Project1 Part1 Sentence Examples and Analysis
CS4740 Team members: Ransen Niu, Weicheng Yu

Problem Statement

Part 1 of this project uses unsmoothed unigram and bigram models to generate random sentences.
We generated three sentences for each model we implemented. Since we have one unigram model and two bigram models (we compared these two as well), there are 63 sentences generates.
For the following sentence examples, we assign each one a symbol to represent if the sentence is not meaningful "&", a little meaningful "@", somehow meaningful "#" at the beginning of the sentence.

Unsmoothed Unigram and Sentence Generation

Sentence Examples

autos:
Unigram model:
& 0. like , over , in real questions All the only US Korea Well one 50 wondering My When called the be con .
@ 1. OS that you a , it has to ntu.ac.sg good tried an Spirit anything in the up hold with it If pump to MG mail Civic I cars considered ?
& 2. on 3272 it cactus.org roomy mike , new price , Philip the to .

graphics:
Unigram model:
& 0. with , frames to that was up a last if n't in domain case ones that from , to !
@ 1. ; check programs , commands said on .
& 2. there method .

religion:
Unigram model:
& 0. to .
& 1. governments he ?
& 2. of of of , Je morality bondage slave " n't mother BYUVM.BITNET steaks be Norse that but , battle prayer Latter I : is stiff mother to key has word way authorities .

motorcycles:
Unigram model:
& 0. execution wish could to ."&"
@ 1. sanjay a That if no input Mx cheaper , I is the every 's my Dean the heard !
& 2. also suppossed had sanding mostly to court protected .

medicine:
Unigram model:
& 0. Inc frequent is section The IN does too T to that artificial mention .. email toner I ny compared and middles high 'm if did practice clinical Deficiency exercising possible Sounds

minutes the disease total of wonders have associated uncontrolled , ub .
& 1. to to their I ones intestinal typing is , he Endoscopy on is : ARE from by infection science beginning Telecom evaluate force configurable described several n't their few series has now update , 's one is called picked good the to , improved cadre.dsl.pitt.edu helpful steak .
& 2. Natalja or this life for in min that I RSI to public , for patients .

atheism:
Unigram model:
@ 0. the absolute be never In major Here although grandeur " I moment evidence Russell Jennifer original desk you suggested am a .
& 1. The in goal may But of that for he L .
& 2. not Jew you , quite expect to true It at " .

space:
Unigram model:
& 0. SPACE serious Graydon facts arc comet pinch to .
@ 1. rocket this Just get End speed 0815 .
& 2. , be Aeronautics Atlantic km foot !

Sentence Analysis

	unigram model
not meaningful "&"	16
a little meaningful "@"	5
some meaningful "#"	0

As we can see from the examples above, we have 21 sentences total, and there are only 5 marked as a little meaningful "@" and the rest are all not meaningful"&". It's clear that the sentences are just randomly words and usually have no meanings.

Unsmoothed Bigram and Sentence Generation

We built two bigram models with different specifications to compare which one is better.

For the first model, we go through the element (word) in the wordlist two by two to count the frequency of each pair of them. When generating a sentence, we randomly choose the beginning word according to the unigram model. Afterwards, we use the bigram model to randomly sample which word is going after the previous word. And the sentence ends when one of the "stop" punctuations is met.

For the second model, there is some preprocessing. We prepend a beginning marker "<" at the beginning of the wordlist and we add the beginning marker "<" after each "stop" punctuation. Then we know once we encounter a "<", it means the end of the previous sentence and the beginning of a new sentence. Then we generate this second bigram model the same way as the first bigram model, but we should notice that here the beginning marker "<" is counted as a new word type. When generating a sentence, we always make "" as the beginning of the sentence, and then use this second bigram model to randomly sample which word is going after

the previous word. And the sentence ends when one of the "stop" punctuations is met. Finally, we remove the beginning marker "<" from the generated sentence because the marker is not really a part of a sentence.

Sentence Examples

autos:
Bigram model 1:
& 0. and cheap R12 systems that happened to threaten to deal with those fancy but very ugly very well built , the engine and in first generation of the product in the oil change , which has an electronic mailing lists for a car that people who will be in the commercial .
@ 1. rust on hand or just something like my car myself , particularly interested in the bottom line and see no leaks in the latest 911s , snip , iguanas , exposing a cold wet My old 2 days on end alignment done all the manufacturer , and a numerically higher price of my second floor like a cure for just building and was n't know BMW : " O'Leary said " Through a compact will probably about 20 , it 's a sports car i also named to redline the dealer shop manager at a little more so it that 1 New Jersey I claimed they consume that I believe it is that anybody 's probably get gouged right ?
@ 2. the adsorbate water to avoid a voltage regulator in a shell , with experience with a new camry station wagons .

