HNRS 3035 / CSC 4700 - Homework #1 Answers

Jonathan Morse

Question 1: Train a Unigram Model

```
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 % python Morse_hnrs3035_cshw1.py
train_ngram --data moby.txt --save unigram_model.pkl --n 1

N-gram model trained and saved to unigram_model.pkl
```

Question 1a: Unigram Prediction (whale, random)

```
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 % python Morse_hnrs3035_cshw1.py
predict_ngram --word whale --nwords 100 --load unigram_model.pkl

must exhale; worse than billiard-tables, and by the life in one another’s
eyes. “Holloa!” he made up sooner trimmed the thick branches of dissimilar
in the hunters. His face or Ramadan, or the spinal theory for a mere vapor
they call the awakened much he seemed theirs; they were almost approaching
to something went the great mundane grandeur. A spy-glass in the half
distrustful of all the magical, sometimes be undecided as well hoisted,
sheeted home the boats! Luff!” cried Bunger, stoopingly walking up the
fastening to make it would turn up to be not what he would try a
```

Question 1b: Unigram Prediction (whale, deterministic)

```
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 % python Morse_hnrs3035_cshw1.py
predict_ngram --word whale --nwords 100 --load unigram_model.pkl --d

is the whale is the whale is the whale is the whale is the whale is the
whale is the whale is the whale is the whale is the whale is the whale is
the whale is the whale is the whale is the whale is the whale is the whale
is the whale is the whale is the whale is the whale is the whale is the
whale is the whale is the whale is the whale is the whale is the whale is
the whale is the whale is the whale is the whale is the whale is
```

Question 1c: Unigram Prediction (tomato, error)

```
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 % python Morse_hnrs3035_cshw1.py
predict_ngram --word tomato --nwords 100 --load unigram_model.pkl

Error: Input word(s) not found in vocabulary. Error: Input word(s) not
found in vocabulary.
```

Question 2: Train a Bigram Model

```
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 % python Morse_hnrs3035_cshw1.py
train_ngram --data moby.txt --save bigram_model.pkl --n 2

N-gram model trained and saved to bigram_model.pkl
```

Question 2a: Bigram Prediction (the harpooneer, deterministic)

```
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 % python Morse_hnrs3035_cshw1.py
predict_ngram --word "the harpooneer" --nwords 100 --load bigram_model.pkl
--d

might be, and that the whale is a thing not to be the first place, you
will be all this as it were, to the deck, and in the same time that the
whale is a thing not to be the first place, you will be all this as it
were, to the deck, and in the same time that the whale is a thing not to
be the first place, you will be all this as it were, to the deck, and in
the same time that the whale is a thing not to be the first place, you
```

Question 3: Train a BPE Tokenizer (k=500)

```
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 % python Morse_hnrs3035_cshw1.py
train_bpe --data moby.txt --save bpe_model_500.pkl

BPE tokenizer trained and saved to bpe_model_500.pkl
```

Question 3a: BPE Tokenization (k=500)

```
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 % python Morse_hnrs3035_cshw1.py
tokenize --text "The bright red tomato was eaten by the whale\!" --load
bpe_model_500.pkl

Tokens: ['T', 'h', 'e', ' ', 'b', 'r', 'i', 'g', 'h', 't', ' ', 'r', 'e',
'd', ' ', 't', 'o', 'm', 'a', 't', 'o', ' ', 'w', 'a', 's', ' ', 'e', 'a',
't', 'e', 'n', ' ', 'b', 'y', ' ', 'th', 'e', ' ', 'w', 'h', 'a', 'l',
'e', '!']
Token IDs: [21, 82, 283, 564, 680, 323, 9, 25, 82, 15, 564, 323, 283, 944,
564, 15, 37, 319, 784, 15, 37, 564, 650, 784, 787, 564, 283, 784, 15, 283,
885, 564, 680, 997, 564, 59, 283, 564, 650, 82, 784, 500, 283, 256]
```

Question 4: Train a BPE Tokenizer (k=3000)

```
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 % python Morse_hnrs3035_cshw1.py
train_bpe --data moby.txt --save bpe_model_3000.pkl

BPE tokenizer trained and saved to bpe_model_3000.pkl
```

Question 4a: BPE Tokenization (k=3000)

```
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 % python Morse_hnrs3035_cshw1.py
tokenize --text "The bright red tomato was eaten by the whale\!" --load
bpe_model_3000.pkl

Tokens: ['T', 'h', 'e', ' ', 'b', 'r', 'i', 'g', 'h', 't', ' ', 'r', 'e',
'd', ' ', 't', 'o', 'm', 'a', 't', 'o', ' ', 'w', 'a', 's', ' ', 'e', 'a',
't', 'e', 'n', ' ', 'b', 'y', ' ', 'th', 'e', ' ', 'w', 'h', 'a', 'l',
'e', '!']
Token IDs: [807, 638, 320, 813, 41, 112, 277, 31, 638, 775, 813, 112, 320,
192, 813, 775, 225, 319, 648, 775, 225, 813, 318, 648, 270, 813, 320, 648,
775, 320, 164, 813, 41, 907, 813, 865, 320, 813, 318, 638, 648, 489, 320,
469]
```

Fin

```
total 17560
-rw-r--r--  1 jomo  staff   5.7K Feb  5 00:21 Morse_hnrs3035_cshw1.py
-rw-r--r--  1 jomo  staff   5.2M Feb  5 00:36 bigram_model.pkl
-rw-r--r--  1 jomo  staff   507B Feb  5 00:37 bpe_model_3000.pkl
-rw-r--r--  1 jomo  staff   507B Feb  5 00:36 bpe_model_500.pkl
-rw-r--r--@ 1 jomo  staff   1.2M Feb  5 00:22 moby.txt
-rw-r--r--  1 jomo  staff   2.1M Feb  5 00:34 unigram_model.pkl
(hnrs3035) jomo@MacBook-Pro-98 hnrs3035 %
```