CmpE 493
Spring 2020
Assignment 1: Spelling Error Corrector

Gurkan Demir
2015400177

March 15, 2020

Contents

1 Introduction

In this assignment, we are expected to implement an isolated word spelling error corrector based on the noisy channel model. We are given set of files which are;

1. ***corpus.txt :*** Used for finding word frequencies.
2. ***spell-errors.txt :*** Used for constructing confusion matrices according to error types.
3. ***test-words-misspelled.txt :*** Example set of misspelled words for testing.
4. ***test-words-correct.txt :*** Example set of corrected versions of misspelled words for testing.

While creating a dictionary, we are expected to tokenize the file and perform case-folding. During correction of misspelled words, we are expected to find candidates with edit distance 1 according to Damerau-Levenshtein.

Also, we are expected to implement two versions of this corrector. First one is calculating probability of candidates with the help of language and noisy channel model without smooting. Second one is using add-one smoothing (Laplace smooting with alpha = 1).

Our program is expected to take a file containing a list of misspelled words (one word per line) as input, and produce a file with the predicted correct spellings of these words (one word per line) as output. If our program can not produce predictions for any of the words in the input file, the corresponding lines in the output file should be printed as blank lines.

2 Implementation

1. Read corpus, tokenize and construct dictionary.
2. Read spelling errors, and construct confusion matrices.
3. Get misspelled word, and find all words in corpus with edit distance 1.
4. Using language and noisy channel model, calculate probability of each candidate.
5. Return most probable candidate.

After performing case-folding while reading corpus, tokenization is done. Tokenization is performed by replacing each non-alpha character with space.

3 How to Run?

In order to run error spelling corrector, execute the following from the command line:

```
python3 corrector.py –corpus [CORPUS_FILE] –spell_errors [SPELL_ERRORS_FILE]
–misspelled [MISSPELLED_FILE] –correct [CORRECT_FILE] –smooth
–print_confusions
```

where;

1. **CORPUS_FILE :** Path of corpus.txt (***required***).
2. **SPELL_ERRORS_FILE :** Path of spell-errors.txt (***required***).
3. **MISSPELLED_FILE :** Path of misspelled words file (***required***).
4. **CORRECT_FILE :** Path of correct words file (***not required***).
5. **smooth :** Whether use alpha smoothing or not (***not required***).
6. **print_confusions :** Whether print confusion matrices or not (***not required***).

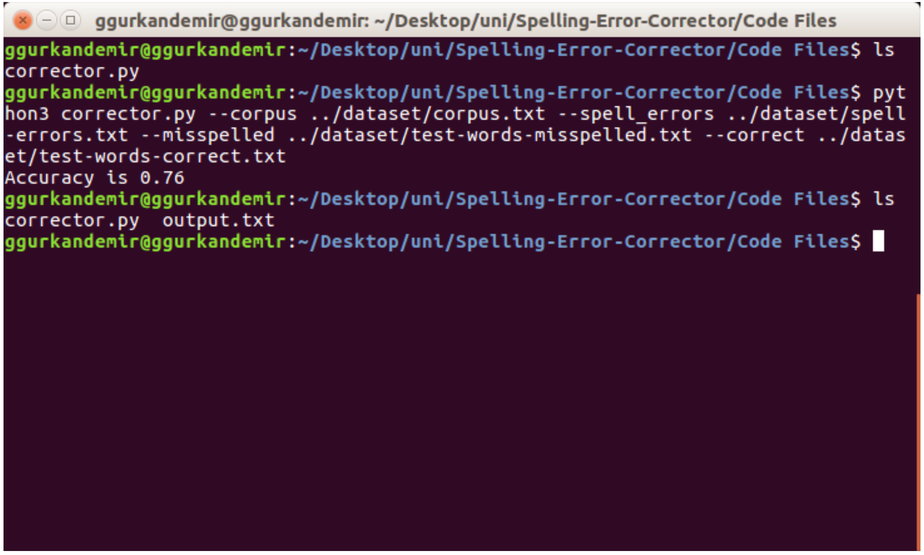
3.1 Notes

1. Algorithm prints corrected versions of misspelled words in a file named output.txt.
2. Output file is located in the same directory with the execution.
3. CORRECT_FILE is not required while execution.
4. It is required when calculating accuracy is needed.
5. smooth is not required.
6. If you execute algorithm with smooth, it implements alpha smoothing. Otherwise, it does not implement smoothing.
7. If you execute algorithm with print_confusions, it prints confusion matrices by creating new text files.

Details of starting execution is mentioned in ReadME file.

