テキスト処理

金子泰之
情報理工学専攻　３年

2019 年 12 月 17 日

5.1

課題 5_1text1 から text9 までの語彙数

text1: 17231

text2: 6403

text3: 2628

text4: 9201

text5: 5441

text6: 1855

text7: 11387

text8: 882

text9: 6349

5.2

課題 5_2 アルファベットの出現率と出現回数

表 1: text1 から text9 までのアルファベットの出現回数

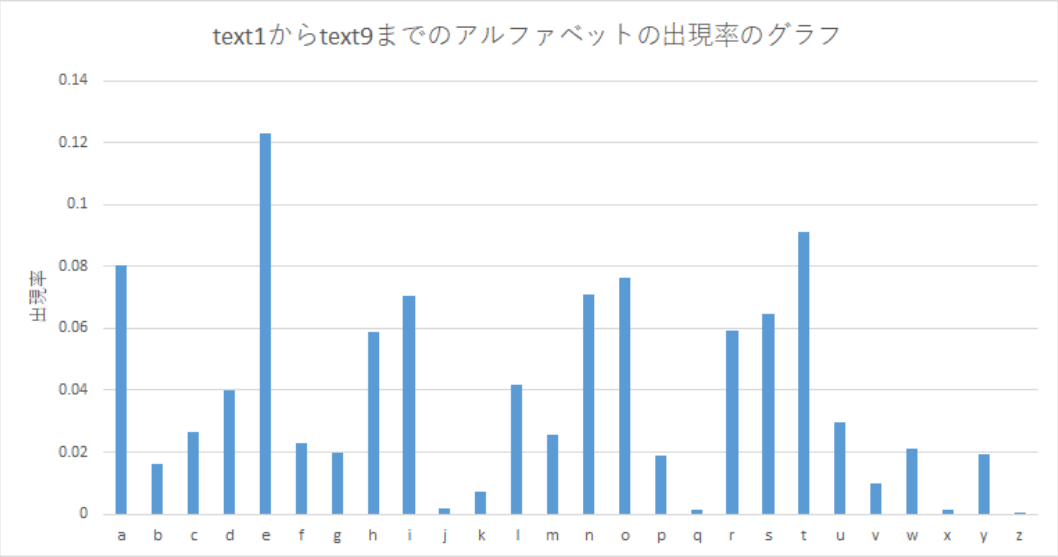
text	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
text1	77916	16877	22507	38219	117092	20833	20820	62896	65434	1082	8059	42793	23277	65617	69326	17255	1556	52134	64231	87996	26697	8598	22222	1030	16872	632
text2	40446	7938	12443	22323	66609	12227	9552	32264	36524	948	2780	20628	14617	38444	42015	7946	604	33246	32772	44995	14717	5849	12655	840	11678	69
text3	15425	2660	2585	9139	19119	3612	2303	13214	8305	488	929	5441	3952	11151	10193	1842	17	7628	8613	13515	3556	1353	3085	68	2681	108
text4	46389	9259	20500	23429	83353	17428	10583	31626	47942	949	2003	24269	15159	48231	51636	13941	630	40258	39572	62681	19507	7608	11817	1311	10563	626
text5	11125	1912	3134	3755	13229	1853	2876	7853	10233	1310	2027	6819	4253	9037	12970	3219	60	6568	7255	10709	7405	1041	3427	206	4540	225
text6	4076	859	1281	1658	5092	701	1430	3252	2970	64	631	2202	1097	2994	3643	621	104	3217	2513	3873	1992	428	1032	30	1062	33
text7	33159	6289	14298	15173	46871	8482	8046	16286	29858	848	3052	16349	10727	28884	28980	9195	428	27336	28382	36465	11569	3983	5703	1156	6652	314
text8	1172	195	336	638	1798	395	512	470	1351	33	285	898	542	1232	1212	335	18	948	1275	1015	473	185	213	19	412	4
text9	21003	3792	6216	10831	30758	5475	5036	15767	17072	182	2267	11019	6500	16239	18374	4250	347	13949	16488	22398	7147	2091	5731	311	5769	97

