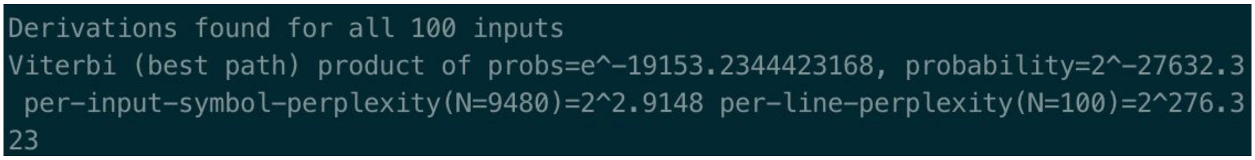
Natural Language Processing HW 6

Jing Huang   933039277

1. Playing with the NLM (0 pts)

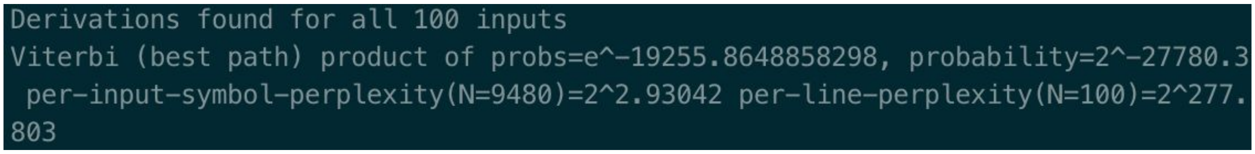
2. Evaluating Entropy (2 pts)

1. base:



```
Derivations found for all 100 inputs
Viterbi (best path) product of probs=e^-19153.2344423168, probability=2^-27632.3
 per-input-symbol-perplexity(N=9480)=2^2.9148 per-line-perplexity(N=100)=2^276.3
23
```

large:



```
Derivations found for all 100 inputs
Viterbi (best path) product of probs=e^-19255.8648858298, probability=2^-27780.3
 per-input-symbol-perplexity(N=9480)=2^2.93042 per-line-perplexity(N=100)=2^277.
803
```

No, because larger dataset will create more path and have higher cost.

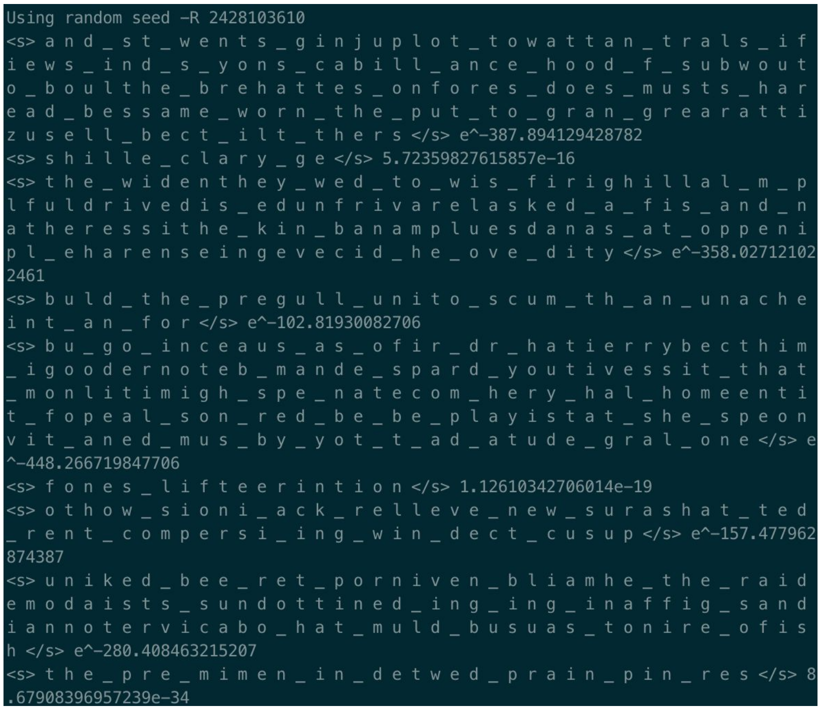
2. Base: 2.6928136114928667
Large: 1.9726965712505398

Large is better. Because it's accuracy increases while it has larger dataset.

