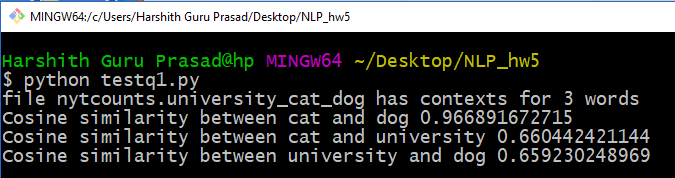
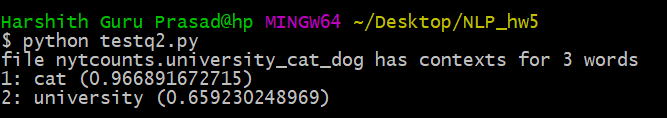
**Answer 1**

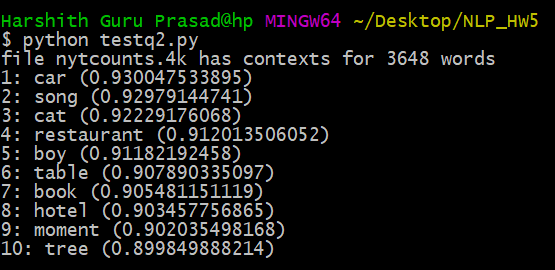


Since cat and dog are used frequently in a similar context (animals) and have a similar meaning in natural language, they have a high cosine similarity of 0.967. However, cat and university have a low cosine similarity of 0.66 since they are semantically quite different. Similarly, dog and university have a lower cosine similarity of 0.659 as they are almost never used in the same context. Hence, the relative similarity scores among the 3 words confirms the intuitive sense of similarity.

**Answer 2**

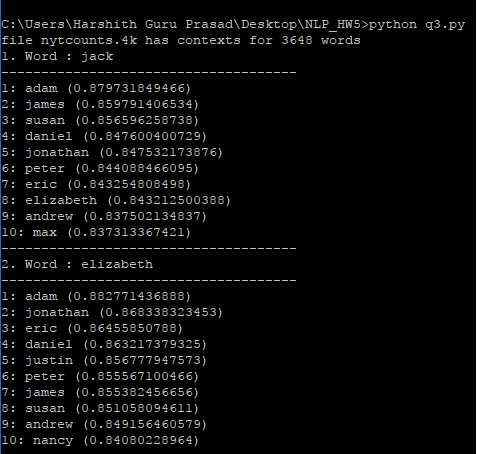


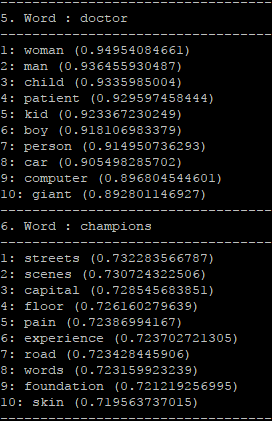
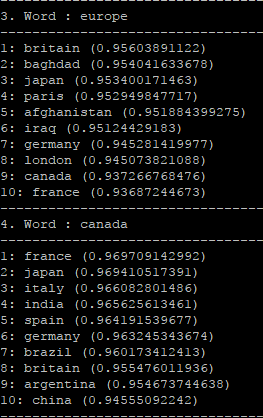
Found the 10 most similar words to the given query word(dog) ranked in decreasing order of cosine similarities generated from the contexts for words in the file ‘nytcounts.4k’. The word ‘cat’ was previously expected to be the most similar to ‘dog’ as per the context vectors in the smaller nytcounts file, but the results indicate that ‘car’ is more similar to ‘dog’. This is because of different context vectors for dog and cat present in nytcounts.4k file.

