0.00	0.01	0.09	0.02	0.00
q	0.05	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.51	0.00	0.00	0.01	0.00	0.06	0.01	0.00	0.00	0.02	0.06	0.00	0.01	0.00	0.00	0.01	0.00	0.00
r	1.30	0.21	0.31	0.34	2.02	0.11	0.38	0.01	0.87	0.00	0.00	0.00	0.00	0.01	1.27	1.35	0.00	0.16	0.02	1.27	1.97	0.15	0.00	0.00	0.01	0.00
s	0.44	0.06	0.01	0.05	5.98	0.07	0.00	0.00	1.68	0.00	0.00	0.13	0.01	1.96	2.50	0.03	0.00	0.55	0.82	0.77	0.58	0.00	0.00	0.00	0.18	0.00
t	2.25	0.01	0.66	0.00	1.70	0.00	0.01	0.02	1.89	0.00	0.00	0.12	0.00	3.77	0.27	0.20	0.00	0.83	1.07	0.27	0.76	0.00	0.00	0.06	0.01	0.00
u	1.26	0.15	0.35	0.90	0.79	0.04	0.11	0.06	0.00	0.13	0.00	0.36	0.13	0.22	1.65	0.20	1.68	0.13	0.76	0.36	0.00	0.06	0.00	0.01	0.00	0.00
v	0.40	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.45	0.00	0.00	0.01	0.00	0.18	0.07	0.00	0.00	0.15	0.00	0.00	0.32	0.00	0.00	0.00	0.00	0.00
w	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
x	0.01	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.52	0.00	0.00	0.00	0.00	0.00
y	0.15	0.00	0.01	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.01	0.08	0.00	0.00	0.01	0.07	0.02	0.01	0.00	0.00	0.00	0.00	0.00
z	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Character LM Example

- Let’s say we want to build a simple language ID system
- Bigram model over 2GB of Spanish UN text produces the following counts



	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	0.00	0.32	1.06	1.38	0.25	0.12	0.40	0.40	1.08	0.07	0.01	3.35	0.94	1.55	0.01	1.10	0.00	2.24	0.57	1.72	0.31	0.36	0.02	0.06	0.08	0.35
b	0.49	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.34	0.00	0.49	0.00	0.00	0.02	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00
c	1.86	0.01	0.18	0.00	1.31	0.00	0.00	0.00	1.31	0.00	0.00	0.02	0.00	0.82	0.41	0.03	0.00	0.18	0.13	0.00	0.35	0.00	0.00	0.02	0.00	0.01
d	1.73	0.00	0.00	0.01	0.42	0.00	0.00	0.00	1.08	0.00	0.00	0.03	0.00	0.67	0.30	0.00	0.00	0.14	0.03	0.00	0.19	0.00	0.00	0.00	0.00	0.00
e	0.03	0.21	0.54	5.33	0.03	0.23	0.25	0.08	0.87	0.20	0.01	1.05	1.00	1.19	0.01	0.72	0.00	2.66	1.52	1.88	1.40	0.35	0.01	0.01	0.10	0.01
f	0.10	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.30	0.10	0.00	0.00	0.01	0.06	0.00	0.02	0.00	0.00	0.00	0.00	0.00
g	0.11	0.00	0.00	0.00	0.43	0.01	0.00	0.00	0.29	0.00	0.00	0.04	0.00	0.11	0.21	0.00	0.00	0.22	0.01	0.00	0.05	0.00	0.00	0.00	0.00	0.00
h	0.02	0.00	0.29	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.01	0.00	0.00
i	0.03	0.37	3.35	0.91	0.04	0.43	0.23	0.06	0.03	0.01	0.02	0.82	0.94	0.92	0.01	0.13	0.00	1.08	1.07	1.23	0.24	0.45	0.00	0.07	0.00	0.00
j	0.11	0.04	0.00	0.01	0.18	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.00	0.01	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00
k	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
l	1.75	0.38	0.19	0.00	2.22	0.04	0.05	0.00	0.32	0.00	0.00	0.22	0.00	0.01	0.50	0.32	0.00	0.04	0.04	0.00	0.35	0.00	0.00	0.00	0.00	0.00
m	0.70	0.00	0.00	0.04	0.51	0.00	0.00	0.00	0.45	0.00	0.00	0.05	0.01	0.03	0.71	0.00	0.00	0.46	0.11	0.00	0.26	0.00	0.00	0.00	0.00	0.00
n	1.58	0.00	0.05	0.00	3.85	0.00	0.03	0.00	1.58	0.00	0.00	0.01	0.01	0.01	2.54	0.01	0.00	0.24	0.01	0.01	1.09	0.00	0.00	0.00	0.00	0.00
o	0.01	0.12	2.10	1.44	0.07	0.40	0.20	0.25	1.59	0.18	0.01	1.39	0.47	0.69	0.06	0.93	0.00	1.24	0.84	1.32	0.05	0.22	0.00	0.02	0.09	0.08
p	0.32	0.00	0.01	0.00	0.22	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.35	0.00	0.26	0.00	0.00	0.03	0.35	0.00	0.17	0.00	0.00	0.08	0.00	0.00
q	0.02	0.00	0.00	0.01	0.05	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
r	2.28	0.48	0.23	0.07	2.39	0.11	0.35	0.00	0.32	0.00	0.00	0.00	0.00	0.00	1.63	1.18	0.00	0.32	0.06	0.94	0.39	0.00	0.00	0.00	0.00	0.00
s	2.46	0.06	0.00	0.00	4.08	0.00	0.00	0.00	0.85	0.00	0.00	0.01	0.00	0.47	3.02	0.01	0.00	0.26	0.02	0.01	0.27	0.00	0.00	0.00	0.00	0.00
t	0.52	0.01	0.54	0.00	0.31	0.00	0.00	0.01	0.59	0.00	0.00	0.17	0.00	2.19	0.29	0.08	0.00	0.56	1.42	0.01	0.15	0.00	0.00	0.06	0.00	0.00
u	0.14	0.10	0.69	0.17	0.07	0.18	0.33	0.10	0.01	0.15	0.01	0.28	0.26	0.18	0.02	0.23	1.00	0.18	0.58	0.36	0.00	0.01	0.00	0.01	0.02	0.00
v	0.07	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.47	0.00	0.00	0.02	0.00	0.13	0.10	0.00	0.00	0.12	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
w	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
x	0.00	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
y	0.10	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.12	0.00	0.00	0.00	0.00	0.01	0.06	0.00	0.00	0.00	0.00	0.00
z	0.10	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.05	0.01	0.00	0.00	0.06	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00

Character LM Experiment

- Let’s try the following sentences and see the results
  - Today is the second day of class
  - La universidad tienen 50 años.
  - Marius se sentit fier de cet inconnu.
  - Melissa Villaseñor is on SNL
- What happens when we put in a language we don’t have a model for
  - What can this tell us?

Reminder

- HW 0 “Due” Tuesday