Homework Project #4

Yunxing Lu

ylu446@gatech.edu

***Abstract*—**In this homework assignment, I have utilized the Word2Vec model to assess word similarities and analogies, evaluating the model's performance and the correlation of similarity scores. Additionally, the "I01 [noun - plural_reg].txt" file was selected from the provided dataset to analyze and examine distributions across age, gender, and race. Findings reveal significant underrepresentation of elderly and minority groups, highlighting the need for data balancing to prevent biased algorithm outcomes.

1 THE APPLICATION OF WORD2VEC SYSTEM

1.1 Rank results between the 15 selected words with the targeted words - man and woman

Given the selected words provided by the homework instructions, the similarity results between the target and selected words are shown in Table 1 and Table 2.

Table 1 - Similariy results between target words and selected words

Word	Similarity to man
man	1.0
woman	0.587694
child	0.333422
doctor	0.289247
wife	0.283479

king	0.264497
husband	0.234116
nurse	0.153481
birth	0.123439
scientist	0.112269
scientist	0.110419
professor	0.107622
teacher	0.098740
president	0.094579
engineer	0.087364

Table 2 - Correlation Coefficients of Each Protected Class and Their Strength

Word	Similarity to man
woman	1.0
child	0.587694
man	0.587694
husband	0.449643
birth	0.420309
wife	0.300689
nurse	0.254358
queen	0.228572
teacher	0.204078
doctor	0.196134

scientist	0.137311
king	0.122529
professor	0.105199
president	0.084627
engineer	0.044264

2 THE PRACTICE WITH BATS

The file I selected is named 'I01 [noun - plural_reg].txt' under the folder '1_Inflectional_morphology.' This file contains the following words and their plurals: 'Target Plural form album application area car college council customer day death department development difference director event example fact friend god government hour idea language law member month night office period player population problem product resource river road role science solution song street student system thing town user version village website week year.'

According to the discussions on Piazza given by the TA, I selected the first word of each row as the target word, and the similarity between the target word and the other words in the row is shown in Table 3. Then, I selected three words: 'Hawaiian,' 'Asian,' and 'American' from the race category to compute the similarity between the target word and each of the three words above. The results are shown in Table 4.

The noticeable differences I have observed from Table 4 are, for example, for the target word 'customer,' only 'Hawaiian' has a positive value, whereas the other two both have negative values. Similarly, for the target word 'street,' only 'Hawaiian' has a negative value, whereas the other two both have positive values. Another example is the target word 'development,' where 'Asian' has the highest score of 0.31 among the values, while 'American' has a score of 0.13, and 'Hawaiian' even has a negative score of -0.02.

Table 3 - Word similarity in each row within the file "I01 [noun - plural_reg].txt"

Word	Similarity to man
album	0.806836
application	0.687333
area	0.577292
car	0.687196
college	0.566934
council	0.668369
customer	0.672182
day	0.380304
death	0.368868
department	0.408221
development	0.428566
difference	0.656968
example	0.373018
fact	0.243712
friend	0.650437
god	0.545838
government	0.627761
hour	0.595015
idea	0.493462
language	0.793717
law	0.716712

member	0.656933
month	0.635300
night	0.488998
office	0.468929
period	0.642558
player	0.724513
population	0.448160
problem	0.640801
product	0.425838
resource	0.532047
river	0.734379
road	0.581594
role	0.558104
science	0.478589
solution	0.656865
song	0.723103
street	0.544748
student	0.688544
system	0.689320
thing	0.574539
town	0.533612
user	0.598270
version	0.671480

village	0.397475
website	0.449784
week	0.622991
year	0.474169

Table 3 - Word similarity in each row within the file "I01 [noun - plural_reg].txt"

