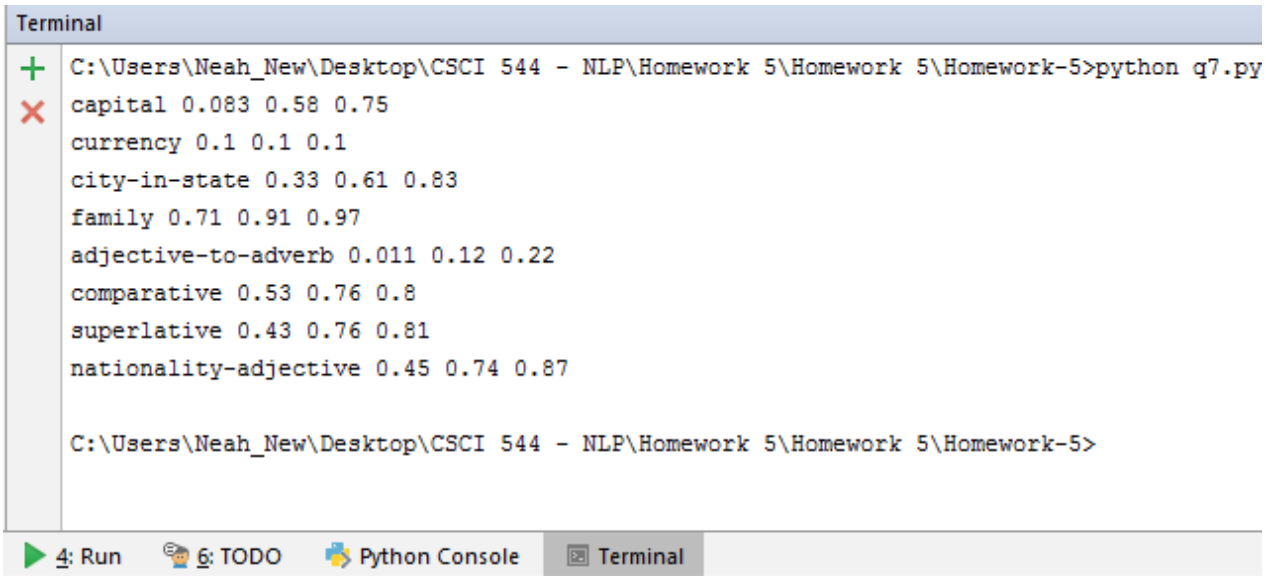
Q7

The table generated by q7.py is shown below:



The incorrectly predicted 1-best analogy items for each of the relation groups are as follows:

RELATION GROUP	INCORRECTLY PREDICTED ANALOGY	CORRECT ANSWER
Capital	baghdad, iraq, queens, england	baghdad, iraq , london, england
Currency	brazil, real, wins, won	brazil, real, korea, won
City-in-state	philadelphia, pennsylvania, houston, texas	philadelphia, pennsylvania, dallas, texas
Family	boy, girl, sons, sisters	boy, girl, brothers, sisters
Adjective-to-adverb	typical, typically, soon, quickly	typical, typically, quick, quickly
Comparative	bad, worse, big, larger	bad, worse, large, larger
Superlative	hot, hottest, low, highest	hot, hottest, high, highest
Nationality-adjective	china, chinese, japan, italian	china, chinese, italy, italian

Capital and Family are the 2 relation groups that have the highest prediction accuracy. Currency and Adjective-to-adverb relation groups have lowest prediction accuracies.

So, in general, the prediction accuracy for capital and family relation groups are high because these words approximate well and there is probably large amount of data in the training corpus which contains these contexts of words. For instance, there may be many sentences in the corpus where

London is associated with England (capital relation group), and mom is related to dad (family relation group).

However in the case of Currency and Adjective-to-adverb relation groups, there are probably many contexts that occur with the three other words in the analogy and hence there are many possibilities for the third word to be predicted.

For example, consider the following analogy: complete, completely, quick, quickly

The resulting nearest words got are as follows:
[('members', 159.75070501209726), ('p.m.', 135.73398414255348), ('article', 111.72377551913816), ('like', 94.08279681570305), ('percent', 92.43414423367739), ('songs', 91.1234020756762), ('game', 85.00566336068411), ('(', 84.39069000299675), ('feet', 82.30725346228324), ("'ve", 80.85326284497307)]

Here, the word 'quick' is not predicted as there many be many contexts in input where the word 'quick' does not occur with the other three words in the analogy.