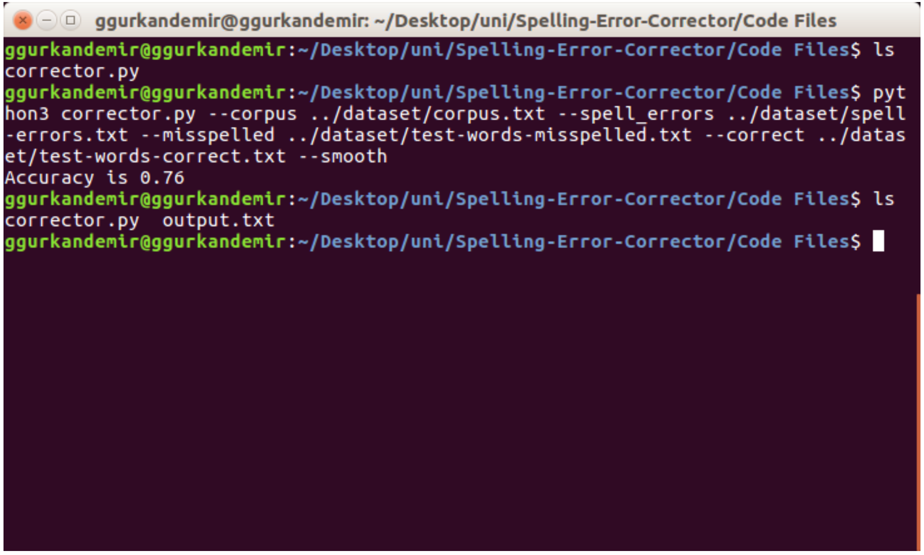
4 Screenshots of Running System

- Without Smoothing


ggurkandemir@ggurkandemir: ~/Desktop/uni/Spelling-Error-Corrector/Code Files

```
ggurkandemir@ggurkandemir:~/Desktop/uni/Spelling-Error-Corrector/Code Files$ ls
corrector.py
ggurkandemir@ggurkandemir:~/Desktop/uni/Spelling-Error-Corrector/Code Files$ pyt
hon3 corrector.py --corpus ../dataset/corpus.txt --spell_errors ../dataset/spell
-errors.txt --misspelled ../dataset/test-words-misspelled.txt --correct ../datas
et/test-words-correct.txt
Accuracy is 0.76
ggurkandemir@ggurkandemir:~/Desktop/uni/Spelling-Error-Corrector/Code Files$ ls
corrector.py  output.txt
ggurkandemir@ggurkandemir:~/Desktop/uni/Spelling-Error-Corrector/Code Files$ 
```

- Smoothing


ggurkandemir@ggurkandemir: ~/Desktop/uni/Spelling-Error-Corrector/Code Files

```
ggurkandemir@ggurkandemir:~/Desktop/uni/Spelling-Error-Corrector/Code Files$ ls
corrector.py
ggurkandemir@ggurkandemir:~/Desktop/uni/Spelling-Error-Corrector/Code Files$ pyt
hon3 corrector.py --corpus ../dataset/corpus.txt --spell_errors ../dataset/spell
-errors.txt --misspelled ../dataset/test-words-misspelled.txt --correct ../datas
et/test-words-correct.txt --smooth
Accuracy is 0.76
ggurkandemir@ggurkandemir:~/Desktop/uni/Spelling-Error-Corrector/Code Files$ ls
corrector.py  output.txt
ggurkandemir@ggurkandemir:~/Desktop/uni/Spelling-Error-Corrector/Code Files$ 
```

5 Assumptions

- All words in corpus are spelled correctly.
- Empty string is returned as output for misspelled words which have edit

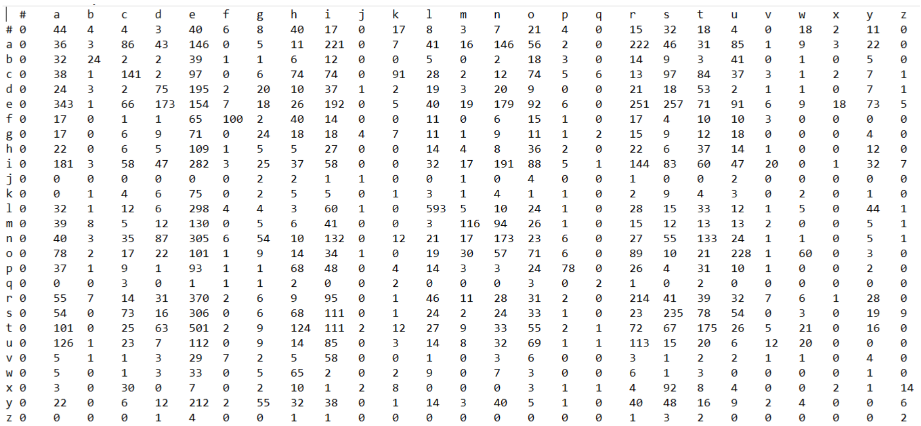
distance more than 1.

- Laplace smoothing is implemented using alpha = 1.