3. Yes. It makes sense because NLM has more grams than trigram.

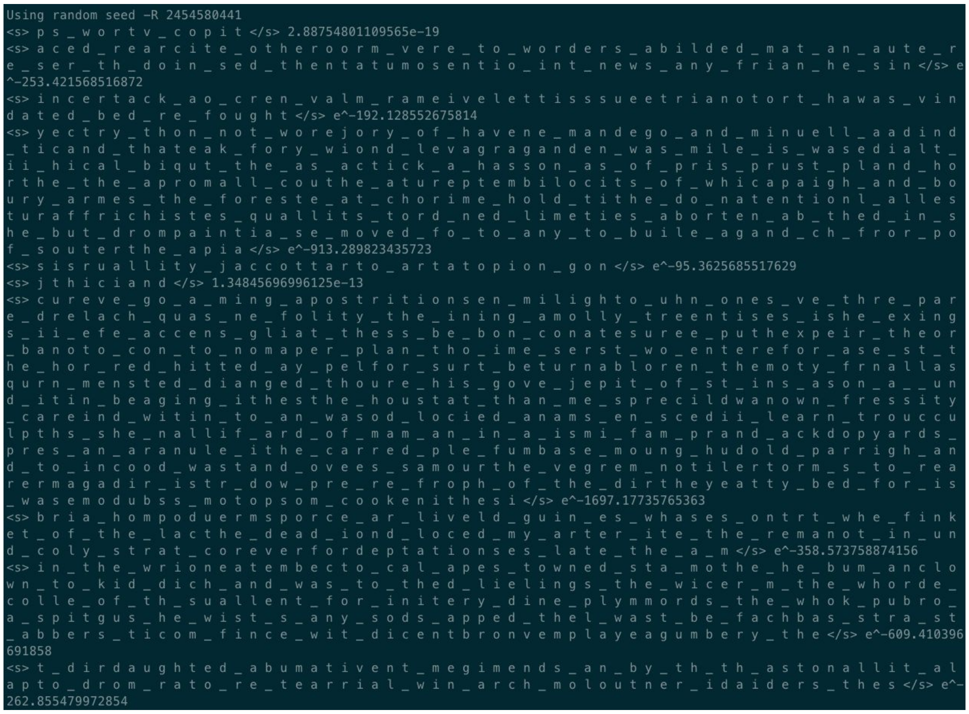
3. Random Generation (2 pts)