Word	hawaiian	asian	american
album	-0.086506	0.011041	0.048847
application	-0.093452	-0.010450	-0.085342
area	0.037764	0.205240	0.088694
car	-0.026932	0.055382	0.008751
college	0.107527	0.055432	0.107527
council	0.006199	0.091531	0.020328
customer	0.005862	-0.046081	-0.107423
day	-0.005895	0.013080 0	0.005658
death	-0.056239	-0.013778	0.017204
department	0.064260	0.022601	0.050891
development	-0.020309	0.310679	0.135233
difference	-0.070238	0.008447	-0.018690
director	-0.101055	0.103038	0.254937
event	-0.029367	0.127034	0.060735
example	-0.023405	0.042405	0.148859
fact	0.021395	0.049319	-0.003892

friend	0.025331	-0.107984	-0.053945
god	-0.027932	-0.067055	-0.077925
government	0.099934	0.165486	0.025794
hour	-0.118002	0.008062	0.028810
idea	-0.092626	0.020504	-0.001258
language	0.386572	0.200506	0.112349
law	0.062031	0.003038	0.059859
member	0.070393	0.284183	0.176842
month	-0.045229	0.038896	-0.091339
night	-0.104372	-0.069018	0.000971
office	-0.035546	0.035003	-0.044742
period	-0.072450	0.127523	0.005719
player	-0.096201	0.065644	0.127904
population	0.125615	0.213949	0.051187
problem	-0.036307	0.023480	-0.057099
product	-0.104131	0.051126	-0.057305
resource	0.087321	0.087657	0.039271
river	0.121428	0.107039	-0.031567
road	-0.011037	0.108642	0.038840
role	-0.086633	0.067789	0.086696
science	0.016395	0.102789	0.179021
solution	-0.157857	-0.069958	-0.095943
song	-0.035707	-0.052423	0.025049

street	0.158370	0.158370	0.119915
student	0.012774	0.047290	0.027426
system	0.005327	0.045575	0.062304
thing	-0.081733	-0.073591	-0.051719
town	0.082324	0.027401	-0.057419
user	-0.036455	-0.045575	-0.186814
version	-0.002018	0.030945	0.033264
village	0.182826	0.131560	0.095665
ebsite	0.038097	0.103888	0.105828
week	-0.076408	0.040031	-0.043074
year	0.016383	0.064849	-0.016960

3 SENTENCE COMPLETION PRACTICE AND EVALUATION

3.1 Complete the sentences with my own words

king is to throne as judge is to _code__?

giant is to dwarf as genius is to __dump___?

college is to dean as jail is to _judge__?

arc is to circle as line is to _sqaure__?

French is to France as Dutch is to _German__?

man is to woman as king is to _queen__?

water is to ice as liquid is to __rock_?

bad is to good as sad is to _happy__?

nurse is to hospital as teacher is to _school__?

usa is to pizza as japan is to _sushi__?

human is to house as dog is to _mud__?

grass is to green as sky is to _blue__?

video is to cassette as computer is to _hardrive___?

universe is to planet as house is to _land___?

poverty is to wealth as sickness is to _poor__?

3.2 Similarity calculation results from my own words and the words selection from Word2Vec

In this section, I will present the similarity scores from my word selection from the previous subsection. Additionally, the results of using the Word2Vec model to find the word analogy and the corresponding similarity scores are also presented in the last two columns of Table 4.

Table 4 - Word similarity in each row within the file 'I01 [noun - plural_reg].txt' based on my own word selection and the word analogy selection from the Word2Vec model

Word from sentence	Word of my choice	Similarity Score	Word from Model	Similarity Score
judge	code	0.107	prosecution	0.518
genius	dump	0.029	theorist	0.428
jail	judge	0.255	peress	0.544
line	sqaure	0.125	lines	0.428
Dutch	Netherlands	0.419	netherlands	0.604
king	queen	0.568	queen	0.553

liquid	rock	0.157	solid	0.450
sad	happy	0.448	glory	0.440
teacher	school	0.532	institution	0.482
japan	sushi	0.011	dishes	0.576
dog	mud	0.156	hound	0.423
sky	blue	0.443	blue	0.547
computer	hardrive	0.358	peripherals	0.665
house	land	-0.002	houses	0.426
sickness	poor	0.282	impious	0.496

3.3 Correlation strength calcualtions

I computed and printed the correlation between the vector of similarity scores from my analogies and the Word2Vec analogy-generated similarity scores. Based on the provided categories, I found a correlation of 0.34 and a p-value of 0.245. The correlation is weak, according to the defined category.

3 PRACTICES WITH UTK DATA

Based on my calculations, the frequency of images associated with each subgroup for age is shown in Figures 1 through 3. From these figures, we can answer the following questions from the homework:

- Age group with the largest representation: 0-20.
- Age group with the least representation: 81-116.
- Gender with the largest representation: 1.
- Gender with the least representation: 0.
- Race with the largest representation: 0.
- Race with the least representation: 1.

