Marcus J. Anderson

Task 1 - Question 1:

Rank	Categories	Target: Man - Similarity
1	Man	1.0000000
2	Woman	0.5876938
3	Child	0.33342198
4	Doctor	0.28924733
5	Wife	0.2834791
6	King	0.26449707
7	Husband	0.2341164
8	Nurse	0.15348099
9	Birth	0.12343917
10	Scientist	0.112269185
11	Queen	0.11041949
12	Professor	0.10762214
13	Teacher	0.09874005
14	President	0.09457928
15	Engineer	0.08736356

Rank	Categories	Target: Woman - Similarity
1	Woman	0.99999994
2	Child	0.5898086
3	Man	0.5876938
4	Husband	0.4496431
5	Birth	0.42030877
6	Wife	0.3006885
7	Nurse	0.25435838
8	Queen	0.22857244
9	Teacher	0.2040782
10	Doctor	0.19613354
11	Scientist	0.13731062
12	King	0.122528546
13	Professor	0.105198584
14	President	0.08462686
15	Engineer	0.044264376

Task 1 - Question 2A:

File path: *.\BATS_3.0\3_Encyclopedic_semantics\E05 [name - occupation]*

Target	Comparable Words	Similarities
andersen	writer/poet/author	0.44000086/0.47731832/0.43809497
aristotle	philosopher	0.47283885

balzac	novelist/writer	0.46924716/0.35855865
beethoven	composer	0.4549362
caesar	emperor/commander/leader	0.41094264/0.19608694/0.17489773
columbus	explorer	0.05722825
confucius	philosopher	0.27792928
dante	poet	0.3467124
darwin	naturalist/biologist/geologist	0.23077369/0.24940835/0.13494161
depp	actor/producer/musician	0.34732425/0.25316972/0.41825163
descartes	mathematician/philosopher	0.43866587/0.475962
dickens	novelist/writer/critic/author	0.22047377/0.1609583/0.2721855/0.18532011
edison	inventor/businessman	0.37050515/0.38618717
einstein	physicist/scientist	0.2677085/0.15051119
euler	mathematician/physicist/astronomer/logician/engineer	0.34382334/0.2268403/0.24956332/0.19567546/0.025778608
goethe	poet/playwright/novelist/writer/author	0.44742927/0.45079568/0.43308097/0.3253155/0.35410023
hawking	physicist/scientist	0.19972192/0.20476863
haydn	composer	0.44596532
hegel	philosopher	0.42656144
hitler	dictator/politician/nazi	0.35221964/0.08640241/0.6447126
hume	philosopher/politician	0.46019194/0.20357871
jolie	actress/filmmaker/director/humanitarian/activist	0.34859166/0.33165818/0.21478063/0.13035092/0.14838143
kant	philosopher	0.48908925
kepler	mathematician/physicist/astronomer/astrologer	0.3492771/0.2649766/0.3729087/0.3244883
lincoln	president	0.28677016
locke	philosopher	0.39266244
marx	philosopher/communist	0.37862042/0.39729604
maxwell	physicist/scientist	0.21211179/0.087509245
mencius	philosopher	0.25081363
michelangelo	sculptor/painter/architect/artist/poet/engineer	0.48655382/0.3384021/0.29149815/0.25512904/0.36025062/0.08379826
moses	prophet/leader	0.66229564/0.12382772
mozart	composer	0.46777067
napoleon	emperor/leader/politician/commander	0.49515596/0.1972515/0.15574002/0.30917433
newton	scientist/mathematician/psysicist/philosopher	0.16116776/0.30103683/NA/0.1281153
pacino	actor/director/filmmaker	0.39157745/0.35383892/0.55067736
pascal	mathematician/philosopher	0.27585015/0.19397451

picasso	painter/artist/sculptor/designer	0.5138691/0.3944299/0.59064513/0.22270757
plato	philosopher	0.45162803
raphael	painter/artist/architect	0.4156319/0.31178018/0.415896
rembrandt	painter/etcher/artist	0.46058512/NA/0.33832243
rousseau	writer/author/philosopher	0.29637927/0.3557737/0.39546445
schwarzenegger	actor/politician/governor	0.22532345/0.22924289/0.39775336
shakespeare	playwright/poet	0.30283284/0.28986806
spinoza	philosopher	0.5268511
stalin	dictator/politician/leader/statesman	0.36515763/0.10883279/0.34110874/0.11084464
strauss	composer	0.31038108
tolstoi	novelist/writer/philosopher	NA/NA/NA
truman	president	0.40689662
wagner	composer	0.49797028
wittgenstein	philosopher	0.3858583

Task 1 - Question 2B:

