AI/ML Part 2 Assignment Report

Juejing Han
jhan446@gatech.edu

1 AI/ML PART 2 ASSIGNMENT REPORT

1.1 Word2vec System (Task Set 1)

1.1.1 *Similarity Score in Descending Order (Q1)*

Table 1 — Similarity Scores for Target Words “man” and “woman”.

Target Word “man”		**Target Word “woman”**	
Word	Similarity Score	Word	Similarity Score
man	1.0	woman	1.0
woman	0.588	child	0.590
child	0.333	man	0.588
doctor	0.289	husband	0.450
wife	0.283	birth	0.420
king	0.264	wife	0.301
husband	0.234	nurse	0.254
nurse	0.153	queen	0.229
birth	0.123	teacher	0.204
scientist	0.112	doctor	0.196
queen	0.110	scientist	0.137
professor	0.108	king	0.123
teacher	0.099	professor	0.105
president	0.095	president	0.085
engineer	0.087	engineer	0.044

In this section, 15 words are examined with two target words ("man" and "woman") using a pre-trained Word2Vec model. Table 1 displays the similarity scores **in descending order (from the most similar word to the least similar word)**.

1.1.2 *Bigger Analogy Test Set (Q2)*

File "I01 [noun - plural_reg].txt" is selected for the study in this section. In this dataset, the first column represents the singular form of a noun, and the second column denotes its plural form. **The singular noun is selected as the target word**.

Table 2 shows the similarity between the target word (in the first column) and the other word (in the second column) within the same row from the selected dataset (Q2a).

Table 2 — Similarity Scores of Words from Selected Dataset.

Target Word (Singular)	**Word (Plural)**	**Similarity Score**
album	albums	0.807
application	applications	0.687
area	areas	0.577
car	cars	0.687
college	colleges	0.567
council	councils	0.668
customer	customers	0.672
day	days	0.38
death	deaths	0.369
department	departments	0.408
development	developments	0.429
difference	differences	0.657
director	directors	0.539

Target Word (Singular)	Word (Plural)	Similarity Score
event	events	0.533
example	examples	0.373
fact	facts	0.244
friend	friends	0.65
god	gods	0.546
government	governments	0.628
hour	hours	0.595
idea	ideas	0.493
language	languages	0.794
law	laws	0.717
member	members	0.657
month	months	0.635
night	nights	0.489
office	offices	0.469
period	periods	0.643
player	players	0.725
population	populations	0.448
problem	problems	0.641
product	products	0.426
resource	resources	0.532
river	rivers	0.734

Target Word (Singular)	Word (Plural)	Similarity Score
road	roads	0.582
role	roles	0.558
science	sciences	0.479
solution	solutions	0.657
song	songs	0.723
street	streets	0.545
student	students	0.689
system	systems	0.689
thing	things	0.575
town	towns	0.534
user	users	0.598
version	versions	0.671
village	villages	0.397
website	websites	0.45
week	weeks	0.623
year	years	0.474

“Christian”, “Buddhist”, and “Muslim” are chosen as the three words from the protected class Religion. Table 3 demonstrates the similarity scores between the target word and these three religious identifiers (Q2b).

A noticeable difference is observed when negative and positive similarity scores exist among the three chosen words. In Table 3, **similarity scores that indicate a noticeable difference are highlighted in red**.

Table 3 — Similarity Scores of Target Word and Protected Class.

Target Word	**Christian**	**Buddhist**	**Muslim**
album	0.017	-0.053	-0.072
application	-0.003	-0.01	0.007
area	0.06	0.021	0.181
car	-0.061	0.016	-0.046
college	0.135	0.065	0.028
council	0.173	0.046	0.094
customer	-0.084	-0.045	-0.143
day	0.195	0.119	0.095
death	0.128	0.098	0.085
department	-0.096	-0.028	0.011
development	0.07	0.104	0.058
difference	0.018	0.003	0.046
director	-0.007	-0.074	-0.027
event	0.074	0.046	0.02
example	0.04	-0.076	0.058
fact	0.069	0.006	0.039
friend	-0.025	0.085	-0.033
god	0.322	0.27	0.158
government	-0.009	0.023	0.123
hour	0.022	-0.001	-0.009

Target Word	Christian	Buddhist	Muslim
idea	0.179	0.172	0.053
language	0.063	0.059	0.129
law	0.041	0.153	0.144
member	0.132	0	0.101
month	0.047	0.127	0.044
night	0.059	0.006	-0.057
office	-0.008	-0.083	-0.001
period	0.085	0.17	0.169
player	-0.068	-0.033	-0.061
population	0.126	0.04	0.351
problem	-0.037	-0.089	-0.063
product	-0.042	-0.071	-0.097
resource	0.019	0.032	-0.03
river	-0.006	0.018	-0.032
road	0.006	-0.012	-0.023
role	0.083	0.066	0.075
science	0.14	0.046	0.01
solution	0.001	-0.125	-0.092
song	0.073	0.025	-0.037
street	-0.011	-0.012	-0.14
student	0.147	0.12	0.105

