Q5

The 6 words chosen are: jack (a name), people (a common noun), beautiful (an adjective), walk (a verb), California, sister.

A list of the words, their 10 most similar words obtained using dense vectors from word2vec and comparison with q3 results (where in similar words using sparse vectors were determined) are in the table below:

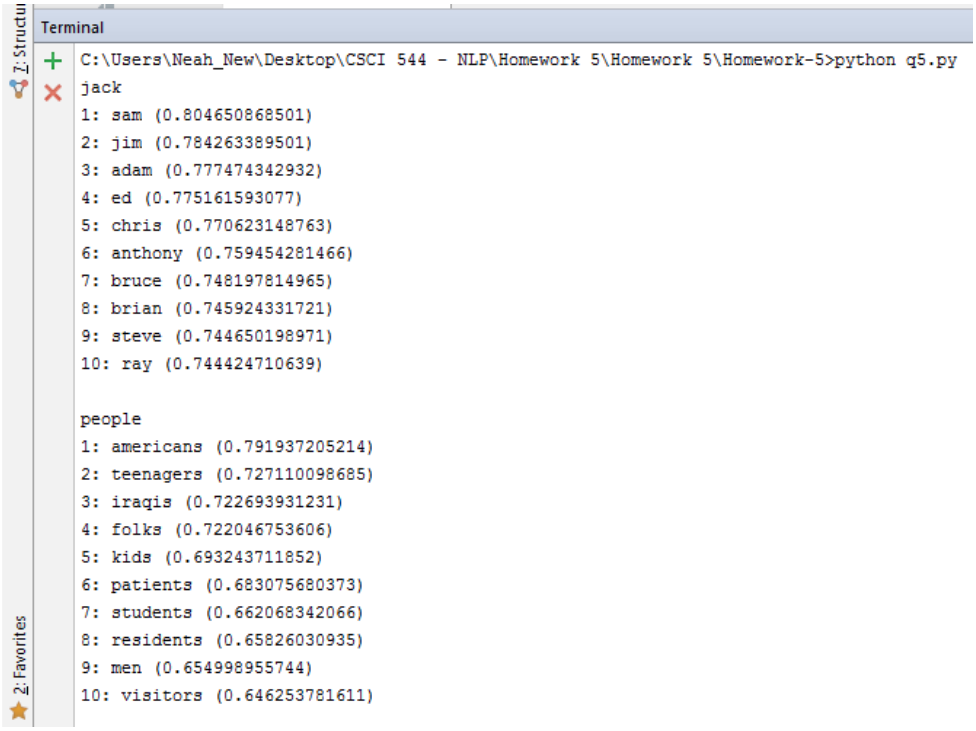
| WORD | 10 MOST SIMILAR WORDS | COMPARISION WITH Q3 |
|---|---|---|
| Jack | 1: sam (0.804650868501)<br>2: jim (0.784263389501)<br>3: adam (0.777474342932)<br>4: ed (0.775161593077)<br>5: chris (0.770623148763)<br>6: anthony (0.759454281466)<br>7: bruce (0.748197814965)<br>8: brian (0.745924331721)<br>9: steve (0.744650198971)<br>10: ray (0.744424710639) | The list for similar words for 'jack' includes a list of ten proper names (i.e. person names) present in the nytcounts.4k.<br>The list of similar words returned (i.e. sam, jim, adam and so on) may be used in similar contexts as 'jack' and hence they are similar.<br>Ex. Jack rode a bicycle, Sam rode a bicycle, and Jim rode a bicycle and so on. Hence, the output makes sense.<br>Hence, I think this approach of distributional similarity works in this case, and the similarity scores make good enough sense.<br><br>Q3 too returns a list of person names. I think both the approaches (sparse vectors in Q3 and dense vectors in Q5) work equally well. The similarity scores returned in Q5 is much lower than that of Q3 as it prevents overfitting. |
| People | 1: americans (0.791937205214)<br>2: teenagers (0.727110098685)<br>3: iraqis (0.722693931231)<br>4: folks (0.722046753606)<br>5: kids (0.693243711852)<br>6: patients (0.683075680373)<br>7: students (0.662068342066)<br>8: residents (0.65826030935)<br>9: men (0.654998955744)<br>10: visitors (0.646253781611) | The list for similar words for 'people' includes a list of ten plural common nouns/collective nouns present in the nytcounts.4k.<br>Since the term 'people' essentially refers to humans being considered collectively, the list of similar words returned as patients, folks, students, teenagers, and so on makes sense as they each refer to a group of people in a different way. In fact most of the similar words returned are hyponyms of 'people' ('people' is a hypernym of 'students', 'americans', 'residents' and so on). Also, it is possible for these words to be used in several similar contexts as 'people'. Ex: students visited the museum, folks visited the museum, and so on. Hence, the returned words are similar and the output makes sense.<br>However, I think 'folks', 'men', 'students', 'kids' and so on must be ranked higher than 'patients', 'americans', 'iraqis'.<br><br>I think dense vectors (Q5) works better for this case. Here, 'americans' is considered most similar to 'people', as compared to 'patients' in Q3. It |

