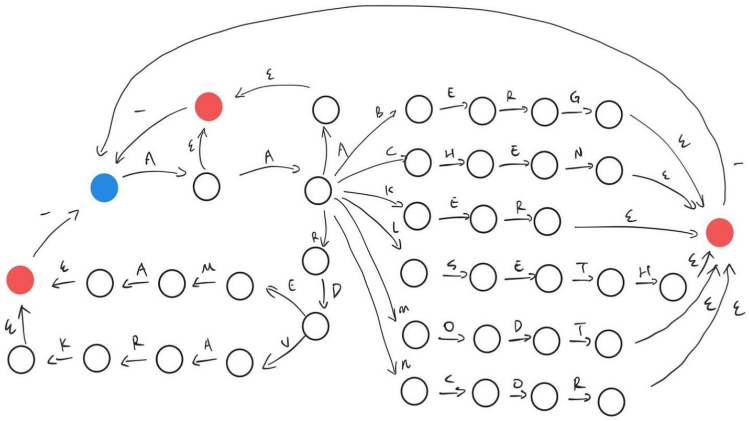
Natural Language Processing HW 1

Jing Huang 933039277
Chao-Ting Wen 933311010
Yu-Hsiang Liu 933951866

1. Finite-state acceptors

1.1.



1.2.

a.

```
Number of states in result: 256331
Number of arcs in result: 361732
Number of paths in result (valid for acyclic only; a cycle means infinitely many
): 105402.000000005
Number of cycle-causing arcs in result: 1
```

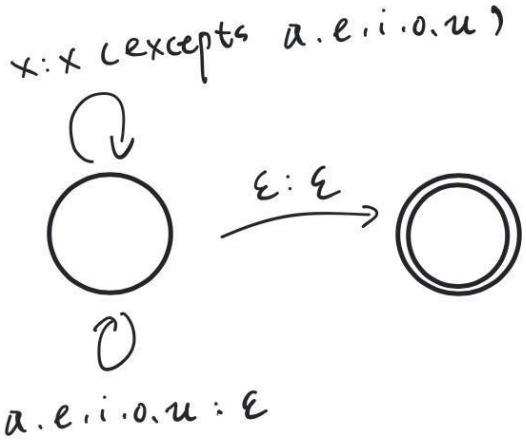
b.

```
Input line 1: LIST_THE_FLIGHTS_FROM_BALTIMORE_TO_SEATTL
E_THAT_STOP_IN_MINNEAPOLIS
     (102 states / 101 arcs reduce-> 79/78)
LIST_THE_FLIGHTS_FROM_BALTIMORE_TO_SEATTLE_THAT
_STOP_IN_MINNEAPOLIS
Input line 2: DOES_THIS_FLIGHT_SERVE_DINNER
     (48 states / 47 arcs reduce-> 35/34)
DOES_THIS_FLIGHT_SERVE_DINNER
Input line 3: I_NEED_A_FLIGHT_TO_SEATTLE_LEAVING_FROM_B
ALTIMORE_MAKING_A_STOP_IN_MINNEAPOLIS
     (119 states / 118 arcs reduce-> 93/92)
I_NEED_A_FLIGHT_TO_SEATTLE_LEAVING_FROM_BALTIMO
RE_MAKING_A_STOP_IN_MINNEAPOLIS
Input line 4: I_NEED_TO_HAVE_DINNER_SERVED
     (48 states / 47 arcs reduce-> 35/34)
I_NEED_TO_HAVE_DINNER_SERVED
Input line 5: I_HAVE_TWO_FRIENDS_THAT_WOULD_LIKE_TO_VIS
IT_ME_ON_WEDNESDAY_HERE_IN_WASHINGTON_D_C
     (128 states / 127 arcs reduce-> 100/99)
I_HAVE_TWO_FRIENDS_THAT_WOULD_LIKE_TO_VISIT_ME_
ON_WEDNESDAY_HERE_IN_WASHINGTON_D_C
```

```
Input line 1: I_WUNT_TO_LEEVE_MONDAY_MORNING
      (0 states / 0 arcs)
Empty or invalid result of composition with transducer "english.fsa".

Input line 2: NOW_I_NEAD_A_FLIGHT_ON_TOOSDAY_FROM_PHE
ENIX_TO_DETROIT
      (0 states / 0 arcs)
Empty or invalid result of composition with transducer "english.fsa".

Input line 3: WHICH_ONES_LEEVE_IN_THE_MORNING
      (0 states / 0 arcs)
Empty or invalid result of composition with transducer "english.fsa".

Input line 4: WHICH_ONES_ARRIVE_ERLY_IN_THE_DAY
      (0 states / 0 arcs)
Empty or invalid result of composition with transducer "english.fsa".

Input line 5: I_NEAD_A_FLIGHT_FROM_PHEENIX_TO_DETROIT_L
EEVING_MONDAY_EEVENING
      (0 states / 0 arcs)
Empty or invalid result of composition with transducer "english.fsa".
```