表 2 :text1 から text9 までのアルファベットの出現回数

```
text1
  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   |
0.082 |0.018 |0.024 | 0.04 |0.123 |0.022 |0.022 |0.066 |0.069 |0.001 |0.008 |0.045 |0.024 |0.069 |0.073 |0.018 |0.002 |0.055 |0.067 |0.092 |0.028 |0.009 |0.023 |0.001 |0.018 |0.001 |
text2
  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   |
0.077 |0.015 |0.024 |0.043 |0.127 |0.023 |0.018 |0.061 | 0.07 |0.002 |0.005 |0.039 |0.028 |0.073 | 0.08 |0.015 |0.001 |0.063 |0.062 |0.086 |0.028 |0.011 |0.024 |0.002 |0.022 | 0.0  |
0.999
text3
  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   |
0.102 |0.018 |0.017 |0.061 |0.127 |0.024 |0.015 |0.088 |0.055 |0.003 |0.006 |0.036 |0.026 |0.074 |0.068 |0.012 | 0.0  |0.051 |0.057 | 0.09 |0.024 |0.009 | 0.02 | 0.0  |0.018 |0.001 |
text4
  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   |
0.072 |0.014 |0.032 |0.037 | 0.13 |0.027 |0.017 |0.049 |0.075 |0.001 |0.003 |0.038 |0.024 |0.075 |0.081 |0.022 |0.001 |0.063 |0.062 |0.098 | 0.03 |0.012 |0.018 |0.002 |0.016 |0.001 |
text5
  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   |
0.081 |0.014 |0.023 |0.027 |0.097 |0.014 |0.021 |0.057 |0.075 | 0.01 |0.015 | 0.05 |0.031 |0.066 |0.095 |0.023 | 0.0  |0.048 |0.053 |0.078 |0.054 |0.008 |0.025 |0.002 |0.033 |0.002 |
text6
  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   |
0.087 |0.018 |0.027 |0.035 |0.109 |0.015 |0.031 |0.069 |0.063 |0.001 |0.013 |0.047 |0.023 |0.064 |0.078 |0.013 |0.002 |0.069 |0.054 |0.083 |0.043 |0.009 |0.022 |0.001 |0.023 |0.001 |
text7
  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   |
0.083 |0.016 |0.036 |0.038 |0.118 |0.021 | 0.02 |0.041 |0.075 |0.002 |0.008 |0.041 |0.027 |0.072 |0.073 |0.023 |0.001 |0.069 |0.071 |0.092 |0.029 | 0.01 |0.014 |0.003 |0.017 |0.001 |
text8
  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   |
0.073 |0.012 |0.021 | 0.04 |0.113 |0.025 |0.032 |0.029 |0.085 |0.002 |0.018 |0.056 |0.034 |0.077 |0.076 |0.021 |0.001 |0.059 | 0.08 |0.064 | 0.03 |0.012 |0.013 |0.001 |0.026 | 0.0  |
text9
  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   |
0.084 |0.015 |0.025 |0.043 |0.123 |0.022 | 0.02 |0.063 |0.069 |0.001 |0.009 |0.044 |0.026 |0.065 |0.074 |0.017 |0.001 |0.056 |0.066 | 0.09 |0.029 |0.008 |0.023 |0.001 |0.023 | 0.0  |
```

a::0.08043838439839733
b:0.015971789086783657
c:0.026726060764731097
d:0.04015807197620129
e:0.12317762274737488
f:0.02278164070420764
g:0.01962199789014915
h:0.058915403194550325
i:0.07048525286126062
j:0.001894245651320197
k:0.0070690911984312155
l:0.041843450093813926
m:0.025707069540375924
n:0.07117185274165107
o:0.07647214714541287
p:0.018802569808599054
q:0.001207645770929746
r:0.059446715999188916
s:0.0645214591338318
t:0.09100560573509822
u:0.029858432088213326
v:0.009989707418615455
q:0.0211386136072546
x:0.0015949009371125844
y:0.01932393654020395
z:0.000676332966291154

図 1: text1 から text9 までのアルファベットの合計出現率



5.3

課題 5_3 text1 から text9 までの出現率上位 5 番目までの 4 文字以上の単語

```
text 1
     that      |   3085
     with      |   1722
     this      |   1394
     whale     |   1226
     from      |   1088
text 2
     that      |   1385
     with      |   992
     have      |   819
    elinor     |   685
     which     |   593
text 3
     unto      |   598
     that      |   520
     said      |   477
     with      |   293
     thou      |   284
```

```
text 4
     that       |   1793
     have       |   1011
     which      |   1006
     with       |   958
     will       |   914
text 5
     part       |   1022
     join       |   1021
    action      |   347
     that       |   284
     what       |   201
text 6
    arthur      |   261
     that       |   106
   launcelot    |   101
     what       |    92
    knight      |    84
text 7
     that       |   848
     *t*-1      |   806
     said       |   628
     with       |   398
     from       |   391
text 8
     lady       |    88
     seeks      |    72
     with       |    46
     male       |    42
    looking     |    34
text 9
     that       |   871
     with       |   561
     syme       |   519
     said       |   507
     they       |   318
```

出現回数 l 位は that であった。

5.4

課題 5_4 text1 から text9 までのストップワードの比率

```
text 1
  0.4137045230600531

text 2
  0.4714570266146805

text 3
  0.4503618979537128

text 4
  0.47647149141838624

text 5
  0.29022439457898247

text 6
  0.28042671067366065

text 7
  0.302812984226628

text 8
  0.17444010684199712
```

```
text 9
  0.4391660526201725
```

5.5

課題 5_5 wagahaiha_nekodearu.txt に含まれる名詞の比率
形態素解析を行いすべての品詞をすべて足したものを分母とし、分子は名詞の数とした。

分母:212662
分子:60263
割合:0.283374556808

5.6

課題 5_6 wagahaiha_nekodearu.txt に含まれる出現頻度上位 30 位までの名詞

1	位	の	1612	回			
2	位	事	1207	回			
3	位	もの	981	回			
4	位	君	973	回			
5	位	主人	933	回			
6	位	ん	707	回			
7	位	よう	697	回			
8	位	人	604	回			
9	位	一	561	回			
10	位	何	540	回			
11	位	吾輩	481	回			
12	位	これ	414	回			
13	位	それ	395	回			
14	位	時	344	回			
15	位	迷亭	343	回			
16	位	傍点	318	回			
17	位	ところ	315	回			
18	位	方	314	回			
19	位	三	311	回	——	311	回
20	位	二	303	回			
21	位	上	297	回			
22	位	寒月	286	回			
23	位	そう	284	回			
24	位	顔	283	回			
25	位	先生	274	回			
26	位	人間	272	回			
27	位	僕	268	回			
28	位	さん	261	回			
29	位	気	251	回			

5.7

課題 5_7 wagahaiha_nekodearu.txt に top30_jnoun が含まれる頻度

```
単語の総数:  60263
トップ３０の名詞の数:  15140
 top 30 の出現率の名詞が全体の単語に占める割合:  0.251232099298
```