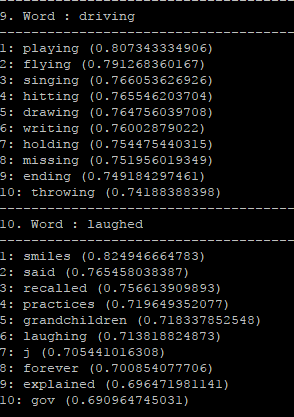
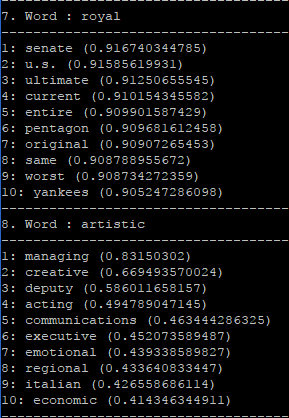


**Answer 3**

**Top 10 results for query words using sparse context count vectors**







**Analysis:**

For the first two **proper nouns** – **jack** and **elizabeth**, the most similar words returned makes sense as they are all proper nouns of **names**. The word ‘elizabeth’ appears in the top ten results of the word ‘jack’ but the converse is not true as the top ten results for Elizabeth have higher cosine similarity scores than jack. This is on account of the relative positions of the two words while traversing the semantics graph.

For the **locations** such as **europe** and **canada**, the most similar words are all locations (names of countries) and hence the similarity metric used makes sense. Canada appears in the list of top ten most similar words for Europe. However, Europe does not feature in Canada’s top ten list of semantically similar words as canada has more semantically similar words(countries) than Europe(continent).

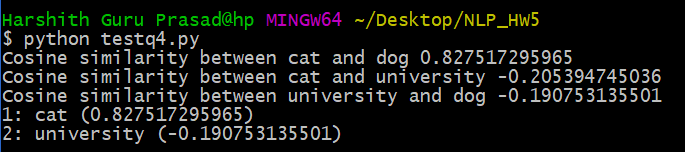
However, for the **common noun doctor**, the words returned are generic and not completely relevant to the semantic meaning of a doctor. The most relevant result returned for doctor is patient. Similarly, for **champions**, the results returned have less similarity with respect to the query word and most of the results are more irrelevant.

For **adjectives** as well, like **‘royal’** and **‘artistic’**, the output makes less sense on account of context ambiguity and hence makes less sense. Adjectives did not provide relevant or semantically meaningful results. ‘Royal’ yields results from a different context than what was expected.

**Verbs** like **‘driving’** are similar to other verbs in the same tense. The verb **‘laughed’** also fetches other verbs or nouns that it is commonly used with and makes comparatively more sense than adjectives.

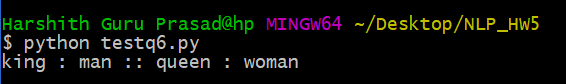
Overall, proper nouns have good results while common nouns have a more diverse context. Verbs are semantically similar to other verbs especially the ones in the same tense. Adjectives yield results that are not entirely relevant to the context of the word.

**Answer 4.**



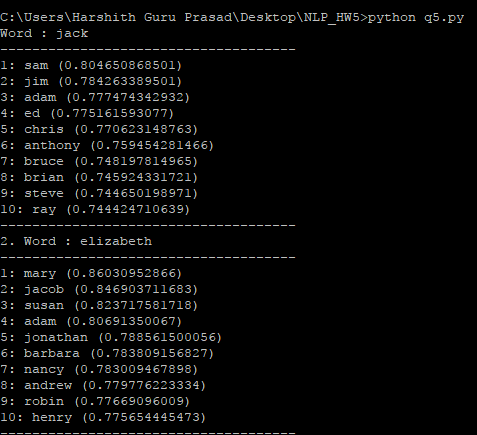
Similarity values were generated using dense numpy arrays of word vectors from the file ‘nyt\_word2vec.university\_cat\_dog’

