Q3

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 and observations are in the table below:

WORD	10 MOST SIMILAR WORDS	OBSERVATIONS
Jack	1: adam (0.879731849466) 2: james (0.859791406534) 3: susan (0.856596258738) 4: daniel (0.847600400729) 5: jonathan (0.847532173876) 6: peter (0.844088466095) 7: eric (0.843254808498) 8: elizabeth (0.843212500388) 9: andrew (0.837502134837) 10: max (0.837313367421)	The list for similar words for 'jack' includes a list of ten proper names (i.e. person names) present in the nytcounts.4k. The list of similar words returned (i.e. adam, james, susan and so on) may be used in similar contexts as 'jack' and hence they are similar. Ex. Jack rode a bicycle, Peter rode a bicycle, and James rode a bicycle and so on. Hence, the output makes sense. However, it does not make sense to claim that 'adam' is more similar to 'jack' than 'james' and so on, as suggested by the similarity scores as it would be possible to use all the similar words in the same context as 'jack'. On the contrary, it can be argued that though the relative similarity scores are varied for each of the 10 nearest proper names returned, their similarity scores are quite close. Hence, I think this approach of distributional similarity works in this case, and the similarity scores make good enough sense.
People	1: patients (0.867294111075) 2: folks (0.866750095106) 3: teenagers (0.855805609674) 4: voters (0.821146247577) 5: iraqis (0.815548130187) 6: americans (0.815415124846) 7: drivers (0.813110803837) 8: fans (0.806085684459) 9: investors (0.802723567553) 10: students (0.800886294386)	The list for similar words for 'people' includes a list of ten plural common nouns/collective nouns present in the nytcounts.4k. 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', 'voters', 'fans' 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. With respect to the relative scores, I think 'folks' is much more similar to 'people' than 'patients', as 'patients' essentially form an extremely small subset of 'people'. Hence, I think 'folks' must be the nearest term to 'people' and also the words 'teenagers', 'students', 'drivers' need to be ranked higher in the list of nearest words (i.e. have higher similarity scores to 'people') as they form larger

		subsets of 'people' than specific words like 'iraqis', 'americans' and so on. Here, the metric and the relative scores are partially right. The distributional similarity approach works fairly well here.
Beautiful	1: quiet (0.934589918762) 2: healthy (0.93423795718) 3: brilliant (0.932141976947) 4: simple (0.930179899943) 5: strange (0.929305413501) 6: handsome (0.928821115404) 7: lovely (0.928245206214) 8: gorgeous (0.925278064971) 9: lonely (0.91021444701) 10: wonderful (0.90310565038)	The list for similar words for 'beautiful' includes a list of ten adjectives present in the nytcounts.4k. Among the ten similar words obtained, only 'handsome', 'lovely', and 'gorgeous' have a close meaning to 'beautiful'. Most of the other words like quiet, healthy, brilliant, lonely, do not have a close meaning to 'beautiful', but they can be used in similar contexts. Since a majority of the nearest words can be used in the similar contexts as 'beautiful', the words are similar, and the output makes sense. Ex: A beautiful setting, a lonely setting, a simple setting, and so on. With respect to relevant scores, I think words 'handsome', 'lovely', 'gorgeous' must be given a higher similarity score to 'beautiful' as they are close in meaning to 'beautiful' and they can be used in similar contexts as that of 'beautiful' (Note: as opposed to other nearest words which can only be used in a similar context, but do not have a close meaning). Hence, the relative scores are satisfactory. Hence, the approach of distributional similarity seems to work fairly well in this case.
Walk	1: ride (0.880255328217) 2: break (0.871605646046) 3: fly (0.854025102725) 4: move (0.852437639954) 5: drive (0.851046742116) 6: pass (0.834080356904) 7: play (0.82916119478) 8: kick (0.822211201538) 9: block (0.819866843465) 10: throw (0.816818633604)	The list for similar words for 'walk' includes a list of ten verbs present in the nytcounts.4k. Each of the words obtained are verbs indicative of an action not similar to walking. The words are not close in meaning to 'walk'. In some cases, they do occur in similar contexts (Ex: walk by the house, ride by the house, drive by the house, pass by the house) and in some cases they don't (Ex: pass the ball, move the ball, walk the ball (WRONG – does not make sense)). Hence, the output only makes partial sense. With respect to relative scores, I think the words 'drive', 'move', 'pass' need to be given a higher similarity score as they can be used in similar contexts as that of 'walk'. Hence, the relative scores too make partial sense. Overall, distributional similarity works partially here.
California	1: massachusetts (0.929528816433)	The list for similar words for 'california' includes a list of ten places present in the nytcounts.4k.