Bigram model 2:
& 0. mil US Military sites if it .
@ 1. If you honk the greatest predators that thing you .
& 2. OH 44334 0582 remember , overdrive , sayoonara o .

graphics:
Bigram model 1:
& 0. Routines are lucky , though , ellipses , Living in the Mac version somewhere .
@ 1. finally decided upon as will be cycled through .
@ 2. few files on or if it 'll look acceptable surface plotting package .

Bigram model 2:
0. Thanks for image processing systems that I shall do .
@ 1. Then , stereo graphics mirrors ftp.informatik.uni oldenburg.de pub gnuplot without introducing visible spectrum .
@ 2. If you might have had .

religion:
Bigram model 1:
@ 0. This " What does not all they have someone , Koresh and can buy land in my trunk that if he is thinking that there are no need to the Law or dead .
@ 1. agree on my personal Lord 's " By the term .
@ 2. find , reconciling the Messiah , the second century through religion , thy lord and he has been through Christ proven illegal weapons is translated perhaps the author of a sin .

Bigram model 2:
0. Such can die even agree .

1. William H .
2. second “ Jesus ’ religions !

motorcycles:
Bigram model 1:
0. your local white Honda does n’t know that close to DC if you twist of any ideas , but they attended , Walnut .
@ 1. support I would put the Hurt study .
# 2. dog chasing me the ride safe as a problem !

Bigram model 2:
0. GS1100E .
1. You got into the insults you like stroking and they stereotype me a red light special 20 or entertainment .
2. No Bras M’Lud .

medicine:
Bigram model 1:
@ 0. with insect repellent 17 110 15 and quixotic endeavor .
@ 1. per 100,000 Americans are carried out of HIV , blubird penguin.equinox.gen.nz , but it while back , again items : If I tend to keep away at a little effect .
# 2. psychiatrist claims related issues .

Bigram model 2:
0. ; Fleming ’s circumference was told is funny business .
1. I have been shown that I am still have results as a fool .
2. Chromium is actually ill students .

atheism:
Bigram model 1:
@ 0. extra death penalty , “ Freemasonry is of women in it seems to be ready , you have some good enough .
# 1. else .
@ 2. fiction , that be the FAQ that the needs and what little to do with no such as I could handle the error free , Simon Simon has no value , and that the marv’lous peace nik if you ca n’t judge other means ?

Bigram model 2:
0. And one .
1. This pursuit of fulfilling a little about them .
2. ’ I ’ve read Italian , ’ a good science .

space:
Bigram model 1:
@ 0. JSC or three or even want everyone for satellites of the moon ’s roughly a little lengthy .
# 1. run .
@ 2. apples and it is here ’s top “ The Orion concept .

Bigram model 2:

0. Diaspar is only indicative of our tents , it is designed to get the reaction controlled by Air Force Calculation Algorithm " record .
@ 1. Each letter corresponds to see the whole continent , though I 'm just to believe that would be more expensive and systems do that anything like they do not sci.space feed plumbing : and could double magnum of the station presented .
# 2. 3 5 11 , is " Astronomical Algorithms available space .

Sentence Analysis

	Total count	bigram model 1	bigram model 2
not meaningful "&"	4	2	2
a little meaningful "@"	18	14	4
some meaningful "#"	20	5	15

The above table shows the classification of each sentence based on our understanding. Since both models use bigram models, they all seem have some meanings. However, it is clear that the second bigram(15 #) outperforms the first model(5 #). Because of our beginning marker, the sentences from second model usually start with words that have the first letter capitalized, which is normally accepted in writing. Once the first word is meaningful, the sentence gets meaningful. To summarize, the second bigram model is better and we are going to use this bigram model for further analysis.

Comparisons between Unigram and Bigram models

From the sentence examples, it is obvious that bigram model outperforms unigram model, where bigram model has 15 some meaningful sentences and unigram has none. The reason is that bigram model captures connections between a pair of words and it generates sentences occur in real text.

Contribution

Ransen Niu: preprocessing, second bigram model, report.

Weicheng Yu: preprocessing, unigram model, first bigram model.