6 Outputs

6.1 Confusion Matrices

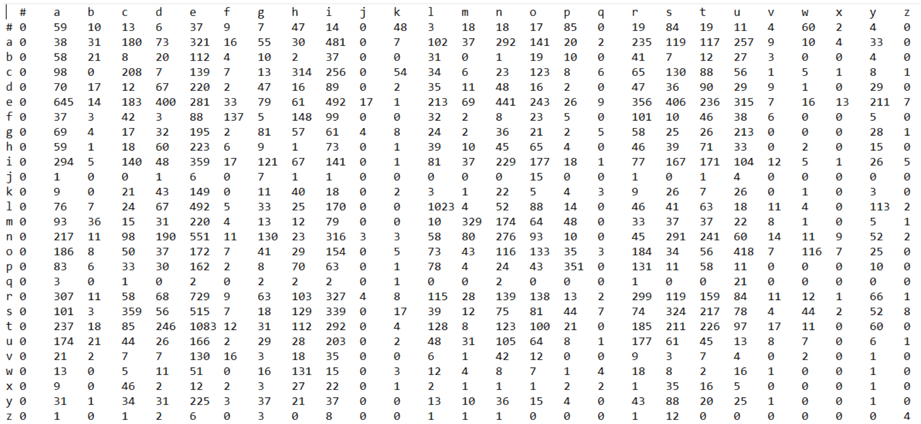
- Insertion



	#	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
#	0	44	4	4	3	40	6	8	40	17	0	17	8	3	7	21	4	0	15	32	18	4	0	18	2	11	0
a	0	36	3	86	43	146	0	5	11	221	0	7	41	16	146	56	2	0	222	46	31	85	1	9	3	22	0
b	0	32	24	2	2	39	1	1	6	12	0	0	5	0	2	18	3	0	14	9	3	41	0	1	0	5	0
c	0	38	1	141	2	97	0	6	74	74	0	91	28	2	12	74	5	6	13	97	84	37	3	1	2	7	1
d	0	24	3	2	75	195	2	20	10	37	1	2	19	3	20	9	0	0	21	18	53	2	1	1	0	7	1
e	0	343	1	66	173	154	7	18	26	192	0	5	40	19	179	92	6	0	251	257	71	91	6	9	18	73	5
f	0	17	0	1	1	65	100	2	40	14	0	0	11	0	6	15	1	0	17	4	10	10	3	0	0	0	0
g	0	17	0	6	9	71	0	24	18	18	4	7	11	1	9	11	1	2	15	9	12	18	0	0	0	4	0
h	0	22	0	6	5	109	1	5	5	27	0	0	14	4	8	36	2	0	22	6	37	14	1	0	0	12	0
i	0	181	3	58	47	282	3	25	37	58	0	0	32	17	191	88	5	1	144	83	60	47	20	0	1	32	7
j	0	0	0	0	0	0	0	2	2	1	1	0	0	1	0	4	0	0	1	0	0	2	0	0	0	0	0
k	0	0	1	4	6	75	0	2	5	5	0	1	3	1	4	1	1	0	2	9	4	3	0	2	0	1	0
l	0	32	1	12	6	298	4	4	3	60	1	0	593	5	10	24	1	0	28	15	33	12	1	5	0	44	1
m	0	39	8	5	12	130	0	5	6	41	0	0	3	116	94	26	1	0	15	12	13	13	2	0	0	5	1
n	0	40	3	35	87	305	6	54	10	132	0	12	21	17	173	23	6	0	27	55	133	24	1	1	0	5	1
o	0	78	2	17	22	101	1	9	14	34	1	0	19	30	57	71	6	0	89	10	21	228	1	60	0	3	0
p	0	37	1	9	1	93	1	1	68	48	0	4	14	3	3	24	78	0	26	4	31	10	1	0	0	2	0
q	0	0	0	3	0	1	1	1	2	0	0	2	0	0	0	3	0	2	1	0	2	0	0	0	0	0	0
r	0	55	7	14	31	370	2	6	9	95	0	1	46	11	28	31	2	0	214	41	39	32	7	6	1	28	0
s	0	54	0	73	16	306	0	6	68	111	0	1	24	2	24	33	1	0	23	235	78	54	0	3	0	19	9
t	0	101	0	25	63	501	2	9	124	111	2	12	27	9	33	55	2	1	72	67	175	26	5	21	0	16	0
u	0	126	1	23	7	112	0	9	14	85	0	3	14	8	32	69	1	1	113	15	20	6	12	20	0	0	0
v	0	5	1	1	3	29	7	2	5	58	0	0	1	0	3	6	0	0	3	1	2	2	1	1	0	4	0
w	0	5	0	1	3	33	0	5	65	2	0	2	9	0	7	3	0	0	6	1	3	0	0	0	0	1	0
x	0	3	0	30	0	7	0	2	10	1	2	8	0	0	0	3	1	1	4	92	8	4	0	0	2	1	14
y	0	22	0	6	12	212	2	55	32	38	0	1	14	3	40	5	1	0	40	48	16	9	2	4	0	0	6
z	0	0	0	0	1	4	0	0	1	1	0	0	0	0	0	0	0	0	1	3	2	0	0	0	0	0	2

Insertion(x, y) refers to number of errors ***x typed as xy***.