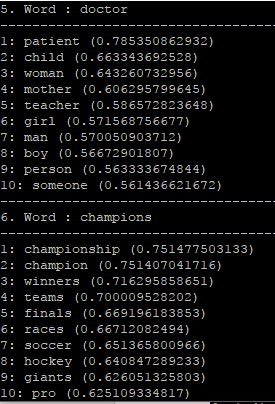
**Answer 6.**

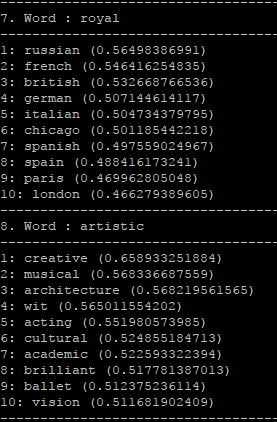
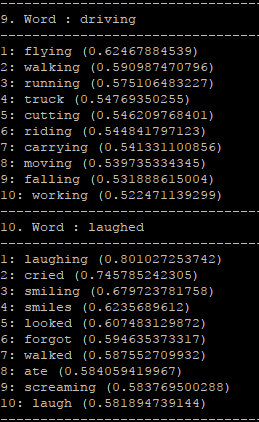


**Answer 5**

**Top 10 results for query words using Dense word embeddings Vectors**





**Analysis**

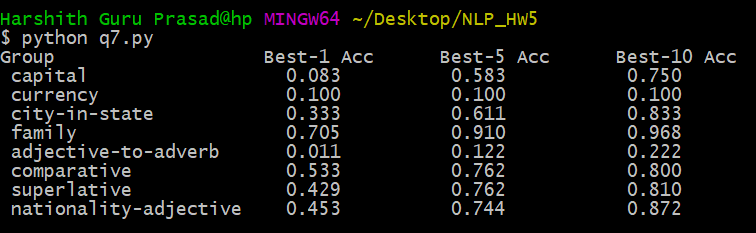
**Proper Nouns:** People names – The semantically similar results are more meaningful in the case of dense word embedding vectors as they are all names of males, while sparse context count vectors return names of both genders. Similarly for elizabeth, the top sematic result is mary(female) as per dense word embedding but adam in the case of sparse context count vectors. Locations – Both sparse context count and dense word embedding vectors return relevant results for names of locations(countries). For proper nouns likes names of people, both vector models provide good semantic results as names of people are used in the same context of an person entity. Different countries can be associated with different contexts.

**Common Nouns**: doctor – the dense word embedding vectors returns patient as the most similar word for the common noun doctor and both dense and context count vectors return similar results for the word doctor. However, for the common noun champions, dense word embedding technique provides far more relevant and semantically similar results than the context count vectors, which returns vague and highly irrelevant results. Common nouns are used more frequently, and the context count vectors yield poor results as a result of word frequencies with respect to various contexts. Dense vectors yield better results as they are language modeling and feature learning techniques where words or phrases from the vocabulary are mapped to vectors of real numbers.

**Adjectives:** For the adjective ‘royal’, context count vectors provide poor semantic results while dense word embedding vectors provide nationalities as semantically similar words. Similarly, for the adjective ‘artistic’ dense word embeddings provide relevant adjectives that are semantically more similar to ‘artistic’ than the results generated from the context count vectors. Both methods return adjectives however; the relevance between the results returned differs. Dense vectors perform better as they are prediction based embeddings and not frequency based like the context count vectors. They predict words that are most similar in context and relevance.

**Verbs:** The verb‘driving’has more relevant results generated from the dense word embedding vectors and are mostly verbs concerned with motion. The results are mostly verbs in the same tense. Context count vectors provide verbs that are semantically less similar and in the same tense as the query word. The verb laughed has more semantically similar and relevant words generated from the dense word embeddings as compared to context count vectors. Since context count vectors use context frequency to determine similarity, verbs which frequently used tend to be more similar but less relevant. However, dense vectors are prediction based and consider the weight of the context and its relevance and do not rely entirely on context frequency which can provide deceptive results based on the words with greatest frequencies in a text .

**Answer 7**



From the table of best-n accuracies, it is evident that some group of relations are predicted more accurately, while some others are predicted less accurately using the dense vectors for the cosine similarity comparison metric.

