Kevin King
Professor Rolando Coto-Solano
COSC 72: Accelerated Computational Linguistics
May 2, 2023

Exercise 4.1A: Text Generation Using N-Grams

Program Output:

Unigram Sequence:
by to are vnder , 'd all for for of montague to i i too precisely so ; my in , i rosencrantz severally one bolingbroke are mrs. april disturbed o , device own off than ! am made it in king silken farewell child comes more ' a , for will it nature of plummet to what that , you and richmond , ? . my ; will nobleman men , [. rich from and all with that it . the his will better nose not this all do him these my not thee thine . hast be
Bigram Sequence:
slender accident , and not ; which wear hair at 'em be a liberal rewarder of saucy friar lodowick , they shall we see the devil ! we go home to leave you be not ours of that is a wicked , and such reasons .] saturninus . no such inevitable prosecution of flesh—you have employ a bargain . king of heaven , i pleas 'd against the wit going back , and life-preserving rest ? soothsayer . perge , which rather give me , to be not . he too far as levels with your pardon me so
Trigram Sequence:
not ; they have privilege to live , shall we dine . this business soundly . duke . my masters , for here comes the better at proverbs by how much i am very well , bully doctor ! shallow . i go ; i can not fight ; the duke he shall feel , to smile again ; for whose sake did i ne'er endured . cerimon . madam , the king , unto the worms were hallow 'd that ; and easy it is not thy kindness last longer telling than thy master here i am not gamesome
Four-gram Sequence:
doth run his course . if heaven do dwell . exeunt clown , who commands them , for a search , seek , but now it is not . portia . is there , diomed . call him hither . re-enter troilus . what , out of his thoughts , wherein i see on thee , prithee , pretty youth , and courtezan say now , sir , stands in record , and , that all , that comes a-wooing , _priami_ , is done ; and let poor volke pass . [within] who 's here ! let

== Counts ==
Unigram Count ('my'): 12618
Bigram Count ('my good'): 228
Trigram Count ('my good lord'): 77

== Probabilities ==
Unigram Prob ('my'): 0.010839898868846502

Bigram Prob ('my good'): 0.01806942463147884
Trigram Prob ('my good lord'): 0.5614035087719298

The four-gram model seems to perform better than the initial three as it considers more context and produces a more articulate output. What makes the outputs “bad” is the lack of grammatical correctness, logical coherence, and overall meaning, and the models fail to capture the complex relationships between words and lack natural sentence structure.

Exercise 4.1B: Perplexity

Program Output:

MLE Estimates: [(('is', ('this',)), 0.07187454624655147), (('a', ('is',)), 0.0656846396146007), (('dreadful', ('a',)), 0.0003203895937459951), (('sentence', ('dreadful',)), 0.01639344262295082)]

MLE Estimates: [(('but', ('put',)), 0.005802707930367505), (('a', ('but',)), 0.03361085414739439), (('losing', ('a',)), 0.00012815583749839805), (('office', ('losing',)), 0.05)]

MLE Estimates: [(('loving', ('love',)), 0.00046490004649000463), (('not', ('loving',)), 0.008695652173913044), (('dogs', ('not',)), 0.0)]

PP(this is a dreadful sentence):79.68987102980361

PP(put but a losing office):168.18812749362763

PP(love loving not dogs):inf

The sentence with the lowest perplexity is "this is a dreadful sentence" with a perplexity value of 79.68987102980361, which means that the model has a relatively better ability to predict this sentence based on the given training data.

On the other hand, the sentence "put but a losing office" has the highest perplexity value of 168.18812749362763. This indicates that the model struggles to predict this sentence accurately, and it has higher uncertainty or confusion when encountering this sequence of words.

The sentence "love loving not dogs" has a perplexity value of infinity, suggesting that the model has not encountered this sequence of words during training and is unable to assign a probability to it. As a result, the perplexity value becomes infinite.

Overall, a lower perplexity translates to better model performance, as it indicates a greater ability to predict the given sentence based on the trained language model.

Exercise 4.2: Spell Check

Program Output:

Sample Sentences (provided in instructions)

```
Sample Input 1: Kia orana kotoo mai i Rarotoga!
== Possible misspelling ==
 kotoo: {'kotou'}
== Possible misspelling ==
 rarotoga: {'rarotonga'}


Sample input 2: Kua aere au ki Mauke.
== Possible misspelling ==
 aere: {'qaere', 'tere', 'mere', 'rere'}
== Possible misspelling ==
 mauke: {'maquke'}
```

Prompt User for Sentence

```
Please write a sentence in Cook Islands Maori and press ENTER to check the spelling:
 Kia orana kotoo mai i Rarotoga!
== Possible misspelling ==
 kotoo: {'kotou'}
== Possible misspelling ==
 rarotoga: {'rarotonga'}
```