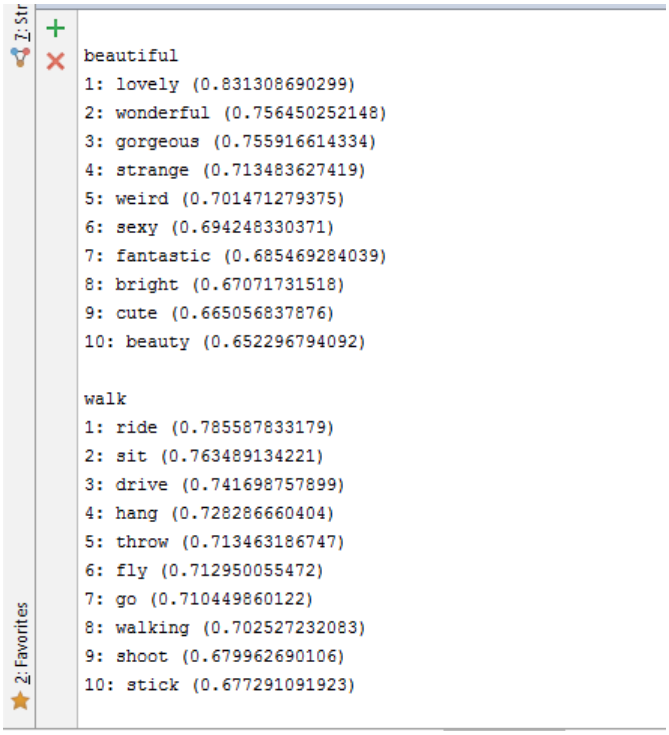
| | | |
|---|---|---|
| | | makes more sense for 'people' and 'americans' to be used in same contexts, than 'people' and 'patients'. Also, Q5 includes words like 'men' and 'residents', 'kids' which form a large subset of people. So, basically in Q5 the output includes words like – 'residents', 'men', 'kids', and so on which are more relatable to 'people'.<br>Hence, the nearest words returned by Q5 make more sense than Q3 and Q5 generalises better and it captures synonymy too. However, the scores of these words are much lower than that of Q3 to prevent overfitting. |
| Beautiful | 1: lovely (0.831308690299)<br>2: wonderful (0.756450252148)<br>3: gorgeous (0.755916614334)<br>4: strange (0.713483627419)<br>5: weird (0.701471279375)<br>6: sexy (0.694248330371)<br>7: fantastic (0.685469284039)<br>8: bright (0.67071731518)<br>9: cute (0.665056837876)<br>10: beauty (0.652296794092) | The list for similar words for 'beautiful' includes a list of ten adjectives present in the nytcounts.4k.<br>Among the ten similar words obtained, only words like 'lovely', and 'gorgeous' have a close meaning to 'beautiful'.<br>The other words like 'cute', 'wonderful', 'fantastic', and so on, can be used in similar contexts. Ex: A beautiful dress, a fantastic dress, a cute dress, a wonderful dress, and so on.<br>Hence, the output makes sense here.<br><br>Both Q3 and Q5 return similar adjectives. But Q5 captures synonymy better i.e. it returns words like 'lovely', 'wonderful', 'gorgeous', which are close synonyms to 'beautiful' a higher similarity score than for other words returned that do not have a close meaning but can only be used in the same context as 'beautiful'. |
| Walk | 1: ride (0.785587833179)<br>2: sit (0.763489134221)<br>3: drive (0.741698757899)<br>4: hang (0.728286660404)<br>5: throw (0.713463186747)<br>6: fly (0.712950055472)<br>7: go (0.710449860122)<br>8: walking (0.702527232083)<br>9: shoot (0.679962690106)<br>10: stick (0.677291091923) | The list for similar words for 'walk' includes a list of ten verbs present in the nytcounts.4k.<br>Each of the words obtained are verbs indicative of an action not similar to walking.<br>The words are not close in meaning to 'walk'.<br>In some cases, they do occur in similar contexts (Ex: walk the child to school, drive the child to school) and in some cases they don't (Ex: walk the dog, sit the dog, shoot the dog (WRONG – does not make sense)).<br>Hence, the output makes partial sense here.<br><br>I think Q3 performs better here – it outputs many more words that can be used in a similar context as that for 'walk'. |
| California | 1: connecticut (0.775855966702)<br>2: florida (0.732917165034) | The list for similar words for 'california' includes a list of ten places present in the nytcounts.4k. |