**Best 1 Accuracies:**

family > comparative > nationality-adjective > superlative > city-in-state > currency > capital > adj-to-adverb

**Best 5 Accuracies:**

family > nationality-adjective > comparative = superlative > city-in-state > capital > adj-to-adverb > currency

**Best 10 Accuracies:**

family > nationality-adjective > city-in-state > superlative > comparative > capital > adj-to-adverb > currency

**Conclusions**

**currency** has a poor accuracy of 10%. Only 1 of the analogies were correctly predicted while the other analogies failed to produce semantically meaningful words that satisfied the semantics of the analogy for best-1,5,10 results.

Currencies like ‘won’,’real’ tend to have different semantic meanings while dollar is used only in a financial context. The words ‘won’ is mostly used as a verb while ‘real’ is used as adjectives or adverbs. Hence currencies has poor prediction accuracies based on the given vectors.

**adjective to adverb** performs poorly as well with a top 10 vector accuracy of 22.22%. Adjectives and adverbs can be used in diverse contexts and hence don’t provide good prediction accuracies. Need larger datasets from which better context scores or vector embeddings can be derived and used to compute similarity metrics.

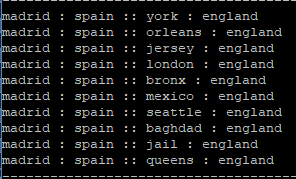
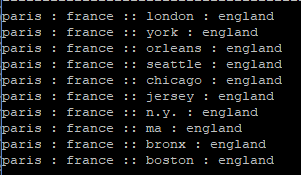
**family** performs extremely well and predicts the missing word in the analogy with a best-10 accuracy of 96.795. Family tops the accuracy charts for best-1,5,10 (Family comes first always). Excellent predictions are on account of frequently used pronouns and common nouns that are often used in the same context as each other. Family relations are close in the semantic graph and hence have a high prediction accuracy.

**city-in-state, comparative, superlative, nationality-adjective** perform well as well and produce a best-10 prediction accuracy of 80-87%. Comparative and superlative forms of English words have essentially the same stem(lemma). They are often used in the same context with a varying degree of intensity. Hence, they have good prediction accuracies. City and state are all locations and hence have good accuracies on account of their limited context. Similarly, nationality and adjective have a common stem(lemma) and used in a similar context. Hence their relevance to each other in most contexts accounts for good accuracies.

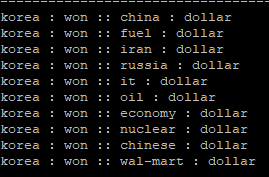
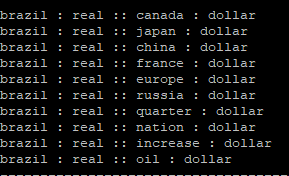
Word predictions ultimately depends on the context scores / word embeddings of each word relative to the other words in a huge corpus. The prediction accuracy depends on the representation of the various contexts in which a word has a semantically related meaning.

**Examples of Incorrect ( Best-1 ) and Correct Predictions for Each Relation Group**

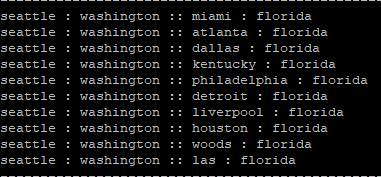
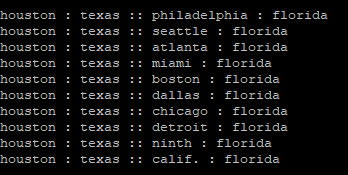
**1. capital**

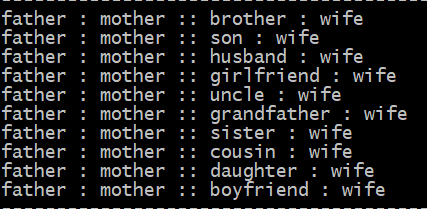
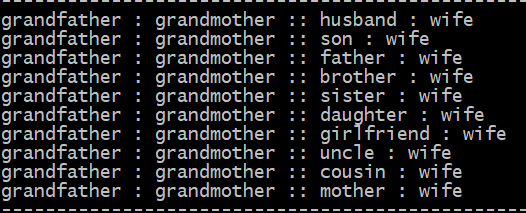
**2. currency**

