Q7:

The output table of q7.py as follow:

|  |  |  |  |
| --- | --- | --- | --- |
| Word | Top-1 | Top-5 | Tom[-10 |
| capital | 0.083 | 0.583 | 0.75 |
| currency | 0.1 | 0.1 | 0.1 |
| city-in-state | 0.333 | 0.611 | 0.833 |
| family | 0.705 | 0.91 | 0.968 |
| adjective-to-adverb | 0.11 | 0.122 | 0.222 |
| comparative | 0.533 | 0.762 | 0.8 |
| superlative | 0.429 | 0.762 | 0.81 |
| nationality-adjective | 0.453 | 0.744 | 0.872 |

The output’s screenshot as follow:

capital: 0.083 0.583 0.75  
currency: 0.1 0.1 0.1  
city-in-state: 0.333 0.611 0.833  
family: 0.705 0.91 0.968  
adjective-to-adverb: 0.011 0.122 0.222  
comparative: 0.533 0.762 0.8  
superlative: 0.429 0.762 0.81  
nationality-adjective: 0.453 0.744 0.872

According to the sheet, its obvious city-in-state and Family are the 2 relation groups that have the highest prediction accuracy. In the other hand, Currency and Adjective-to-adverb relation groups have lowest prediction accuracies.

In general, obvious city-in-state and Family relation groups have highest prediction accuracy due to these words appear at very similar positions and may have large amount of data in the corpus.

For explanation, I will show examples of each relation groups.

The sheet of analogies that can’t find a correct answer in the top-10 nearest words.

|  |  |  |
| --- | --- | --- |
| Relation group | Incorrect predict analogy | Correct answer |
| capital | ['london', 'england', 'paris', 'spain'] | ['london', 'england', 'madrid', 'spain'] |
| currency | ['brazil', 'real', 'wins', 'won'] | ['brazil', 'real', 'korea', 'won'] |
| city-in-state | ['austin', 'texas', 'ma', 'massachusetts'] | ['austin', 'texas', 'boston', 'massachusetts'] |
| family | ['boy', 'girl', 'wilson', 'princess'] | ['boy', 'girl', 'prince', 'princess'] |
| adjective-to-adverb | ['usual', 'usually', 'personality', 'rarely'] | ['usual', 'usually', 'rare', 'rarely'] |
| comparative | ['low', 'lower', 'strong', 'greater'] | ['low', 'lower', 'great', 'greater'] |
| superlative | ['big', 'biggest', 'bright', 'hottest'] | ['big', 'biggest', 'hot', 'hottest'] |
| nationality-adjective | ['china', 'chinese', 'language', 'english'] | ['china', 'chinese', 'england', 'english'] |

Relation group: capital

london england madrid spain

the top-10 nearest words as follow:  
[('paris', 0.6311241126440104), ('australia', 0.5399716351261978), ('argentina', 0.5338104591322531), ('japan', 0.5231119662965942), ('brazil', 0.522790896789085), ('new', 0.5175616375322507), ('canada', 0.5130099036788842), ('germany', 0.5076974357914728), ('france', 0.5000276485361321), ('italy', 0.49936367881202376)]

given that the capital of england is london, the capital of spain can not found in top-10 words, that’s may due to spain mentioned more with paris instead of madrid.

Relation group: currency

brazil real korea won

the top-10 nearest words as follow:

[('wins', 0.5465353150116258), ('scored', 0.5297751445161243), ('joined', 0.5177210290359301), ('races', 0.512965790396326), ('champions', 0.4975167753833556), ('championship', 0.4860926923754578), ('champion', 0.48308661170858663), ('argentina', 0.46890200320595127), ('earned', 0.46868863650066017), ('victory', 0.46163448453145534)]

given that the currency of brazil is real, the currency won belongs to which country can not found correctly. Due to won has another meaning of ‘win’, which may use more times.