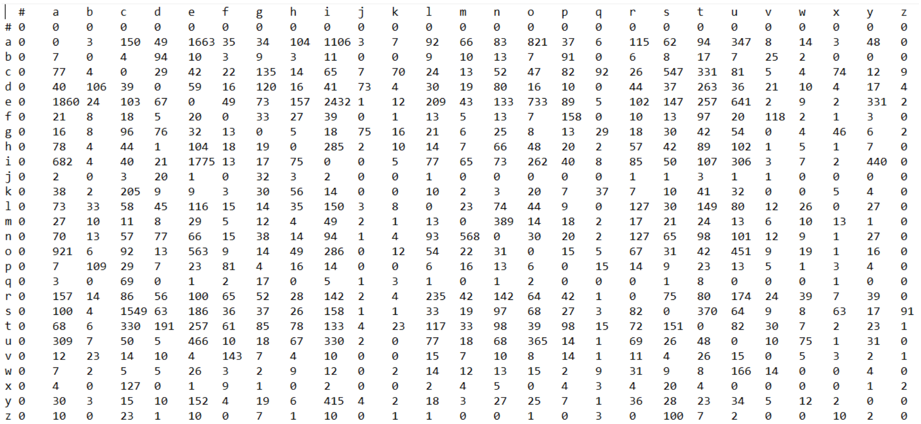
- Deletion



	#	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
#	0	59	10	13	6	37	9	7	47	14	0	48	3	18	18	17	85	0	19	84	19	11	4	60	2	4	0
a	0	38	31	180	73	321	16	55	30	481	0	7	102	37	292	141	20	2	235	119	117	257	9	10	4	33	0
b	0	58	21	8	20	112	4	10	2	37	0	0	31	0	1	19	10	0	41	7	12	27	3	0	0	4	0
c	0	98	0	208	7	139	7	13	314	256	0	54	34	6	23	123	8	6	65	130	88	56	1	5	1	8	1
d	0	70	17	12	67	220	2	47	16	89	0	2	35	11	48	16	2	0	47	36	90	29	9	1	0	29	0
e	0	645	14	183	400	281	33	79	61	492	17	1	213	69	441	243	26	9	356	406	236	315	7	16	13	211	7
f	0	37	3	42	3	88	137	5	148	99	0	0	32	2	8	23	5	0	101	10	46	38	6	0	0	5	0
g	0	69	4	17	32	195	2	81	57	61	4	8	24	2	36	21	2	5	58	25	26	213	0	0	0	28	1
h	0	59	1	18	60	223	6	9	1	73	0	1	39	10	45	65	4	0	46	39	71	33	0	2	0	15	0
i	0	294	5	140	48	359	17	121	67	141	0	1	81	37	229	177	18	1	77	167	171	104	12	5	1	26	5
j	0	1	0	0	1	6	0	7	1	1	0	0	0	0	0	15	0	0	1	0	1	4	0	0	0	0	0
k	0	9	0	21	43	149	0	11	40	18	0	2	3	1	22	5	4	3	9	26	7	26	0	1	0	3	0
l	0	76	7	24	67	492	5	33	25	170	0	0	1023	4	52	88	14	0	46	41	63	18	11	4	0	113	2
m	0	93	36	15	31	220	4	13	12	79	0	0	10	329	174	64	48	0	33	37	37	22	8	1	0	5	1
n	0	217	11	98	190	551	11	130	23	316	3	3	58	80	276	93	10	0	45	291	241	60	14	11	9	52	2
o	0	186	8	50	37	172	7	41	29	154	0	5	73	43	116	133	35	3	184	34	56	418	7	116	7	25	0
p	0	83	6	33	30	162	2	8	70	63	0	1	78	4	24	43	351	0	131	11	58	11	0	0	0	10	0
q	0	3	0	1	0	2	0	2	2	2	0	1	0	0	2	0	0	0	1	0	0	21	0	0	0	0	0
r	0	307	11	58	68	729	9	63	103	327	4	8	115	28	139	138	13	2	299	119	159	84	11	12	1	66	1
s	0	101	3	359	56	515	7	18	129	339	0	17	39	12	75	81	44	7	74	324	217	78	4	44	2	52	8
t	0	237	18	85	246	1083	12	31	112	292	0	4	128	8	123	100	21	0	185	211	226	97	17	11	0	60	0
u	0	174	21	44	26	166	2	29	28	203	0	2	48	31	105	64	8	1	177	61	45	13	8	7	0	6	1
v	0	21	2	7	7	130	16	3	18	35	0	0	6	1	42	12	0	0	9	3	7	4	0	2	0	1	0
w	0	13	0	5	11	51	0	16	131	15	0	3	12	4	8	7	1	4	18	8	2	16	1	0	0	1	0
x	0	9	0	46	2	12	2	3	27	22	0	1	2	1	1	1	2	2	1	35	16	5	0	0	0	1	0
y	0	31	1	34	31	225	3	37	21	37	0	0	13	10	36	15	4	0	43	88	20	25	1	0	0	1	0
z	0	1	0	1	2	6	0	3	0	8	0	0	1	1	1	0	0	0	1	12	0	0	0	0	0	0	4

Deletion(x, y) refers to number of errors ***xy typed as x***.