Target Word	Christian	Buddhist	Muslim
system	-0.04	-0.047	-0.035
thing	0.073	0.062	0.042
town	0.082	0.051	0.052
user	-0.069	-0.053	-0.089
version	0.108	0.032	-0.032
village	0.119	0.161	0.085
website	0.143	0.08	0.123
week	0.061	0.099	0.031
year	0.04	0.116	0.086

1.1.3 *Human-based Word Analogy (Q3a)*

In this section, 15 analogy sentences are tested with human estimation (Table 4).

Table 4 — Human-based Word Analogy.

Original Sentence	Word Picked by Human	Similarity Score
king is to throne as judge is to ____?	bench	0.303
giant is to dwarf as genius is to ____?	idiot	0.344
college is to dean as jail is to ____?	warden	0.278
arc is to circle as line is to ____?	triangle	0.256
French is to France as Dutch is to ____?	netherlands	0.419
man is to woman as king is to ____?	queen	0.569
water is to ice as liquid is to ____?	solid	0.655
bad is to good as sad is to ____?	happy	0.449

Original Sentence	Word Picked by Human	Similarity Score
nurse is to hospital as teacher is to ____?	school	0.533
usa is to pizza as japan is to ____?	sushi	0.012
human is to house as dog is to ____?	kennel	0.284
grass is to green as sky is to ____?	blue	0.444
video is to cassette as computer is to ____?	cpu	0.450
universe is to planet as house is to ____?	room	0.250
poverty is to wealth as sickness is to ____?	health	0.195

1.1.4 *Model-based Word Analogy (Q3b)*

In this section, the same analogy sentences in Section 1.1.3 are tested with the Word2Vec model (Table 5).

Table 5 — Model-based Word Analogy.

Original Sentence	Word Picked by Model	Similarity Score
king is to throne as judge is to ____?	prosecution	0.519
giant is to dwarf as genius is to ____?	theorist	0.428
college is to dean as jail is to ____?	peress	0.544
arc is to circle as line is to ____?	lines	0.429
French is to France as Dutch is to ____?	netherlands	0.604
man is to woman as king is to ____?	queen	0.553
water is to ice as liquid is to ____?	solid	0.450
bad is to good as sad is to ____?	glory	0.440
nurse is to hospital as teacher is to ____?	institution	0.483
usa is to pizza as japan is to ____?	dishes	0.576

Original Sentence	Word Picked by Model	Similarity Score
human is to house as dog is to ____?	hound	0.423
grass is to green as sky is to ____?	blue	0.548
video is to cassette as computer is to ____?	peripherals	0.665
universe is to planet as house is to ____?	houses	0.426
poverty is to wealth as sickness is to ____?	impious	0.496

1.1.5 *Correlation between Human-based and Model-based Similarities (Q3c)*

Calculate the correlation between the human-based similarity scores (from Section 1.1.3) and the model-based similarity scores (from Section 1.1.4). The result is **0.028**. According to Evans's study (1996), this correlation strength is classified as **very weak**.

1.2 UTK Dataset (Task Set 2)

Data cleaning: Remove 2 images with missing values representing age, gender, or race.

Number of entries (images) of original dataset: 9,780

Number of entries (images) after data cleaning: 9,778

Table 6 — Frequency of Images from UTK Dataset.

Age Group		0-20	21-40	41-60	61-80	81-116	Total
Gender	Male	1,941	901	914	502	114	4,372
	Female	2,326	1,632	751	465	232	5,406
Race	White	1,931	1,034	1,252	793	255	5,265
	Black	160	100	75	55	15	405
	Asian	1,017	349	88	47	52	1,553
	Indian	607	598	162	63	22	1,452
	Others	552	452	88	9	2	1,103
Total		4,267	2,533	1,665	967	346	9,778

Table 6 displays the distribution of images corresponding to each subgroup within the categories of age, gender, and race.

1) For age, subgroup (0-20) has the largest representation, and subgroup (81-116) has the least representation.

2) For gender, subgroup "female" has the largest representation, and subgroup "male" has the least representation.

3) For race, subgroup "White" has the largest representation, and subgroup "Black" has the least representation.

4) If an algorithm is trained based on this dataset, subgroups "age (81-116)" and "Black" will be impacted most due to their disproportionately small sample sizes. Theses subgroups make up only 3.5% to 4% of the population, leading to their underrepresentation, which could result in inaccurate outcomes from the algorithm. It is essential to ensure representative samples to mitigate bias.

2 REFERENCES

1. Evans, J.D. (1996). Straightforward Statistics for the Behavioral Sciences. Thomson Brooks/Cole Publishing Co.