1. Base:



```
Using random seed -R 2428103610
<s> a n d _ s t _ w e n t s _ g i n j u p l o t _ t o w a t t a n _ t r a l s _ i f
i e w s _ i n d _ s _ y o n s _ c a b i l l _ a n c e _ h o o d _ f _ s u b w o u t
o _ b o u l t h e _ b r e h a t t e s _ o n f o r e s _ d o e s _ m u s t s _ h a r
e a d _ b e s s a m e _ w o r n _ t h e _ p u t _ t o _ g r a n _ g r e a r a t t i
z u s e l l _ b e c t _ i l t _ t h e r s </s> e^-387.894129428782
<s> s h i l l e _ c l a r y _ g e </s> 5.72359827615857e-16
<s> t h e _ w i d e n t h e y _ w e d _ t o _ w i s _ f i r i g h i l l a l _ m _ p
l f u l d r i v e d i s _ e d u n f r i v a r e l a s k e d _ a _ f i s _ a n d _ n
a t h e r e s s i t h e _ k i n _ b a n a m p l u e s d a n a s _ a t _ o p p e n i
p l _ e h a r e n s e i n g e v e c i d _ h e _ o v e _ d i t y </s> e^-358.02712102
2461
<s> b u l d _ t h e _ p r e g u l l _ u n i t o _ s c u m _ t h _ a n _ u n a c h e
i n t _ a n _ f o r </s> e^-102.81930082706
<s> b u _ g o _ i n c e a u s _ a s _ o f i r _ d r _ h a t i e r r y b e c t h i m
_ i g o o d e r n o t e b _ m a n d e _ s p a r d _ y o u t i v e s s i t _ t h a t
_ m o n l i t i m i g h _ s p e _ n a t e c o m _ h e r y _ h a l _ h o m e e n t i
t _ f o p e a l _ s o n _ r e d _ b e _ b e _ p l a y i s t a t _ s h e _ s p e o n
v i t _ a n e d _ m u s _ b y _ y o t _ t _ a d _ a t u d e _ g r a l _ o n e </s> e
^-448.266719847706
<s> f o n e s _ l i f t e e r i n t i o n </s> 1.12610342706014e-19
<s> o t h o w _ s i o n i _ a c k _ r e l l e v e _ n e w _ s u r a s h a t _ t e d
_ r e n t _ c o m p e r s i _ i n g _ w i n _ d e c t _ c u s u p </s> e^-157.477962
874387
<s> u n i k e d _ b e e _ r e t _ p o r n i v e n _ b l i a m h e _ t h e _ r a i d
e m o d a i s t s _ s u n d o t t i n e d _ i n g _ i n g _ i n a f f i g _ s a n d
i a n n o t e r v i c a b o _ h a t _ m u l d _ b u s u a s _ t o n i r e _ o f i s
h </s> e^-280.408463215207
<s> t h e _ p r e _ m i m e n _ i n _ d e t w e d _ p r a i n _ p i n _ r e s </s> 8
.67908396957239e-34
```