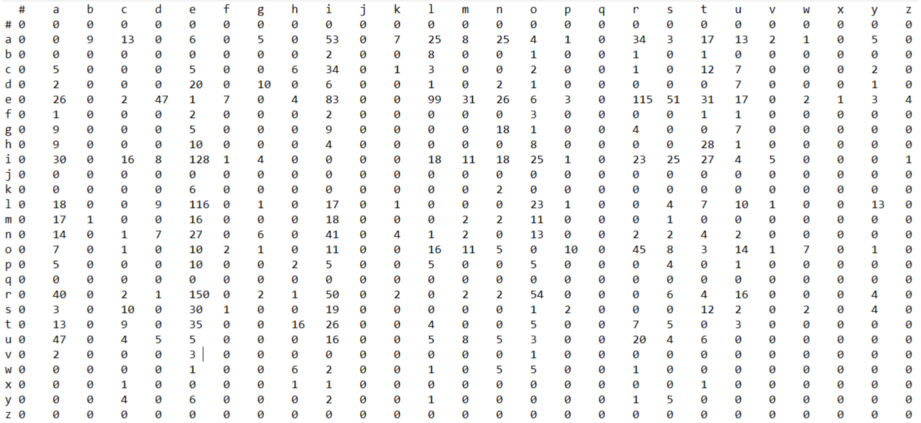
- Substitution



	#	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	0	0	3	150	49	1663	35	34	104	1106	3	7	92	66	83	821	37	6	115	62	94	347	8	14	3	48	0
b	0	7	0	4	94	10	3	9	3	11	0	0	9	10	13	7	91	0	6	8	17	7	25	2	0	0	0
c	0	77	4	0	29	42	22	135	14	65	7	70	24	13	52	47	82	92	26	547	331	81	5	4	74	12	9
d	0	40	106	39	0	59	16	120	16	41	73	4	30	19	80	16	10	0	44	37	263	36	21	10	4	17	4
e	0	1860	24	103	67	0	49	73	157	2432	1	12	209	43	133	733	89	5	102	147	257	641	2	9	2	331	2
f	0	21	8	18	5	20	0	33	27	39	0	1	13	5	13	7	158	0	10	13	97	20	118	2	1	3	0
g	0	16	8	96	76	32	13	0	5	18	75	16	21	6	25	8	13	29	18	30	42	54	0	4	46	6	2
h	0	78	4	44	1	104	18	19	0	285	2	10	14	7	66	48	20	2	57	42	89	102	1	5	1	7	0
i	0	682	4	40	21	1775	13	17	75	0	0	5	77	65	73	262	40	8	85	50	107	306	3	7	2	440	0
j	0	2	0	3	20	1	0	32	3	2	0	0	1	0	0	0	0	0	1	1	3	1	1	0	0	0	0
k	0	38	2	205	9	9	3	30	56	14	0	0	10	2	3	20	7	37	7	10	41	32	0	0	5	4	0
l	0	73	33	58	45	116	15	14	35	150	3	8	0	23	74	44	9	0	127	30	149	80	12	26	0	27	0
m	0	27	10	11	8	29	5	12	4	49	2	1	13	0	389	14	18	2	17	21	24	13	6	10	13	1	0
n	0	70	13	57	77	66	15	38	14	94	1	4	93	568	0	30	20	2	127	65	98	101	12	9	1	27	0
o	0	921	6	92	13	563	9	14	49	286	0	12	54	22	31	0	15	5	67	31	42	451	9	19	1	16	0
p	0	7	109	29	7	23	81	4	16	14	0	0	6	16	13	6	0	15	14	9	23	13	5	1	3	4	0
q	0	3	0	69	0	1	2	17	0	5	1	3	1	0	1	2	0	0	0	1	8	0	0	0	1	0	0
r	0	157	14	86	56	100	65	52	28	142	2	4	235	42	142	64	42	1	0	75	80	174	24	39	7	39	0
s	0	100	4	1549	63	186	36	37	26	158	1	1	33	19	97	68	27	3	82	0	370	64	9	8	63	17	91
t	0	68	6	330	191	257	61	85	78	133	4	23	117	33	98	39	98	15	72	151	0	82	30	7	2	23	1
u	0	309	7	50	5	466	10	18	67	330	2	0	77	18	68	365	14	1	69	26	48	0	10	75	1	31	0
v	0	12	23	14	10	4	143	7	4	10	0	0	15	7	10	8	14	1	11	4	26	15	0	5	3	2	1
w	0	7	2	5	5	26	3	2	9	12	0	2	14	12	13	15	2	9	31	9	8	166	14	0	0	4	0
x	0	4	0	127	0	1	9	1	0	2	0	0	2	4	5	0	4	3	4	20	4	0	0	0	0	1	2
y	0	30	3	15	10	152	4	19	6	415	4	2	18	3	27	25	7	1	36	28	23	34	5	12	2	0	0
z	0	10	0	23	1	10	0	7	1	10	0	1	1	0	0	1	0	3	0	100	7	2	0	0	10	2	0

Substitution(x, y) refers to number of errors ***y typed as x***.