2. Finite-state transducers

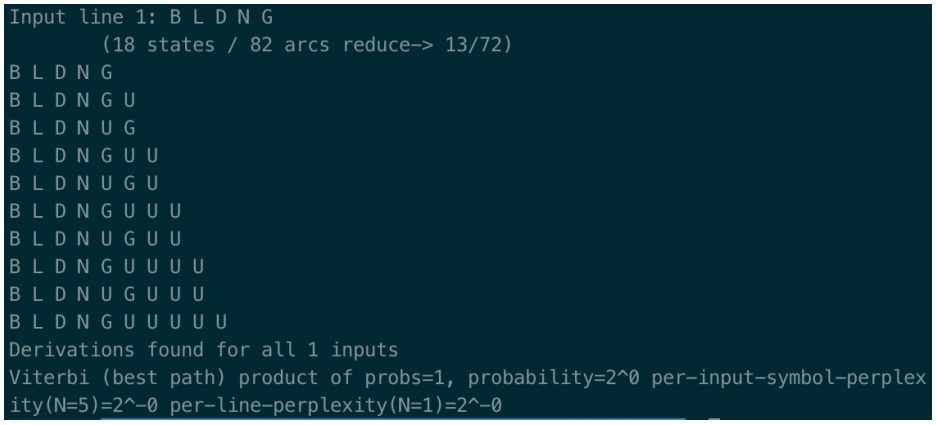
3. a)



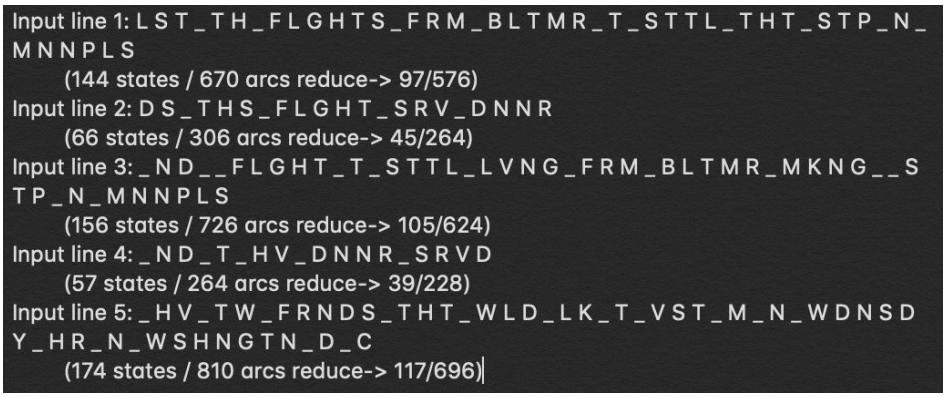
b)

```
Input line 1: LIST_THE_FLIGHTS_FROM_BALTIMORE_TO_SEATTL
E_THAT_STOP_IN_MINNEAPOLIS
      (136 states / 135 arcs reduce-> 69/68)
LST_TH_FLGHTS_FRM_BLTMR_T_STTL_THT_STP_N_MNNPL
S
Input line 2: DOES_THIS_FLIGHT_SERVE_DINNER
      (60 states / 59 arcs reduce-> 31/30)
DS_THS_FLGHT_SRV_DNNR
Input line 3: I_NEED_A_FLIGHT_TO_SEATTLE_LEAVING_FROM_B
ALTIMORE_MAKING_A_STOP_IN_MINNEAPOLIS
      (158 states / 157 arcs reduce-> 80/79)
_ND__FLGHT_T_STTL_LVNG_FRM_BLTMR_MKNG__STP_N_M
NNPLS
Input line 4: I_NEED_TO_HAVE_DINNER_SERVED
      (58 states / 57 arcs reduce-> 30/29)
_ND_T_HV_DNNR_SRVD
Input line 5: I_HAVE_TWO_FRIENDS_THAT_WOULD_LIKE_TO_VIS
IT_ME_ON_WEDNESDAY_HERE_IN_WASHINGTON_D_C
      (166 states / 165 arcs reduce-> 84/83)
_HV_TW_FRNDS_THT_WLD_LK_T_VST_M_N_WDNSDY_HR_N_
WSHNGTN_D_C
```