| | | |
|---|---|---|
| | 3: pennsylvania (0.725749564901)<br>4: texas (0.725168724457)<br>5: massachusetts (0.715188050659)<br>6: ohio (0.698119898535)<br>7: virginia (0.684975746318)<br>8: southern (0.587589066007)<br>9: westchester (0.551599496908)<br>10: state (0.543420680191) | It is possible for these words to be used in a similar context as 'california'.<br>Ex: California is a great place, Texas is a great place, and so on.<br>The output makes sense here.<br><br>Q3 performs better in this case. Q5 outputs 2 words – 'southern' and 'state', which cannot appreciably be used in the same context as 'california'. Hence, Q3 outputs more words that can be used in same context as 'california'<br>Ex: california is a great place, state/southern is a great place (does not make sense). |
| sister | 1: daughter (0.920502516502)<br>2: grandmother (0.898335572332)<br>3: brother (0.886538329098)<br>4: grandfather (0.877567077522)<br>5: wife (0.843730040311)<br>6: son (0.8416104942)<br>7: husband (0.8352816469)<br>8: cousin (0.820411963512)<br>9: uncle (0.798096014684)<br>10: mother (0.786693177584) | The list for similar words for 'sister' includes a list of nouns, most of them denoting relations, present in the nytcounts.4k.<br>Most of the words returned (except voice and vision) can be used in similar context as 'sister'. Hence, output makes almost complete sense.<br><br>Q5 gives words that are similar to 'sister' (Q3 returned some words not similar to 'sister'). However, I think the output of Q5 makes more sense and it captures similar words better. |

In general, dense vectors provide more sensible nearest words than sparse vectors in Q3 because they generalize better and help avoid overfitting. Synonyms are also captured pretty well using dense vectors.

A screenshot of the output of the show_nearest() for the above chosen words is as follows:


Terminal
C:\Users\Neah_New\Desktop\CSCI 544 - NLP\Homework 5\Homework 5\Homework-5>python q5.py
jack
1: sam (0.804650868501)
2: jim (0.784263389501)
3: adam (0.777474342932)
4: ed (0.775161593077)
5: chris (0.770623148763)
6: anthony (0.759454281466)
7: bruce (0.748197814965)
8: brian (0.745924331721)
9: steve (0.744650198971)
10: ray (0.744424710639)
people
1: americans (0.791937205214)
2: teenagers (0.727110098685)
3: iraqis (0.722693931231)
4: folks (0.722046753606)
5: kids (0.693243711852)
6: patients (0.683075680373)
7: students (0.662068342066)
8: residents (0.65826030935)
9: men (0.654998955744)
10: visitors (0.646253781611)
2: Favorites


beautiful
1: lovely (0.831308690299)
2: wonderful (0.756450252148)
3: gorgeous (0.755916614334)
4: strange (0.713483627419)
5: weird (0.701471279375)
6: sexy (0.694248330371)
7: fantastic (0.685469284039)
8: bright (0.67071731518)
9: cute (0.665056837876)
10: beauty (0.652296794092)
walk
1: ride (0.785587833179)
2: sit (0.763489134221)
3: drive (0.741698757899)
4: hang (0.728286660404)
5: throw (0.713463186747)
6: fly (0.712950055472)
7: go (0.710449860122)
8: walking (0.702527232083)
9: shoot (0.679962690106)
10: stick (0.677291091923)
2: Favorites

```
california
1: connecticut (0.775855966702)
2: florida (0.732917165034)
3: pennsylvania (0.725749564901)
4: texas (0.725168724457)
5: massachusetts (0.715188050659)
6: ohio (0.698119898535)
7: virginia (0.684975746318)
8: southern (0.587589066007)
9: westchester (0.551599496908)
10: state (0.543420680191)
```

```
sister
1: daughter (0.920502516502)
2: grandmother (0.898335572332)
3: brother (0.886538329098)
4: grandfather (0.877567077522)
5: wife (0.843730040311)
6: son (0.8416104942)
7: husband (0.8352816469)
8: cousin (0.820411963512)
9: uncle (0.798096014684)
10: mother (0.786693177584)

C:\Users\Neah_New\Desktop\CSCI 544 - NLP\Homework 5\Homework 5\Homework-5>
```