Large:



```
Using random seed -R 2454580441
<s> p s _ w o r t v _ c o p i t </s> 2.88754801109565e-19
<s> a c e d _ r e a r c i t e _ o t h e r o o r m _ v e r e _ t o _ w o r d e r s _ a b i l d e d _ m a t _ a n _ a u t e _ r
e _ s e r _ t h _ d o i n _ s e d _ t h e n t a t u m o s e n t i o _ i n t _ n e w s _ a n y _ f r i a n _ h e _ s i n </s> e
^-253.421568516872
<s> i n c e r t a c k _ a o _ c r e n _ v a l m _ r a m e i v e l e t t i s s s u e e t r i a n o t o r t _ h a w a s _ v i n
d a t e d _ b e d _ r e _ f o u g h t </s> e^-192.128552675814
<s> y e c t r y _ t h o n _ n o t _ w o r e j o r y _ o f _ h a v e n e _ m a n d e g o _ a n d _ m i n u e l l _ a a d i n d
_ t i c a n d _ t h a t e a k _ f o r y _ w i o n d _ l e v a g r a g a n d e n _ w a s _ m i l e _ i s _ w a s e d i a l t _
i i _ h i c a l _ b i q u t _ t h e _ a s _ a c t i c k _ a _ h a s s o n _ a s _ o f _ p r i s _ p r u s t _ p l a n d _ h o
r t h e _ t h e _ a p r o m a l l _ c o u t h e _ a t u r e p t e m b i l o c i t s _ o f _ w h i c a p a i g h _ a n d _ b o
u r y _ a r m e s _ t h e _ f o r e s t e _ a t _ c h o r i m e _ h o l d _ t i t h e _ d o _ n a t e n t i o n l _ a l l e s
t u r a f f r i c h i s t e s _ q u a l l i t s _ t o r d _ n e d _ l i m e t i e s _ a b o r t e n _ a b _ t h e d _ i n _ s
h e _ b u t _ d r o m p a i n t i a _ s e _ m o v e d _ f o _ t o _ a n y _ t o _ b u i l e _ a g a n d _ c h _ f r o r _ p o
f _ s o u t e r t h e _ a p i a </s> e^-913.289823435723
<s> s i s r u a l l i t y _ j a c c o t t a r t o _ a r t a t o p i o n _ g o n </s> e^-95.3625685517629
<s> j t h i c i a n d </s> 1.34845696996125e-13
<s> c u r e v e _ g o _ a _ m i n g _ a p o s t r i t i o n s e n _ m i l i g h t o _ u h n _ o n e s _ v e _ t h r e _ p a r
e _ d r e l a c h _ q u a s _ n e _ f o l i t y _ t h e _ i n i n g _ a m o l l y _ t r e e n t i s e s _ i s h e _ e x i n g
s _ i i _ e f e _ a c c e n s _ g l i a t _ t h e s s _ b e _ b o n _ c o n a t e s u r e e _ p u t h e x p e i r _ t h e o r
_ b a n o t o _ c o n _ t o _ n o m a p e r _ p l a n _ t h o _ i m e _ s e r s t _ w o _ e n t e r e f o r _ a s e _ s t _ t
h e _ h o r _ r e d _ h i t t e d _ a y _ p e l f o r _ s u r t _ b e t u r n a b l o r e n _ t h e m o t y _ f r n a l l a s
q u r n _ m e n s t e d _ d i a n g e d _ t h o u r e _ h i s _ g o v e _ j e p i t _ o f _ s t _ i n s _ a s o n _ a _ _ u n
d _ i t i n _ b e a g i n g _ i t h e s t h e _ h o u s t a t _ t h a n _ m e _ s p r e c i l d w a n o w n _ f r e s s i t y
_ c a r e i n d _ w i t i n _ t o _ a n _ w a s o d _ l o c i e d _ a n a m s _ e n _ s c e d i i _ l e a r n _ t r o u c c u
l p t h s _ s h e _ n a l l i f _ a r d _ o f _ m a m _ a n _ i n _ a _ i s m i _ f a m _ p r a n d _ a c k d o p y a r d s _
p r e s _ a n _ a r a n u l e _ i t h e _ c a r r e d _ p l e _ f u m b a s e _ m o u n g _ h u d o l d _ p a r r i g h _ a n
d _ t o _ i n c o o d _ w a s t a n d _ o v e e s _ s a m o u r t h e _ v e g r e m _ n o t i l e r t o r m _ s _ t o _ r e a
r e r m a g a d i r _ i s t r _ d o w _ p r e _ r e _ f r o p h _ o f _ t h e _ d i r t h e y e a t t y _ b e d _ f o r _ i s
_ w a s e m o d u b s s _ m o t o p s o m _ c o o k e n i t h e s i </s> e^-1697.17735765363
<s> b r i a _ h o m p o d u e r m s p o r c e _ a r _ l i v e l d _ g u i n _ e s _ w h a s e s _ o n t r t _ w h e _ f i n k
e t _ o f _ t h e _ l a c t h e _ d e a d _ i o n d _ l o c e d _ m y _ a r t e r _ i t e _ t h e _ r e m a n o t _ i n _ u n
d _ c o l y _ s t r a t _ c o r e v e r f o r d e p t a t i o n s e s _ l a t e _ t h e _ a _ m </s> e^-358.573758874156
<s> i n _ t h e _ w r i o n e a t e m b e c t o _ c a l _ a p e s _ t o w n e d _ s t a _ m o t h e _ h e _ b u m _ a n c l o
w n _ t o _ k i d _ d i c h _ a n d _ w a s _ t o _ t h e d _ l i e l i n g s _ t h e _ w i c e r _ m _ t h e _ w h o r d e _
c o l l e _ o f _ t h _ s u a l l e n t _ f o r _ i n i t e r y _ d i n e _ p l y m m o r d s _ t h e _ w h o k _ p u b r o _
a _ s p i t g u s _ h e _ w i s t _ s _ a n y _ s o d s _ a p p e d _ t h e l _ w a s t _ b e _ f a c h b a s _ s t r a _ s t
_ a b b e r s _ t i c o m _ f i n c e _ w i t _ d i c e n t b r o n v e m p l a y e a g u m b e r y _ t h e </s> e^-609.410396
691858
<s> t _ d i r d a u g h t e d _ a b u m a t i v e n t _ m e g i m e n d s _ a n _ b y _ t h _ t h _ a s t o n a l l i t _ a l
a p t o _ d r o m _ r a t o _ r e _ t e a r r i a l _ w i n _ a r c h _ m o l o u t n e r _ i d a i d e r s _ t h e s </s> e^-
262.855479972854
```

Almost don’t make sense.