- Transpose



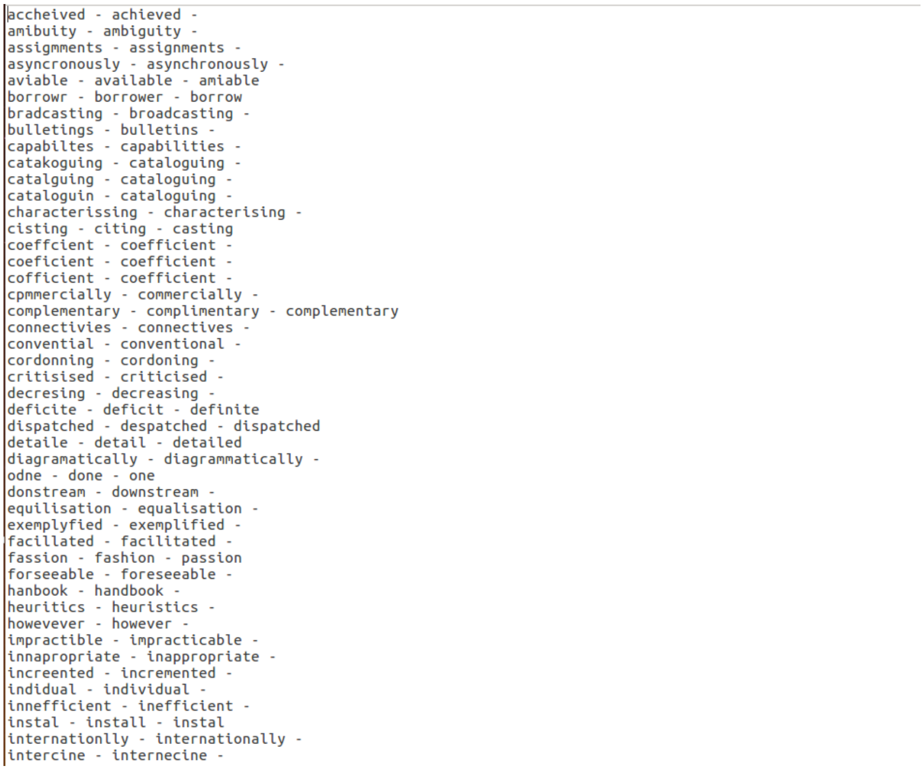
	#	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	0	0	9	13	0	6	0	5	0	53	0	7	25	8	25	4	1	0	34	3	17	13	2	1	0	5	0
b	0	0	0	0	0	0	0	0	0	2	0	0	8	0	0	1	0	0	1	0	1	0	0	0	0	0	0
c	0	5	0	0	0	5	0	0	6	34	0	1	3	0	0	2	0	0	1	0	12	7	0	0	0	2	0
d	0	2	0	0	0	20	0	10	0	6	0	0	1	0	2	1	0	0	0	0	0	7	0	0	0	1	0
e	0	26	0	2	47	1	7	0	4	83	0	0	99	31	26	6	3	0	115	51	31	17	0	2	1	3	4
f	0	1	0	0	0	2	0	0	0	2	0	0	0	0	0	3	0	0	0	0	1	1	0	0	0	0	0
g	0	9	0	0	0	5	0	0	0	9	0	0	0	0	18	1	0	0	4	0	0	7	0	0	0	0	0
h	0	9	0	0	0	10	0	0	0	4	0	0	0	0	0	8	0	0	0	0	28	1	0	0	0	0	0
i	0	30	0	16	8	128	1	4	0	0	0	0	18	11	18	25	1	0	23	25	27	4	5	0	0	0	1
j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
k	0	0	0	0	0	6	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
l	0	18	0	0	9	116	0	1	0	17	0	1	0	0	0	23	1	0	0	4	7	10	1	0	0	13	0
m	0	17	1	0	0	16	0	0	0	18	0	0	0	2	2	11	0	0	0	1	0	0	0	0	0	0	0
n	0	14	0	1	7	27	0	6	0	41	0	4	1	2	0	13	0	0	2	2	4	2	0	0	0	0	0
o	0	7	0	1	0	10	2	1	0	11	0	0	16	11	5	0	10	0	45	8	3	14	1	7	0	1	0
p	0	5	0	0	0	10	0	0	2	5	0	0	5	0	0	5	0	0	0	4	0	1	0	0	0	0	0
q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	40	0	2	1	150	0	2	1	50	0	2	0	2	2	54	0	0	0	6	4	16	0	0	0	4	0
s	0	3	0	10	0	30	1	0	0	19	0	0	0	0	0	1	2	0	0	0	12	2	0	2	0	4	0
t	0	13	0	9	0	35	0	0	16	26	0	0	4	0	0	5	0	0	7	5	0	3	0	0	0	0	0
u	0	47	0	4	5	5	0	0	0	16	0	0	5	8	5	3	0	0	20	4	6	0	0	0	0	0	0
v	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
w	0	0	0	0	0	1	0	0	6	2	0	0	1	0	5	5	0	0	1	0	0	0	0	0	0	0	0
x	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
y	0	0	0	4	0	6	0	0	0	2	0	0	1	0	0	0	0	0	1	5	0	0	0	0	0	0	0
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Transpose(x, y) refers to number of errors ***xy typed as yx***.