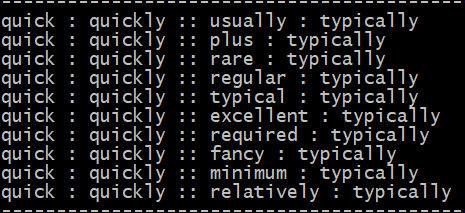
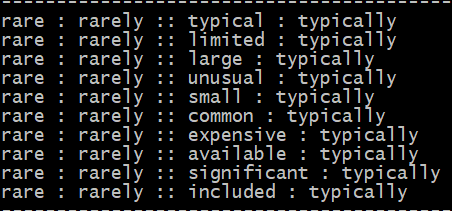
**3. city-in-state**



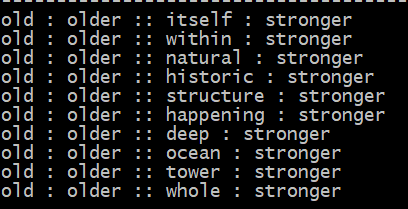
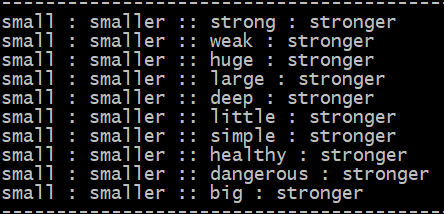
**4. family**

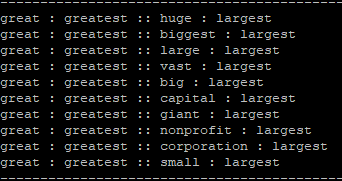
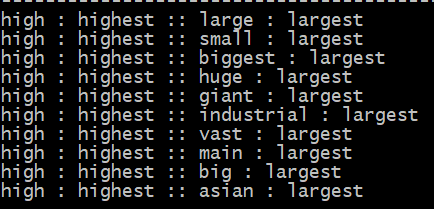
**5. adjective – adverb**

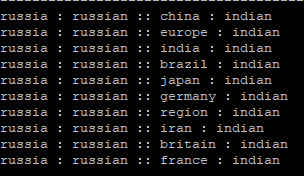
**6.comparative**

**7. superlative**

**8. nationality-adjective**

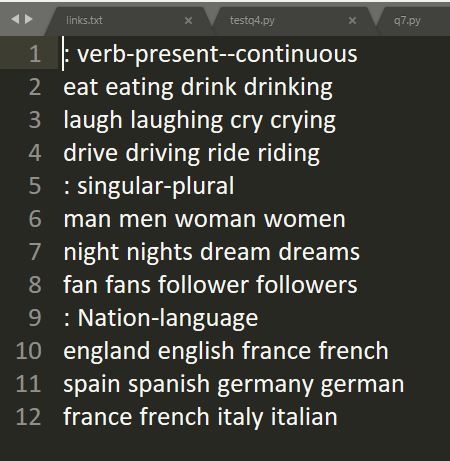
 

**Answer 8**

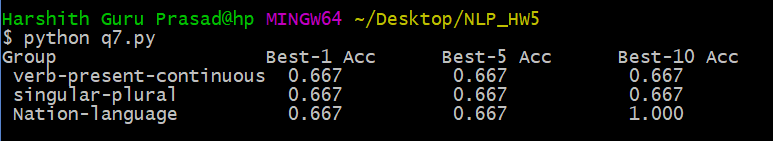
Used the following relations

* 1. verb ( present vs present continuous )
* 2. singular vs plural
* 3. nation vs national language

Created the following analogies (questions) for words present in vocab



Accuracies generated using the dense word embedding vectors using the cosine similarity metric



**Observations:**

The stem(lemma) or root of a word can be used as a powerful feature in determining the similarity of words. Nations, nationalities, languages are easily derivable from each other. Similarly, for the plural forms of words, the common stem form helps in identifying similarities and associated contexts. Proper nouns are associated with semantically closer words than common nouns. This is because of the limited context in which their meaning is defined and used.