2.

_and_the_plan_to_the_and_simper_have_who_on_change_fate_a_gow_the_strest_who_dome_for_the_all_not_many_and_a_seconitions_and_sund_be_ounders_last_start_the_real_childed_the_new_in_has_sucler_of_have_seend_the_regrames_in_the_sparies_are_only_and_the_way_the_whing_the_scarients_and_modies_of_the_complice</s>

ses_and_propers_and_the_and_those_maker_stast_sotse_at_the_prostenting_shained_and_be_make_the_moderned_the_was_sone_with_the_is_and_the_meitting_the_the_stardien_that_you_reis_and_but_comes</s>

_the_for_the_new_for_the_lead_the_oned_her_her_pay_sens_they_politica_a_bely_are_is_lay_is_to_the_will_and_proven_the_ond_her_to_monen_the_been_the_not_propten_to_it_the_steurges_and_be_worke_and_says_stame_of_the_will_and_the_more_wish_the_was_the_monte_as_anter_make_for_the_seppor_of_the_pecosed_the_wererate_to_the_stres_and_of_the_the_for_the_the_unite_and_stasmens_to_was_the_compled_to_wis_the_at_the_read_to_game_and_contress_that_suptorion_we_way_withing_and_to_be_for_the_intort_and_and_and_compled_the_and_and_part_the_were_of_the_speund_in_the_and_make_a_could_for_the_mollected_is_a_stay_and_sund_the_not_on_have_with_the_be_has_mection_that_the_onder_and_the_been_in_the_proscarions_and_bull_to_be_complication_and_and_when_new_has_proses_who_allesions_of_the_bit_and_whe_the_promestions_in_the_confricic_some_of_cheres_of_play_new_so_seate_al_of_the_way_to_beging_marke_this_will_and_doat_a_comenite_content_and_was_seep_not_her_the_onder_a_compept_the_play_make_whinged_and_home_of_in_the_to_being_the_recan_of_the_meathing_the_distarions_manes_the_names_to_a_the_some_and_loak_and_fine_a_and_be_the_new_lise_of_it_and_comples</s>

_have_a_driter_the_will_deforth_to_dementes_to_parting_in_the_ring_of_the_cays_of_the_been_the_and_her_stone_the_chaning_in_uster_can_of_the_promits_to_were_the_recined_the_from_the_sche_work_a_contrets_the_cour_had_componed_the_intsuping_and_the_asting_fins_the_the_dispostan_and_ho_and_recorts_and_comen_to_expervion_as_of_the_stres_and_a_beed_all_and_the_not_the_world_and_and_be_rester_the_ond_the_many_the_patarient_the_strine_would_sitch_and_wh en_the_the_reaw_the_provesting_shaated_and_stinch_with_the_uristing_the_stamed_the_frent_in_the_strike_simply_to_wate_and_and_and_the_beants_of_the_conlect_a_contray_and_and_service_and_her_on_the_conterned_state_thoun_also_allice_to_the_for_the_regions_and_the_like_the_reloper_to_esite_to_send_is_the_dird_in_the_end_of_the_right_in_the_way_the_und_make_must_first_the_to_wate_expellents_protered_the_from_a_comitions_he_presind_can_a_be_to_with_fell_the_with_a_las_is_have_the_underside_with_the_parting_a_trocent_on_the_is_trest_as_the_day_and_in_the_vided_the_contrate_the_proteres_her_for_the_to_years_of_the_practer_the_reses_a_seponicaus_and_the_the_oner_is_was_to_a_stay_helled_the_may_and_when_the_us_fame_for_shout_the_proges_to_players_with_the_sucter_the_ond_a_spen_would_propers_and_the_beer_to_one_for_the_iston_with_the_and_been_was_she_stocher_distames_the_intertioning_lest_when_neame</s>

sing_in_the_say_was_start_the_maran_choites_of_the_draw_can_in_usesting_the_asting_and_be_would_rebict_and_the_malinged_of_the_world_war_and_the_complethre_has_looked_when_when_the_rea_the_wall_in_care_to_you_houshing_some_a_for_the_rege_preser_chald</s>

