CS539_1_NLP
kimdongk@oregonstate.edu
EX2

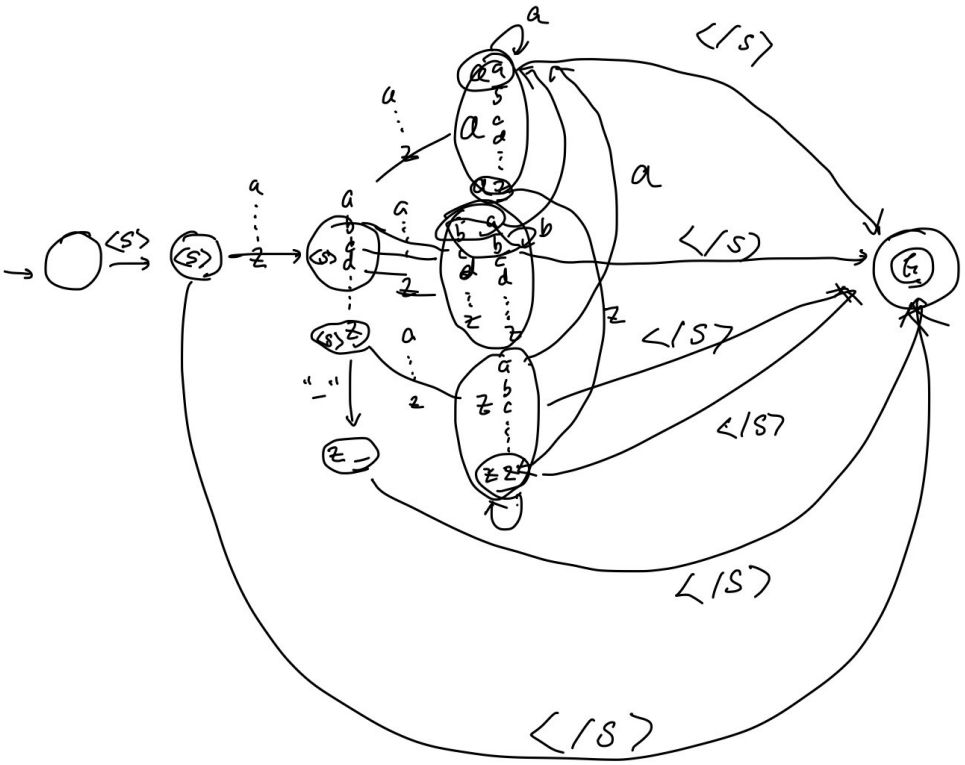
Part 1.

2 & 3.

	Unigram	Bigram	Bigram with smoothing	Trigram	Trigam with smoothing
Corpus probability	2^-39009.6	2^-29843.2	2^-32549.9	2^-5255.71	2^-27939.1
Entropy	4.11494	3.14802	3.43354	0.5544	2.94716
States	3	30	30	497	759
Transitions	29	540	811	3894	21925

4. at the initialization stage,
   1) making lists with a ~ z and <s>, </s>, _
   2) mingle them for bigram and trigram
   3) trying different smoothing possibilities to find a lower entropy values
   4) I set smoothing=0.5

5. drawing a graph



Part 2.

1. the results are in 2_1_results folder in .txt format. I cannot find any meaning in those results.

2.

```
cat test.txt | sed -e 's/[aeiou]//g' > test.txt.novowels

cat test.txt.novowels | sed -e 's/ /_/g;s/\(.\)/\1 /g' | awk '{printf("<s> %s </s>\n", $0)}' | ./carmel -sribI
trigram.wfsa.norm remove-vowels.fst > test.txt.vowel-restored.tri

cat test.txt | sed -e 's/ /_/g;s/\(.\)/\1 /g' | awk '{printf("<s> %s </s>\n", $0)}' > test_new.txt
```

eval.py from hw1

```
python eval.py test_new.txt test.txt.vowel-restored.uni
→ accuracy results
uni: 0.008135168961201502
bi: 0.16270337922403003
tri: 0.42490613266583227
```

eval2.py for ex2

```
python eval2.py test_new.txt test.txt.vowel-restored.tri

uni: recall= 0.688 precision= 0.970 F1= 0.805
bi  : recall= 0.836 precision= 0.971 F1= 0.898
tri : recall= 0.923 precision= 0.959 F1= 0.941
```

3.

```
cat test.txt | sed -e 's/ //g' > test.txt.nospaces

cat test.txt.nospaces | sed -e 's/ /_/g;s/\(.\)/\1 /g' | awk '{printf("<s> %s </s>\n", $0)}' | ./carmel -sribI
trigram_smooth.wfsa.norm remove-spaces.fst > test.txt.space_restored.tri

python eval2.py test_new.txt test.txt.space_restored.uni

uni: recall= 0.842 precision= 0.976 F1= 0.904
bi  : recall= 0.950 precision= 0.978 F1= 0.964
tri : recall= 0.992 precision= 0.973 F1= 0.983

echo "therestcanbeatotalmessandyoucanstillreaditwithoutaproblem" | sed -e 's/ /_/g;s/\(.\)/\1 /g' | awk
'{printf("<s> %s </s>\n", $0)}' | ./carmel -sribI trigram_smooth.wfsa.norm remove-spaces.fst
Input line 1: <s> t h e r e s t c a n b e a t o t a l m e s s a n d y o u c a n s t i l l r e a d i t w i t h o u t a p r
o b l e m  </s>
        (121 states / 240 arcs)
        (234 states / 518 arcs)
<s> t h e _ r e s t _ c a n _ b e a t o _ t a l m e s s _ a n d _ y o u _ c a n s t i l l _ r e a d i t _ w i t h o u t
_ a _ p r o b l e m </s> *e* e^-131.802862957238
Derivations found for all 1 inputs
Viterbi (best path) product of probs=e^-131.802862957238, probability=2^-190.151 per-input-symbol-
perplexity(N=59)=2^3.2229 per-line-perplexity(N=1)=2^190.151
```

```
 echo "thisisbecausethehumanminddoesnotreadeveryletterbyitselfbutthewordasawhole" | sed -e 's/
/_/g;s/\(.\)/\1 /g' | awk '{printf("<s> %s </s>\n", $0)}' | ./carmel -sribI trigram_smooth.wfsa.norm
remove-spaces.fst
Input line 1: <s> t h i s i s b e c a u s e t h e h u m a n m i n d d o e s n o t r e a d e v e r y l e t t e r b y i t
s e l f b u t t h e w o r d a s a w h o l e  </s>
        (153 states / 304 arcs)
        (298 states / 646 arcs)
<s> t h i s _ i s _ b e c a u s e _ t h e _ h u m a n _ m i n d _ d o e s _ n o t _ r e a d e v e r y _ l e t t e r _
b y _ i t _ s e l f b u t _ t h e _ w o r d _ a s _ a _ w h o l e </s> *e* e^-171.248962768691
Derivations found for all 1 inputs
Viterbi (best path) product of probs=e^-171.248962768691, probability=2^-247.06 per-input-symbol-
perplexity(N=75)=2^3.29413 per-line-perplexity(N=1)=2^247.06
```

4.

To decide the more efficient method among two can be considered by two things: difficulty and accuracy. In terms of difficulty, for the restoring spaces, remove-spaces.fst needs to be written, which is easier than writing remove-vowels.fst. Also, intuitively, we can guess the meaning of a string without spaces easier than a string without vowels. With respect to accuracy, I checked recall, precision, and f1 score on both methods. As the results, restoring spaces has slightly higher f1 score. Therefore, I think the second method, restoring spaces, is a more efficient solution for this problem.