Relation group: city-in-state

austin texas boston massachusetts

the top-10 nearest words as follow:

[('ma', 0.5462141983046472), ('n.y.', 0.5212968238154321), ('princeton', 0.5152417134854544), ('greenwich', 0.508978562122771), ('3rd', 0.5038258817332566), ('nyc', 0.49027822199673415), ('cemetery', 0.4795378346315674), ('n.j.', 0.4788772406193382), ('queens', 0.4783688588448897), ('7th', 0.47662621007900147)]

given that the city-in-state of texas is austin, the city-in-state of Massachusetts can not find boston. Due to Massachusetts maybe always be mentioned as ‘ma’for writing convenient.

Relation group: family

boy girl prince princess

the top-10 nearest words as follow:

[('wilson', 0.5469766804473892), ('davis', 0.5372234204065911), ('leo', 0.5180961078079441), ('sir', 0.5147582224501853), ('howard', 0.5124038982580791), ('lawrence', 0.5082068842132594), ('moore', 0.49831456440327326), ('max', 0.4920211460258172), ('carter', 0.48534295956364065), ('miller', 0.48402583752064127)]

given that the boy and girl is family members, the ‘princess’ can not pair similar family member ‘prince’. That may because of princess is not a family member and the corpus has limited amount of word ‘princess’ and ‘prince’.

Relation group: adjective-to-adverb

usual usually rare rarely

the top-10 nearest words as follow:

[('personality', 0.3816663998725839), ('well', 0.37483982504846114), ('celebrity', 0.3665383040802688), ('hilarious', 0.364114440055946), ('blunt', 0.3512041055361574), ('voice', 0.3419348032109326), ('silly', 0.3284552778623868), ('classical', 0.3256226668258607), ('loud', 0.321890020687553), ('legend', 0.31950685512206717)]

given that the adj usual pair to adv usually, rare pair to rarely in theory, however according to the return top-10, I can not find and the most possible word is ‘personality’, which is a synonym word to ‘rare’, and maybe due to the corpus has few cases of ‘rare’.

Relation group: comparative

low lower great greater

the top-10 nearest words as follow:

[('strong', 0.6407787498434203), ('weak', 0.5616044839284206), ('increasing', 0.5573452119698361), ('healthy', 0.5390430513504019), ('good', 0.5385680320281966), ('sharp', 0.5272617664609115), ('positive', 0.5175915845092691), ('lack', 0.5128573243867391), ('negative', 0.5114776825337802), ('significant', 0.5097743575779436)]

Given ‘lower’ is the comparative of ‘low’, the original word of ‘greater’ is ‘great’, however it can not find in the list maybe due to this corpus always use ‘strong’ instead of ‘great’ as this meaning.

Relation group: superlative

big biggest hot hottest

the top-10 nearest words as follow:

[('bright', 0.4808289879529178), ('dumb', 0.4765185823934632), ('loud', 0.4760615181446228), ('cool', 0.47027268447371384), ('cute', 0.46063480903273), ('bunch', 0.456240404425604), ('hungry', 0.45354754364926686), ('shorts', 0.44930804279174763), ('skinny', 0.44845333823145966), ('nice', 0.44416983118049586)]

In the top-10 words list can not found ‘hot’ as ‘biggest’ is superlative of ‘big’. Insteadly the word ‘bright’ is the most probable original word of ‘hottest’.

Relation group: nationality-adjective

china chinese england english

the top-10 nearest words as follow:

[('language', 0.5884056930637268), ('france', 0.5840742509581972), ('germany', 0.5568269012448792), ('britain', 0.5402279695373187), ('japan', 0.5391449356168462), ('europe', 0.5136610340534908), ('italy', 0.5083517602315544), ('russia', 0.4985433595545383), ('australia', 0.49781305863074177), ('india', 0.4617668640733066)]

In the nationality-adjective group, ‘english’ can not pair to ‘england’ in the top-10 words, the most probable word is ‘language’, that may due to ‘english’always be used as a language instead of its other meanings.