_a_chained_and_the_complies</s>

_the_was_the_reast_as_this_for_the_prester_his_never_the_indersite_is_dours_of_bore_in_dort_and_entered_erectes_comet_and_a_computer_the_store_the_dast_your_see_can_with_her_makes_are_the_contres_able_unitient_on_the_shout_the_distrate_the_stargen_what_it_to_a_couter_the_promossed_by_month_a_the_recontented_of_the_wert_in_the_propes_to_a_suppose_the_reably_start_helbonger</s>

_can_to_completic_she_way_the_and_challenger_content_and_hote_a_supperenty_in_the_pulled_for_the_back_the_betion_work_a_chood_that_send_to_get_challal_the_expenty_sotner_seed_thounous_astants_the_the_school_the_all_the_betioners_the_been_the_becens_to_get_to_can_when_i_way_the_strow_the_soge_a_challed_the_more_send_is_a_for_fall_of_stan_her_and_stand_things_work_and_make_not_and_one_part_and_complation_on_the_a_recontert_stars_to_be_growers_on_seston_for_the_not_in_the_with_the_pride_of_the_from_a_some_to_the_direasing_in_the_planions_the_was_daming_to_the_m

```
a s t _ a n d _ f e r t e r s _ a n d _ t h e _ c o n t e n c e _ t h e _ i s _ h a v e _ s o n e _ f o r _ t h e
_ b e s t _ t h e _ s t e p s _ g o v e r s _ o f _ t h e _ p e r s i n d _ u p _ i n _ a _ c o m p i s a l l y _
g e a n t s _ t o _ t h e _ w i l l _ t h e _ m o s t _ c o n t r i c _ m o r e _ a n d _ h o w _ t h e _ n e
w _ c o n t a m e _ i s _ s t a r e d _ i n _ t h e _ a c e r i t a l _ c o m p e t e s _ a n d _ c o n t r o n
_ a _ s e l e d e d _ t o _ s t a r d _ a n d _ s t r a t e _ i s _ t h e _ r e a l _ a s t a n _ i n _ t h e _ p r
o m a r e s _ a t _ f o r _ t h e _ p r o m e p t e r s _ t h e _ s t a r e d _ t h e _ p r o s e s t _ b e e n _
t h e _ o r p e n d _ o f _ t h e _ m a n y _ r e p e n _ a n d _ o n e _ b e e r _ o f _ t h e _ c h i l d e r
s _ a n d _ c a n _ b e _ t h e _ g a r d _ a t _ c a r e _ s e a s t _ f o r _ t h e _ n e w _ s t e n _ t h e
_ t o _ t h e _ p r o m e t i o n s _ a r e _ a m p e a s t _ s t a s e d _ b e t o m e _ u p _ t h e _ f o l l i
n g _ i n _ t h e _ m i s s i n g _ t h e _ t o _ h e _ g e e t _ s t r a t e s _ t o _ c o m p o n i c t _ a r e
d _ y o u r _ l a r g e r _ t o _ j e r e t s _ t h e _ w u r e _ r o a l _ t o _ v e r t i n g _ a n d _ p r o p
e n c e _ t h e _ r e t e r t s _ t o _ t h e _ r e c a r e d _ w h e n _ h e r _ t h e _ p a c e d _ a s _ t h e
_ a n d _ t h e _ b a c k s _ a n d _ w h e n _ f o l l o w _ t h e _ i n _ t h e _ d e s t i o n _ t o _ w o
u l d _ b e _ a n d _ t h e _ w i t h _ t h e _ r e c o m p i n e d _ a n d _ h e _ w o r k i n g _ a _ c o
m p u t e r _ h a s _ a _ c o u l d _ p r e c a r e d _ t h e _ p a n t i n g _ t h e _ r e s e r e d _ f o r _ t
h e _ r e s t _ b e s t _ c o l l e c t _ w e _ a n d _ t h e _ b a c k _ a n d _ e x p a l l e r _ t h e _ o n d
_ a _ s t a r s _ a n d _ d i s t a r t i n g _ t o _ c o n t r i e s s _ a n d _ f e l l e r _ f i r s t _ h a s _ u
n d e r _ a _ r e a l _ t h e _ m e s t _ a n d _ g i n g _ s e e p _ f o r _ t h e _ s t r i t _ t h e _ r e a l _
f e o l l e _ c o u n t y _ t h e _ w a i r e s </s>

o n g _ i n _ t h e _ p r o m o s s i o n _ o f _ t h e _ c o n f o r t h _ a _ k e p p e r t _ a n d _ t h e _
s h o u t _ n e w _ l i m e _ t h e _ f o r t e _ t o _ t h e _ n e w _ a n d _ a n d _ t h e _ s t r a t e r s
_ a n d _ r e s t e r _ a n d _ t h e _ m u s s i n g _ v e r s i a n g _ r e s u a l i v e _ o f _ a r e _ i s _
f o r _ t h e _ p r o c e s i n _ d o n t y _ t h e _ c o n s l e g e _ a n d _ t h e _ s h e _ t o _ w e r e _ i
n _ t h e _ w a y _ t o _ t h e _ s o m e r i c t _ t h e _ s t a n t _ l a t c h _ s p a n t _ s t a n _ a c c i
m e d _ i n _ t h e _ s t a r d _ a n d _ c o m p l e s _ w a s _ t h e _ m a u s i n g _ a n d _ a _ s o m
p e r s _ t o _ w a s _ s e e _ i s _ t h e _ t o _ c o n t e n t s _ a n d _ o n e _ a n d _ m a n k e n t _ i
n _ t h e _ o f f e r s </s>

_ t h e _ b e a n _ t h e _ b e _ w a y _ t h e _ p r o m a n s </s>
```