Target	Comparable Words: Black/White/Asain - Similarities
andersen	0.00037875105/0.0013222431/NA
aristotle	-0.108022325/-0.118745275/NA
balzac	-0.022973424/-0.03589454/NA
beethoven	0.015509715/0.00564903/NA
caesar	-0.023882786/-0.0066797934/NA
columbus	-0.068338536/-0.09856424/NA
confucius	-0.084193766/-0.084193766/NA
dante	-0.046136446/-0.05466225/NA
darwin	0.033535462/-0.0013632694/NA
depp	0.095245674/0.11256091/NA
descartes	-0.12266282/-0.16258872/NA
dickens	-0.02573648/-0.009747116/NA
edison	-0.057153538/-0.0024713709/NA
einstein	-0.0048499377/-0.07277176/NA
euler	-0.13285758/-0.09272886/NA
goethe	-0.027708381/-0.045607492/NA
hawking	0.066887535/0.041638833/NA
haydn	-0.120149605/-0.12509155/NA
hegel	-0.034964185/-0.063780025/NA
hitler	-0.026845356/-0.03134216/NA

hume	-0.013768448/-0.008607665/NA
jolie	0.047783125/0.032714374/NA
kant	-0.112128474/-0.11160187/NA
kepler	-0.018130887/-0.020113958/NA
lincoln	-0.0011282974/0.07899269/NA
locke	-0.045440197/-0.023596/NA
marx	-0.06238781/-0.09504833/NA
maxwell	-0.071487375/-0.08767567/NA
mencius	-0.013193186/-0.045066204/NA
michelangelo	0.0079923915/-0.029830875/NA
moses	-0.043555934/-0.08109552/NA
mozart	-0.09074781/-0.068719685/NA
napoleon	-0.09329044/-0.050651006/NA
newton	0.031159354/-0.053674623/NA
pacino	0.062346604/0.060827672/NA
pascal	-0.13801496/-0.11996582/NA
picasso	0.026902018/0.011744485/NA
plato	-0.033506308/-0.07485206/NA
raphael	-0.055060294/-0.05971957/NA
rembrandt	-0.004071618/-0.030184172/NA
rousseau	-0.09389442/-0.1048965/NA
schwarzenegger	-0.058240555/-0.023354055/NA
shakespeare	-0.009434361/0.022738433/NA
spinoza	-0.064794704/-0.06671906/NA
stalin	0.049267914/-0.009978187/NA
strauss	-0.08073717/-0.055232607/NA
tolstoi	NA/NA/NA
truman	-0.076558486/-0.005981253/NA
wagner	-0.035856385/-0.06647922/NA
wittgenstein	0.004801147/-0.014892161/NA

Observations:

- The term **Asian** returned an error each time within the Word2Vec model.
- The scores between the **black** and **white** comparable terms are relatively close in size.
- A lot of the scores are coming back as negative, which says to me that the similarity between these list of words and the race protected class is extremely weak.
- The targeted word, **tolstoi**, was the only one to return an error for all three comparable terms.

Task 1 – Question 3A:

My Analogy	Similarity
Bench	0.30267337
Idiot	0.34426278
Warden	0.27777424
Square	0.19263238
Netherlands	0.41922888
Queen	0.5685571
Gas	0.5397003
Happy	0.44885093
School	0.5326567
Sushi	0.01186634
Kennel	0.28415978
Blue	0.4439698
Software	0.5089051
Suburb	0.1706722
Health	0.19527604

Task 1 – Question 3B:

Generated Analogy	Similarity
Prosecution	0.5186458230018616
Theorist	0.4280889630317688
Peress	0.5444425940513611
Lines	0.4287526607513428
Netherlands	0.6044681072235107
Queen	0.5532454252243042
Solid	0.4500039219856262
Glory	0.440381795167923
Institution	0.48289817571640015
Dishes	0.5763506293296814
Hound	0.4231664538383484
Blue	0.5478643178939819
Peripherals	0.6654507517814636
Houses	0.4264702796936035
Impious	0.49606096744537354

Task 1 – Question 3C:

Correlation: 0.185185594

Strength: Very Weak Correlation

Task 2 – Question 1:

Age Group	0 - 20	21 - 40	41 - 60	61 - 80	81 - 116	Total
Male (0)	1,941	901	914	502	114	**4,372**
Female (1)	2,326	1,632	751	467	232	**5,408**
White (0)	1,931	1,034	1,252	793	255	**5,265**
Black (1)	160	100	75	56	15	**406**
Asian (2)	1,017	349	88	47	52	**1,553**
Indian (3)	607	598	162	64	22	**1,453**
Others (4)	552	452	88	9	2	**1,103**
Total	**8,534**	**5,066**	**3,330**	**1,938**	**692**	**19,560**

Questions:

Which subgroup in each age, gender, and race category has the largest representation?

Age: The largest represented age group in the data is **0-20**, with a total count of **8,534** people

Gender: The largest represented gender in the data is **Female**, with a total count of **5,408** people

Race: The largest represented race in the data is **White**, with a total count of **5,265** people

Which subgroup in each age, gender, and race category has the least representation?

Age: The smallest represented age group in the data is **81-116**, with a total count of **692** people

Gender: The smallest represented gender in the data is **Male**, with a total count of **4,372** people

Race: The smallest represented race in the data is **Black**, with a total count of **406** people

Based on what you've learned so far, if an algorithm is trained based on this dataset, which group(s) will be impacted the most? Explain why.

Based on the representation displayed on the table, the most favorable representation that the algorithm would be trained with are white females between the ages of 0 and 20 years old. However, the least represented group would be black males between the ages of 81 and 116 years old. This means that running this algorithm with this specific dataset would provide more images of young white females when searching things such as **students**, **beautiful, smart**, **etc**. This would create an unfair imbalance when it comes to minority representation since the data is so lopsided.