Exercise 4.3: Naive Bayes Classification

```
=== AMAZON ===
train on 800 instances, test on 200 instances
accuracy:      0.89
pos precision: 0.90625
pos recall:    0.87
neg precision: 0.875
neg recall:    0.91
neg F-measure: 0.892156862745098
pos F-measure: 0.8877551020408163
Most Informative Features
              ('Great',) = True              pos : neg    =     40.3 : 1.0
               ('nice',) = True              pos : neg    =     13.0 : 1.0
              ('smart',) = True              pos : neg    =     12.3 : 1.0
         ('people', ',') = True              pos : neg    =     11.7 : 1.0
              ('learn',) = True              pos : neg    =     11.0 : 1.0
      ('opportunities',) = True              pos : neg    =      9.8 : 1.0
           ('benefits',) = True              pos : neg    =      9.7 : 1.0
         ('to', 'learn') = True              pos : neg    =      9.0 : 1.0
            ('balance',) = True              neg : pos    =      8.8 : 1.0
                ('Not',) = True              neg : pos    =      7.8 : 1.0
  ('opportunity', 'for') = True              pos : neg    =      7.7 : 1.0
              ('does',) = True               neg : pos    =      7.7 : 1.0
              ('rate',) = True               neg : pos    =      7.7 : 1.0
                ('No',) = True               neg : pos    =      7.4 : 1.0
              ('Good',) = True               pos : neg    =      7.0 : 1.0
          ('You', 'get') = True              pos : neg    =      7.0 : 1.0
        ('work', 'with') = True              pos : neg    =      7.0 : 1.0
       ('long', 'hours') = True              neg : pos    =      7.0 : 1.0
           ('get', 'to') = True              pos : neg    =      7.0 : 1.0
         ('a', 'great') = True               pos : neg    =      7.0 : 1.0
               ('fun',) = True               pos : neg    =      7.0 : 1.0
     ('life', 'balance') = True              neg : pos    =      6.6 : 1.0
          ('.', 'Great') = True              pos : neg    =      6.3 : 1.0
              ('times',) = True              neg : pos    =      6.3 : 1.0
             ('decent',) = True              pos : neg    =      6.3 : 1.0
```

Amazon (above):

- Positive aspects about working at Amazon: The most informative features that indicate a positive sentiment about working at Amazon include "Great," "nice," "smart," "people," "learn," "opportunities," and "benefits." These words suggest that employees appreciate the company culture, opportunities for personal growth, and the positive environment.
- Negative aspects about working at Amazon: The most informative features that indicate a negative sentiment about working at Amazon include "balance," "Not," "does," "rate," "No," "long hours," and "life balance." These words suggest concerns about work-life balance, demanding hours, and dissatisfaction with certain aspects of the job.

```
=== GOOGLE ===
train on 800 instances, test on 200 instances
accuracy:      0.885
pos precision: 0.9230769230769231
pos recall:    0.84
neg precision: 0.8532110091743119
neg recall:    0.93
neg F-measure: 0.8899521531100479
pos F-measure: 0.8795811518324608
Most Informative Features
              ('Great',) = True             pos : neg    =     29.8 : 1.0
              ('perks',) = True             pos : neg    =     25.4 : 1.0
               ('free',) = True             pos : neg    =     21.0 : 1.0
            ('amazing',) = True             pos : neg    =     17.7 : 1.0
          ('hard', 'to') = True             neg : pos    =     15.7 : 1.0
               ('Good',) = True             pos : neg    =     15.0 : 1.0
           ('can', 'be') = True             neg : pos    =     14.2 : 1.0
          ('sometimes',) = True             neg : pos    =     13.7 : 1.0
        ('interesting',) = True             pos : neg    =     13.0 : 1.0
           ('food', ',') = True             pos : neg    =     12.6 : 1.0
          ('difficult',) = True             neg : pos    =     12.3 : 1.0
              ('times',) = True             neg : pos    =     12.3 : 1.0
               ('fun',) = True              pos : neg    =     10.2 : 1.0
            ('politics',) = True            neg : pos    =      9.7 : 1.0
            ('benefits',) = True            pos : neg    =      9.5 : 1.0
        ('culture', ',') = True             pos : neg    =      9.0 : 1.0
             ('awesome',) = True            pos : neg    =      9.0 : 1.0
     ('and', 'benefits') = True             pos : neg    =      8.3 : 1.0
            ('nothing',) = True             neg : pos    =      8.3 : 1.0
    ('environment', ',') = True             pos : neg    =      8.3 : 1.0
              ('food',) = True              pos : neg    =      8.2 : 1.0
              ('Free',) = True              pos : neg    =      7.8 : 1.0
               ('not',) = True              neg : pos    =      7.8 : 1.0
       ('organization',) = True             neg : pos    =      7.7 : 1.0
             ("'s", 'a') = True             neg : pos    =      7.7 : 1.0
```

Google (above):

- Positive aspects about working at Google: The most informative features that indicate a positive sentiment about working at Google include "Great," "perks," "free," "amazing," "Good," "interesting," "food," "fun," "benefits," and "awesome." These words imply that employees appreciate the great benefits, perks, and interesting work culture at Google.
- Negative aspects about working at Google: The most informative features that indicate a negative sentiment about working at Google include "hard to," "sometimes," "difficult," "times," "politics," "nothing," and "not." These words suggest challenges related to the intensity or difficulty of work, office politics, and certain negative aspects that some employees may experience.