	2: parliament (0.926764079682) 3: pennsylvania (0.923939932312) 4: connecticut (0.909955405571) 5: texas (0.903871934072) 6: chicago (0.899457046856) 7: france (0.883675708923) 8: florida (0.880151711863) 9: virginia (0.876764122872) 10: baghdad (0.872795401219)	It is possible for these words to be used in a similar context as 'california'. Hence, the nearest words are similar and the output makes sense. Ex: California is a great place, Texas is a great place, and so on. However, it does not make sense to claim that 'massachusetts' is more similar to 'california' than 'chicago' and so on, as suggested by the similarity scores. On the contrary, it can be argued that though the relative similarity scores are varied for each of the 10 nearest words returned, their similarity scores are quite close. Hence, I think this approach of distributional similarity works in this case, however the similarity scores make good enough sense. Hence, distributional similarity works fairly well here.
sister	1: brother (0.969081228735) 2: cousin (0.931375316603) 3: daughter (0.910062035017) 4: son (0.909114720457) 5: uncle (0.882357541539) 6: mother (0.876661780427) 7: father (0.857117419436) 8: voice (0.795489770211) 9: vision (0.792894496318) 10: wife (0.792880292691)	The list for similar words for 'sister' includes a list of nouns, most of them denoting relations, present in the nytcounts.4k. Most of the words returned (except voice and vision) can be used in similar context as 'sister'. Hence, output makes almost complete sense. With respect to relative score, I think 'wife' should be scored higher than 'voice' and 'vision'. Hence, the scores too are mostly right. Hence, I think distributional similarity works fairly well here.

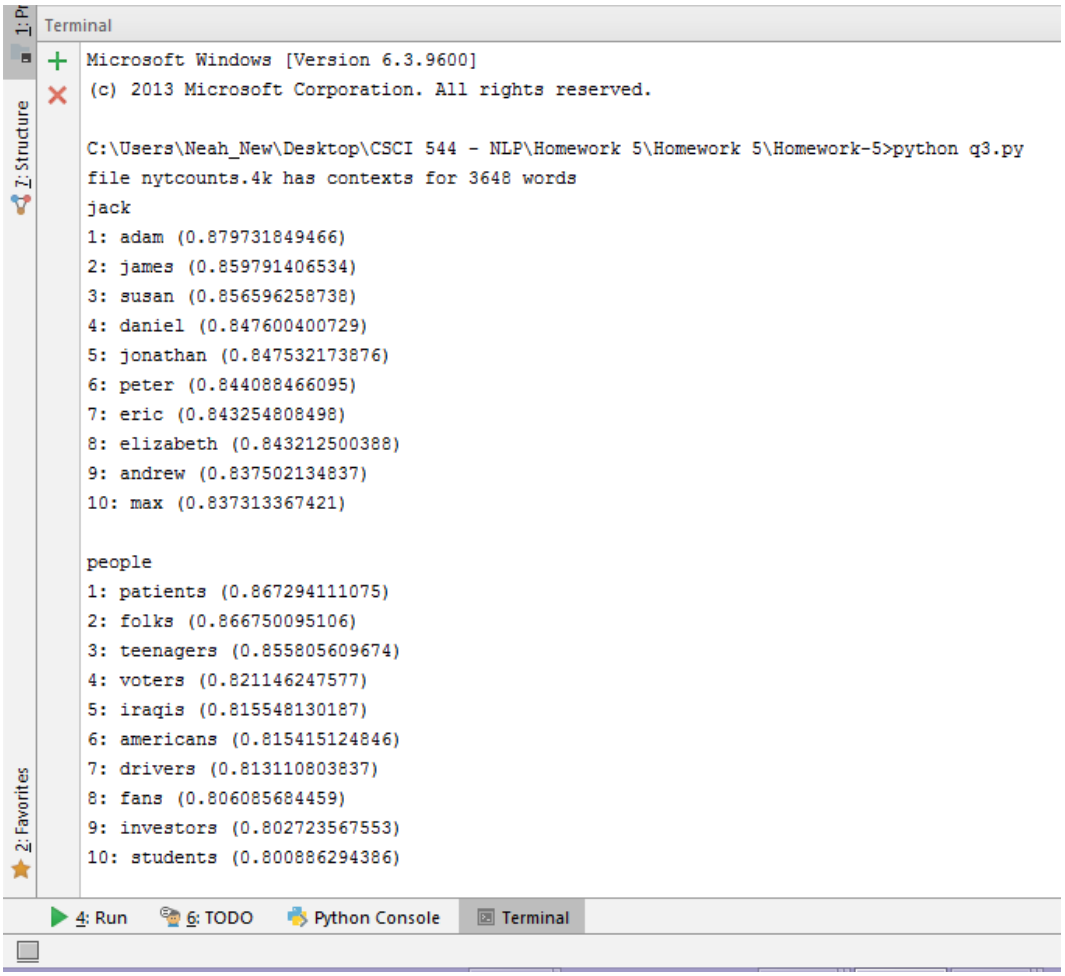
For further analysis, the first nearest word of 'jack' ('adam') was further explored. Then, the first nearest word of 'adam' .i.e. 'susan' was explored. The results are as follows:

WORD	jack	adam	Susan
10 NEAREST WORDS	1: adam (0.879731849466) 2: james (0.859791406534) 3: susan (0.856596258738) 4: daniel (0.847600400729) 5: jonathan (0.847532173876) 6: peter (0.844088466095) 7: eric (0.843254808498) 8: elizabeth (0.843212500388) 9: andrew (0.837502134837)	1: susan (0.906303287797) 2: jonathan (0.903141469464) 3: eric (0.898352454499) 4: daniel (0.890523863894) 5: peter (0.888142865848) 6: andrew (0.887256313909) 7: james (0.883557711366) 8: elizabeth (0.882771436888) 9: david (0.880152191565) 10: jack (0.879731849466)	1: andrew (0.913343234757) 2: eric (0.913104358468) 3: adam (0.906303287797) 4: david (0.904444410231) 5: jonathan (0.904390542417) 6: peter (0.899325777846) 7: robert (0.89618325357) 8: daniel (0.896132117851) 9: steven (0.888358172996) 10: nancy (0.883563171603)

	10: max (0.837313367421)		

It can be observed that the similarity between the pairs will always be the same irrespective of the exploration order .i.e. similarity(jack, adam) = similarity(adam, jack), similarity(adam, susan) = similarity(susan, adam). Also, the set of 10 nearest words for every word will always include the parent of the word (for instance, adam contains jack, susan contains adam).

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



```
beautiful
1: quiet (0.934589918762)
2: healthy (0.93423795718)
3: brilliant (0.932141976947)
4: simple (0.930179899943)
5: strange (0.929305413501)
6: handsome (0.928821115404)
7: lovely (0.928245206214)
8: gorgeous (0.925278064971)
9: lonely (0.91021444701)
10: wonderful (0.90310565038)

walk
1: ride (0.880255328217)
2: break (0.871605646046)
3: fly (0.854025102725)
4: move (0.852437639954)
5: drive (0.851046742116)
6: pass (0.834080356904)
7: play (0.82916119478)
8: kick (0.822211201538)
9: block (0.819866843465)
10: throw (0.816818633604)
```

```
california
1: massachusetts (0.929528816433)
2: parliament (0.926764079682)
3: pennsylvania (0.923939932312)
4: connecticut (0.909955405571)
5: texas (0.903871934072)
6: chicago (0.899457046856)
7: france (0.883675708923)
8: florida (0.880151711863)
9: virginia (0.876764122872)
10: baghdad (0.872795401219)
```

```
sister
1: brother (0.969081228735)
2: cousin (0.931375316603)
3: daughter (0.910062035017)
4: son (0.909114720457)
5: uncle (0.882357541539)
6: mother (0.876661780427)
7: father (0.857117419436)
8: voice (0.795489770211)
9: vision (0.792894496318)
10: wife (0.792880292691)
```