c)



4. a)



b)

0.012865497076023392

c)

There are too many possible combinations.

For example, ‘Y O U’ will become ‘Y’ after calling remove-vowels. While using backward application, ‘Y’ can be ‘Y A’, ‘Y O’, … and so on.

3. Combining FSAs and FSTs

5. a)

We use command ‘carmel fsa fst’ to combine ‘english.fsa’ with ‘remove-vowels.fst’.

b)

c)

The results are acceptable.

Because we know patterns of word-formation. On the other hand, machines can only depend on the repeated test.

6. a)

Finding all the vocabulary in strings. (cheat_list)
Also, sort them with the by frequency.
Creating english.fsa by wordlist in strings. So, it can decrease the amount of vocabulary.
cat vocab | python make.py > english.fsa cat cheat_list | python make.py > english.fsa
Combine English.fsa and remove-vowels.fst to combine.fst.

b)

Implement it and try it on strings.novowels. > strings.restored

```
L I S T _ T H E _ F L I G H T S _ F R O M _ B A L T I M O R E _ T O _ S E A T T
L E _ T H A T _ S T O P _ I N _ M I N N E A P O L I S
D O E S _ T H E S E _ F L I G H T _ S E R V E _ D I N N E R
I _ N E E D _ I _ F L I G H T _ T O _ S E A T T L E _ L E A V I N G _ F R O M _
B A L T I M O R E _ M A K I N G _ I T _ S T O P _ I N _ M I N N E A P O L I S
I _ N E E D _ T O _ H A V E _ D I N N E R _ S E R V E D
I _ H A V E _ T W O _ F R I E N D S _ T H A T _ W O U L D _ L I K E _ T O _ V I
S I T _ M E _ I N _ W E D N E S D A Y _ H E R E _ I N _ W A S H I N G T O N _ D
O _ C
I N _ O F _ T H E M _ L I V E S _ I N _ D E N V E R _ N E E D _ T H E _ T H E R
E _ L I V E S _ I N _ M I A M I
I _ H A V E _ I _ F R I E N D _ L E A V I N G _ I N _ D E N V E R _ T H A T _ W
O U L D _ L I K E _ T O _ V I S I T _ M E _ H E R E _ I N _ W A S H I N G T O N
_ D O _ C
I _ N E E D _ I _ F L I G H T _ F R O M _ I N D I A N A P O L I S _ T O _ H O U
S T O N _ I N _ T O _ W _ I
I _ N E E D _ I _ F L I G H T _ F R O M _ A T L A N T A _ T O _ C H A R L O T T
E _ N O R T H _ C A R O L I N A _ N E X T _ M O N D A Y
I _ N E E D _ T O _ R E T U R N _ I N _ T U E S D A Y
H O W _ M U C H _ E X T R A _ W O U L D _ I _ F I R S T _ C L A S S _ F A R E _
C O S T
H O W _ M U C H _ W O U L D _ T H E _ C O A C H _ F A R E _ C O S T
I _ W O U L D _ L I K E _ T O _ F I N D _ I _ F L I G H T _ T H A T _ G O E S _
F R O M _ B O S T O N _ T O _ O R L A N D O
```

c)

```
(base) 10-248-121-225:documents tony$ python eval.py strings strings.rest
ored
0.9134502923976608  animation.py  carmel  chess.cpp  cl
```