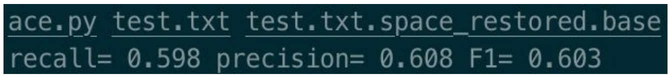
3. The results of NLMs is better than the results of trigrams because most of the generated words in NLM make sense but results of trigrams don’t make any sense.

4. Restoring Spaces (4 pts)

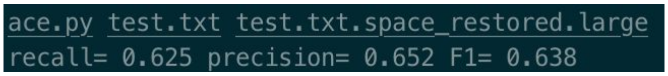
Restoring command:

```
cat test.txt.nospaces | sed -e 's/ /_/g;s/\(.\)/\1 /g' | awk '{printf("<s> %s </s>\n", $0)}' | carmel -sribIEWk 1 restore_spaces.base.wfst > test.txt.space_restored.base
```

1.



```
ace.py test.txt test.txt.space_restored.base
recall= 0.598 precision= 0.608 F1= 0.603
```



```
ace.py test.txt test.txt.space_restored.large
recall= 0.625 precision= 0.652 F1= 0.638
```

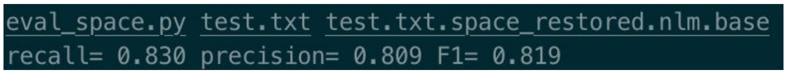
2.

```
H ← NLM MODEL
BEAM ← (0, H)           # SCORE, STATE
B ← SIZE
FOR CHAR IN SENTENCE:
        NEW_BEAM ← [],
        FOR SCORE, STATE IN BEAM:
                CALCULATE NEW SCORE AND STATE,      # W/O SPACE
                NEW_BEAM += (NEW SCORE, NEW STATE)
                CALCULATE NEW SCORE AND STATE,      # W/    SPACE
                NEW_BEAM += (NEW SCORE, NEW STATE)
        BEAM ← SORT_REVERSE(NEW_BEAM)[:B]
```

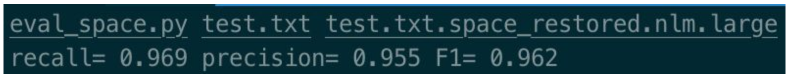
Time: O(2mb)   Space: O(mb)

Where b is the beam width, and m is the maximum depth of any path in the search tree.