Tried to determine the accuracies using sparse context count vector but the results were disappointing with accuracies close to 0. Sparse context count vectors are less helpful for analogies and suffer from performance issues. The dense vectors predicted 2/3 analogies accurately in each group. The relation nation-language has a best-10 accuracy of 1. This is on account of the similar context in which the national language of a nation is used. It can also be used in the context of nationality or as an adjective.

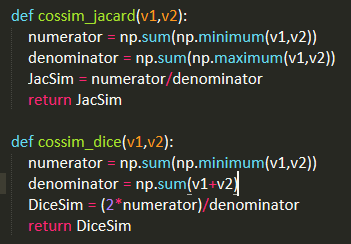
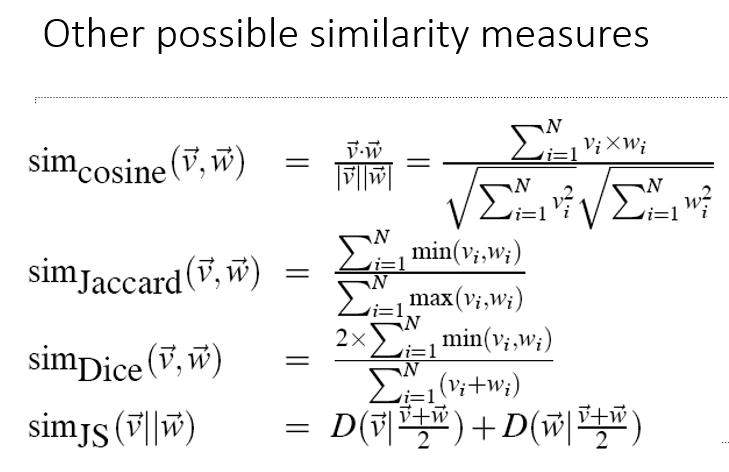
Alternate word tenses and comparative degrees can be predicted with good accuracies on account of same stem and frequency of similar contexts.

Predicting nations, nationalities, languages, adjectives associated with nations, locations and proper nouns is also relatively easy as a results of same stem and context similarity.

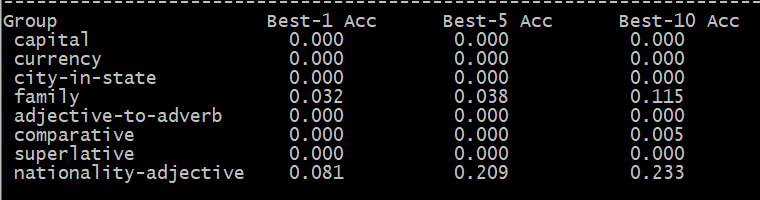
**Extra credit – Tried to Improve Prediction Acuuracy**

**ATTEMPT -1 : Alternate similarity metrics**

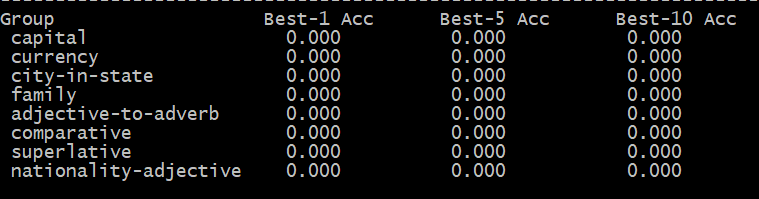
**Implemented Jacard and Dice Similarity ( Bad Accuracies )**



**Poor performance using Jacard similarity – not helpful**



**Poor performance using Dice similarity – not helpful**



**ATTEMPT - 2 : Alternate pre trained vectors**

Used pre trained word vectors from https://nlp.stanford.edu/projects/glove/

Downloaded the vectors for words from Wikipedia 2014 and Gigaword5(822 MB)

**ATTEMPT – 3 : Stemming words**

Tried to add the context of word stem for word vectors. Logic: words having the same stem are more likely to have the same semantic meaning. However, this technique does not generate the words that can be derived or inflected from the stemmed forms. Hence did not improve prediction accuracy as expected.