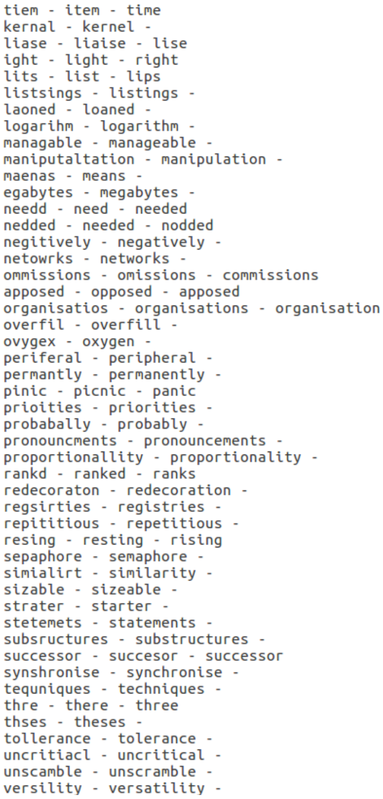
6.2 Without Smoothing

6.2.1 Corrections

Below you can find comparison between my system's output versus real output. There exist 3 columns which are misspelled word, real output, system's output, respectively.



```
accheived - achieved -
amibuity - ambiguity -
assigmments - assignments -
asyncronously - asynchronously -
aviable - available - amiable
borrowr - borrower - borrow
bradcasting - broadcasting -
bulletings - bulletins -
capabiltes - capabilities -
catakoguing - cataloguing -
catalguing - cataloguing -
cataloguin - cataloguing -
characterissing - characterising -
cisting - citing - casting
coeffcient - coefficient -
coeficient - coefficient -
cofficient - coefficient -
cpmmercially - commercially -
complementary - complimentary - complementary
connectivies - connectives -
convential - conventional -
cordonning - cordoning -
critisised - criticised -
decresing - decreasing -
deficite - deficit - definite
dispatched - despatched - dispatched
detaile - detail - detailed
diagramatically - diagrammatically -
odne - done - one
donstream - downstream -
equilisation - equalisation -
exemplyfied - exemplified -
facillated - facilitated -
fassion - fashion - passion
forseeable - foreseeable -
hanbook - handbook -
heuritics - heuristics -
howevever - however -
impractible - impracticable -
innapropriate - inappropriate -
increented - incremented -
indidual - individual -
innefficient - inefficient -
instal - install - instal
internationlly - internationally -
intercine - internecine -
```



```
tiem - item - time
kernal - kernel -
liase - liaise - lise
ight - light - right
lits - list - lips
listsings - listings -
laoned - loaned -
logarihm - logarithm -
managable - manageable -
maniputaltation - manipulation -
maenas - means -
egabytes - megabytes -
needd - need - needed
nedded - needed - nodded
negitively - negatively -
netowrks - networks -
ommissions - omissions - commissions
apposed - opposed - apposed
organisatios - organisations - organisation
overfil - overfill -
ovygex - oxygen -
periferal - peripheral -
permantly - permanently -
pinic - picnic - panic
prioities - priorities -
probabally - probably -
pronouncments - pronouncements -
proportionallity - proportionality -
rankd - ranked - ranks
redecoraton - redecoration -
regsirties - registries -
repititious - repetitious -
resing - resting - rising
sepaphore - semaphore -
simialirt - similarity -
sizable - sizeable -
strater - starter -
stetemets - statements -
subsructures - substructures -
successor - succesor - successor
synshronise - synchronise -
tequniques - techniques -
thre - there - three
thses - theses -
tollerance - tolerance -
uncritiacl - uncritical -
unscamble - unscramble -
versility - versatility -
```

6.2.2 Accuracy of System

There exist 384 misspelled words in test case, my system generates 94 wrong answers. We can calculate accuracy of system with the following function;