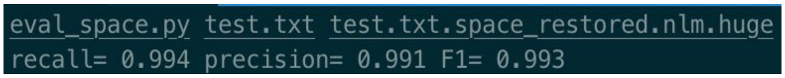
3.



```
eval_space.py test.txt test.txt.space_restored.nlm.base
recall= 0.830 precision= 0.809 F1= 0.819
```



```
eval_space.py test.txt test.txt.space_restored.nlm.large
recall= 0.969 precision= 0.955 F1= 0.962
```



```
eval_space.py test.txt test.txt.space_restored.nlm.huge
recall= 0.994 precision= 0.991 F1= 0.993
```

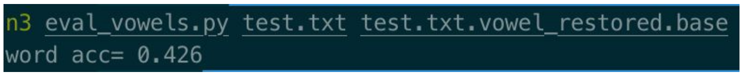
The results are almost the same with the expected results.

5. Restoring Vowels (4 pts)

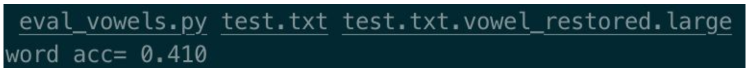
Restoring command:

```
cat test.txt.novowels | sed -e 's/ /_/g;s/\(.\)/\1 /g' | awk '{printf("<s> %s </s>\n", $0)}' | carmel
-sribIEWk 1 restore_vowels.base.wfst > test.txt.vowel_restored.base
```

1.



```
n3 eval_vowels.py test.txt test.txt.vowel_restored.base
word acc= 0.426
```



```
 eval_vowels.py test.txt test.txt.vowel_restored.large
word acc= 0.410
```

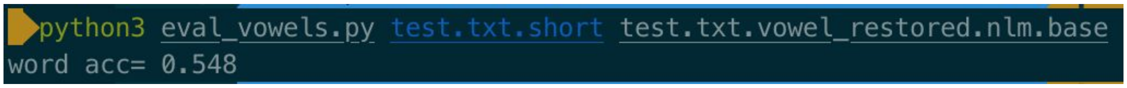
2.

```
H ← NLM MODEL
BEAM ← (0, H)          # SCORE, STATE
B ← SIZE
REPEAT:
        TMP ←  []
        FOR CHAR IN SENTENCE:
                NEW_BEAM ← [],
                PREV ← [BEAM],
                FOR SCORE, STATE IN PREV[-1]:
                        CALCULATE NEW SCORE AND STATE,          # W/O VOWEL
                        NEW_BEAM += (NEW SCORE, NEW STATE)
                        FOR VOWELS:
                                CALCULATE NEW SCORE AND STATE,# W/    VOWEL
                                TMP += (NEW SCORE, NEW STATE)
                PREV += TMP
        BEAM ← SORT_REVERSE(NEW_BEAM)[:B]
```

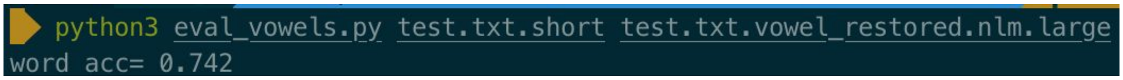
Time: O(2mb)   Space: O(mb)

Where b is the beam width, and m is the maximum depth of any path in the search tree.

3. Since it would take a long time to run restore vowels for 100 examples, I restored the first two sentences instead.



```
python3 eval_vowels.py test.txt.short test.txt.vowel_restored.nlm.base
word acc= 0.548
```



```
 python3 eval_vowels.py test.txt.short test.txt.vowel_restored.nlm.large
word acc= 0.742
```

We can see that the accuracy of base is near 54%.

Although the accuracy of large is slightly lower than 80%, we will get closer if we run it for 100 examples.