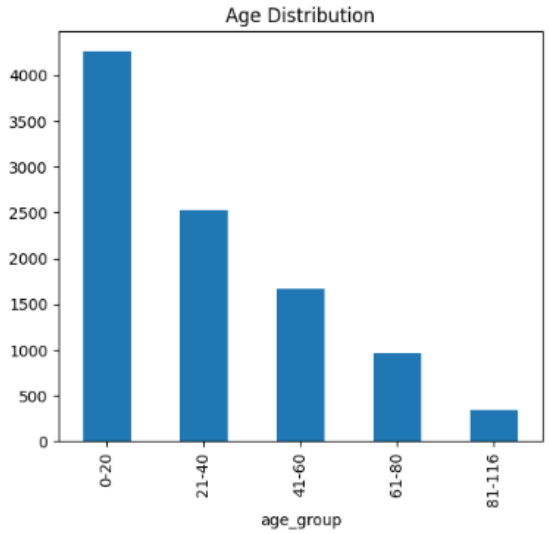

Figure 1. Age gourp distribution

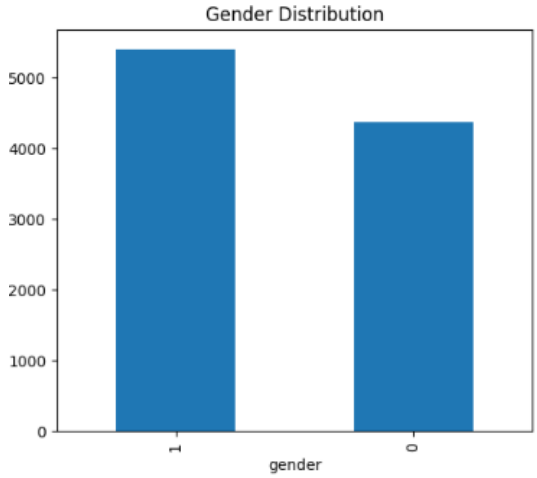

Figure 2. Gneder gourp distribution

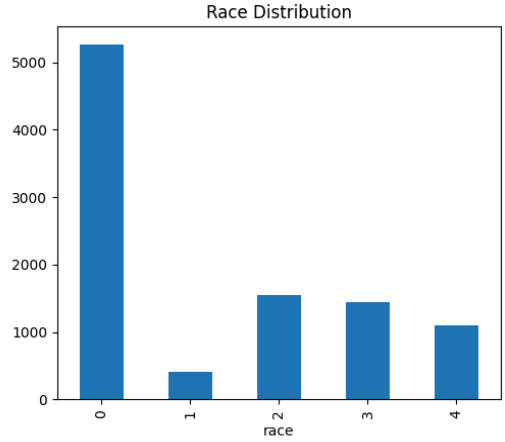

Figure 3. Race group distribution

Inspired by the one discussed in the lecture, we could recreable a table of the age group, gender and race distrbution as follows in Table 5:

Table 5 - Cross-tabulation of Age Group, Gender, and Race:

Race Age	Group Gender	0	1	2	3	4
0-20	0	895	68	521	266	191
	1	1036	92	496	341	361
21-40	0	411	45	110	150	185
	1	623	55	239	448	267
41-60	0	673	26	58	74	73
	1	579	39	30	88	15
61-80	0	388	45	28	35	6
	1	405	10	19	28	3
81-116	0	84	7	16	7	0
	1	1171	8	36	15	2

The elderly, especially those aged 81-116, are not well represented in the dataset, with only 346 people compared to 4,267 in the 0-20 age group. This problem is even worse for elderly minorities, which is important for healthcare and medical purposes. Black individuals (Race = 1) are also underrepresented, with only 405 people compared to 5,265 for Race = 0, especially in older age groups. This creates issues for elderly Black people and other minorities like elderly Asian, Indian, and other women, as well as middle-aged and older Black people, who are not well represented. Older age groups (61-80, 81-116) across all minority races are also lacking. These groups might face less accurate predictions, more errors, and possibly unfair results, which would hurt how well the model works for their specific needs. To fix this before training, it would be a good idea to consider ways to balance the data or add more examples from these groups