Relation group: Capital

Consider the following three analogies and most similar words:

Analogy 1: `['baghdad', 'iraq', 'london', 'england']`
Most similar words: `[('queens', 0.55695019337795815), ('london',0.55403436996081534), ('bronx', 0.55287833986294621), ('n.j.', 0.54298544498308143), ('nearby', 0.5125112097440474), ('york', 0.50957178900636257), ('jersey',0.50722701214057553), ('orleans', 0.49895683169696953), ('orange', 0.49792495511619422), ('el', 0.4946545674812155)]`

Analogy 2: `['madrid', 'spain', 'london', 'england']`
Most similar words: `[('york', 0.67819096214036734), ('orleans', 0.67088342602128015), ('jersey', 0.63660524662632167), ('london', 0.50261500439359386), ('bronx', 0.42147987863337116), ('mexico', 0.41685804739237897), ('seattle', 0.38571932607125986), ('baghdad', 0.37211915207225715), ('jail', 0.36886563205191764), ('queens', 0.36616478683010906)]`

Analogy 3: `['paris', 'france', 'london', 'england']`
Most similar words: `[('london', 0.63642860937749601), ('york', 0.61144588297572311), ('orleans', 0.55311544563174775), ('seattle', 0.52609722300313932), ('chicago', 0.49697272541353527), ('jersey', 0.48837338315584461), ('n.y.', 0.47639082877628508), ('ma', 0.47113532266491731), ('bronx', 0.47009134885850712), ('boston', 0.45614124840114528)]`

In Analogy 1, given that the capital of Iraq is Baghdad, the capital of England is in the second position of the vectors returned.

In Analogy 2, given that the capital of Spain is Madrid, the capital of England, London, is in the fourth position of the vectors returned.

In Analogy 3, given that the capital of France is Paris, the capital of England, London, is in the first position of the vectors returned (1-best - best performance).

In all three analogies, the right capital of England was to be determined. However, the performance of all three analogies varied - correct answer was in top 5% in Analogy 1 and 2, and top 1% in Analogy3. Hence, this shows that the contexts chosen for the first analogy (Analogy 1: Baghdad Iraq,

Analogy 2: Madrid Spain, Analogy 3: Paris France) plays a very important role in how accurately the third item in the analogy is predicted.

In most cases, the correct context for the first pair of analogy items are established, hence, capital has a good 5-best and 10-best accuracy.

Relation group: Family

Analogies under the 'family' relation group seem to be predicted accurately the most number of times. I think this is because in most of the analogies, there is only one logical option for the missing third item in the analogy. For instance, consider the following analogies in the 'family' relation group:

son, daughter, dad, mom

Here, in the above example, the only word that would make sense in the third position is 'dad'.
So, the family relation group has the highest 1-best, 5-best accuracy, and 10 best accuracy.

Relation group: Currency

Consider the following relations:

Analogy 1: `['usa', 'dollar', 'brazil', 'real']`
Most similar words: `[('great', 0.37751334541438919), ('b', 0.34621914929085501), ('wonderful', 0.34168579675160415), ('sports', 0.32581078150226217), ('founder', 0.32495297492401626), ('lawrence', 0.32278974288252504), ('paul', 0.3218019092020567), ('frank', 0.31952550784608558), ('longtime', 0.31536071915610198), ('nyc', 0.3146733710342165)]`

Analogy 2: `['usa', 'dollar', 'korea', 'won']`
Most similar words: `[('joined', 0.63192014804931629), ('played',0.56987678533708963), ('founded', 0.56978042753552127), ('attended', 0.56145856824653884), ('entered', 0.54714927036318783), ('champions', 0.52720352449173147), ('wins', 0.51468683207303001), ('named', 0.47742252890676795), ('established', 0.46973464247892405), ('champion', 0.46881004264321507)]`

Here, in Analogy 1 and 2, the fourth word in the relation has additional word senses, apart from representing currency. For instance in Analogy 2, the word 'won' is taken as the verb form and other verbs that can be used in the same contexts is predicted.

Hence, I think, due to the presence of multiple word senses for currency words in the fourth position, a large number of incorrect predictions were returned as a result of considering the wrong word sense.

Relation group: Adjective-to-Adverb

In most cases, the set of adjectives present in the corpus is large. Hence, in many cases, a synonym, or a completely unrelated word is predicted. Hence, the accuracy of this relation group is low.