$$Accuracy = 1 - \#ofWrong/\#ofMisspelled \tag{1}$$

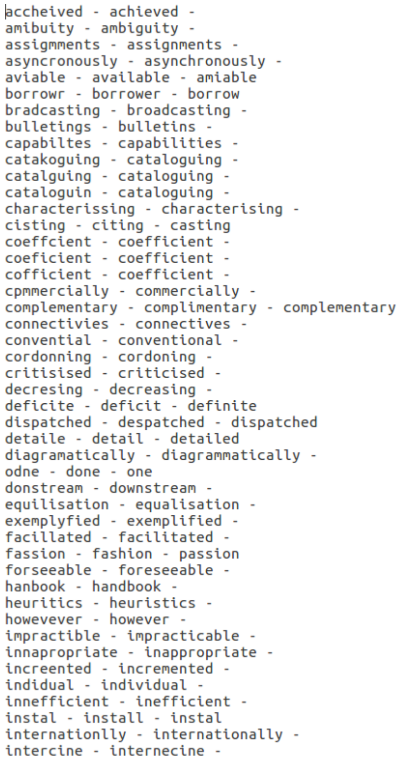
$$Accuracy = 1 - 95/384 \tag{2}$$

$$Accuracy = 0.7526 \tag{3}$$

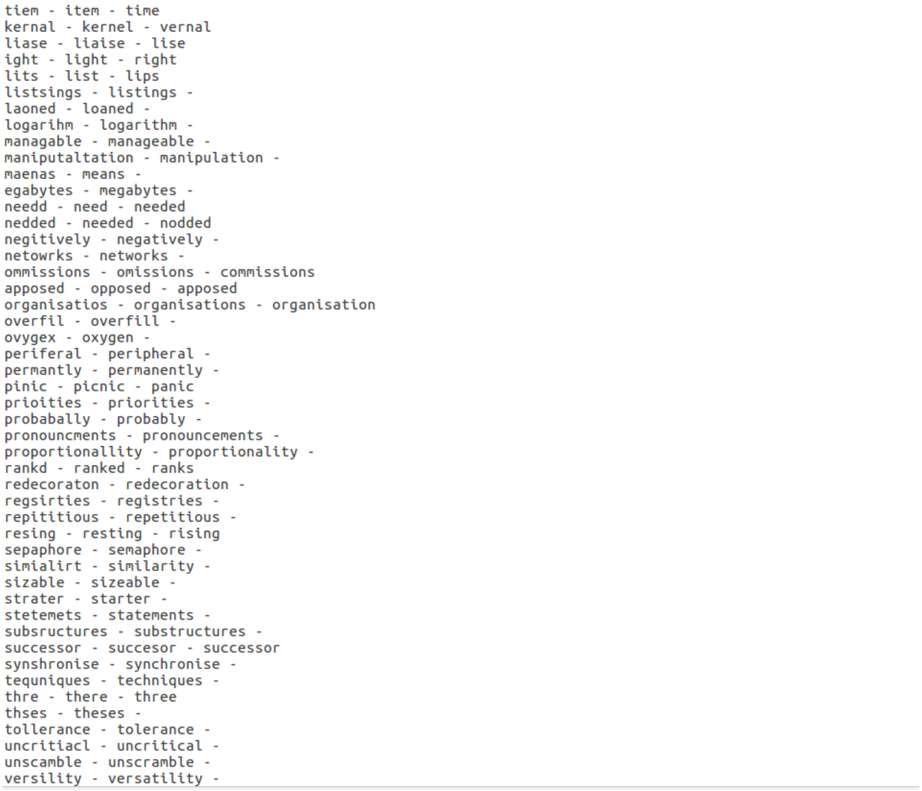
6.3 Smoothing

6.3.1 Corrections

Below you can find comparison between my system's output versus real output. There exist 3 columns which are misspelled word, real output, system's output, respectively.



```
accheived - achieved -
amibuity - ambiguity -
assigmments - assignments -
asyncronously - asynchronously -
aviable - available - amiable
borrowr - borrower - borrow
bradcasting - broadcasting -
bulletings - bulletins -
capabiltes - capabilities -
catakoguing - cataloguing -
catalguing - cataloguing -
cataloguin - cataloguing -
characterissing - characterising -
cisting - citing - casting
coeffcient - coefficient -
coeficient - coefficient -
cofficient - coefficient -
cpmmercially - commercially -
complementary - complimentary - complementary
connectivies - connectives -
convential - conventional -
cordonning - cordoning -
critisised - criticised -
decresing - decreasing -
deficite - deficit - definite
dispatched - despatched - dispatched
detaile - detail - detailed
diagramatically - diagrammatically -
odne - done - one
donstream - downstream -
equilisation - equalisation -
exemplyfied - exemplified -
facillated - facilitated -
fassion - fashion - passion
forseeable - foreseeable -
hanbook - handbook -
heuritics - heuristics -
howevever - however -
impractible - impracticable -
innapropriate - inappropriate -
increented - incremented -
indidual - individual -
innefficient - inefficient -
instal - install - instal
internationlly - internationally -
intercine - internecine -
```



```
tiem - item - time
kernal - kernel - vernal
liase - liaise - lise
ight - light - right
lits - list - lips
listsings - listings -
laoned - loaned -
logarihm - logarithm -
managable - manageable -
maniputaltation - manipulation -
maenas - means -
egabytes - megabytes -
needd - need - needed
nedded - needed - nodded
negitively - negatively -
netowrks - networks -
ommissions - omissions - commissions
apposed - opposed - apposed
organisatios - organisations - organisation
overfil - overfill -
ovygex - oxygen -
periferal - peripheral -
permantly - permanently -
pinic - picnic - panic
prioities - priorities -
probabally - probably -
pronouncments - pronouncements -
proportionallity - proportionality -
rankd - ranked - ranks
redecoraton - redecoration -
regsirties - registries -
repititious - repetitious -
resing - resting - rising
sepaphore - semaphore -
simialirt - similarity -
sizable - sizeable -
strater - starter -
stetemets - statements -
subsructures - substructures -
successor - succesor - successor
synshronise - synchronise -
tequniques - techniques -
thre - there - three
thses - theses -
tollerance - tolerance -
uncritiacl - uncritical -
unscamble - unscramble -
versility - versatility -
```

6.3.2 Accuracy of System

There exist 384 misspelled words in test case, my system generates 93 wrong answers. We can calculate accuracy of system with the following function;

$$Accuracy = 1 - \#ofWrong/\#ofMisspelled \tag{4}$$

$$Accuracy = 1 - 94/384 \tag{5}$$

$$Accuracy = 0.7552 \tag{6}$$

7 Results

7.1 Analysis

According results of test misspelled words, there are some wrong corrections that my system produces. Those wrong can be divided into 3 as:

1. Misspelled word's edit distance is more than 1.
2. Correct version of misspelled word does not exist in corpus.
3. Misspelled word exists in corpus.

Due to the fact that we are expected to find errors with edit distance 1, first class of errors can be neglected.

Because of the fact that, corpus contains all words that we know, corrected versions of misspelled one must exist in corpus. However, some of the corrected versions do not exist in corpus, so my system can not produce output, or can not produce true output.

Since we assume all words in corpus are spelled correctly, third class of errors can be ignored, hence they did not misspelled.

Smoothed version of system is more accurate than non-smoothed version. Due to the fact that there exists a possibility of being overall probability 0 in non-smoothed one, it is too harsh.

7.2 Improvements

There are various ways to tokenize a file. My system implements tokenization by only replacing non-alpha characters with blank. In this case, tokenization is not done perfectly. We need to care about the words that contains apostrophe etc like ***shouldn't***.

Moreover, our dictionary is created according to unigram language model. In unigram language model, it is hard to estimate correct versions of misspelled words, since we have no idea about the words that before or after misspelled ones.

Also, context plays a significant role in spelling error correction. If there exists a way to understand what text means, we can produce more related outputs.