HML: ASSIGNMENT 3 – VECTOR REPRESENTATION AND LSA

TOEFL Synonym Questions

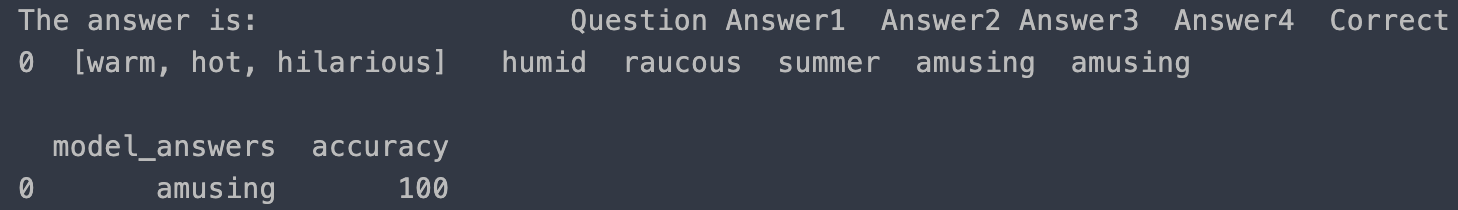
1. How well does the model do?
   1. The model performs at an accuracy rate of 13/20 i.e. 65% which is not bad, but it is not great as well. The below data can be used to see how and in what values does the novel model with cosine similarities work on various words from an English test.

|  |  |
| --- | --- |
|  |  |

1. What kind of items does the model do well on? What kind of items does the model fail on? Why does it fail on those items?
   1. As we can see from the above images, the model fails answering questions that have options of answers which are very closely related to each other. For example, the first question and the 9th question about “large” and “sad” and a synonym is asked for that. The options for the “large” word are too close to one another, during the class as well only one person correctly guessed the answer. So, it is an overall tricky question. This might be the case because while representing a vector addition or any function on a 2D plane, as demonstrated by professor, “cat likes dog” and “dog likes cat” are the same point on the plane but they are two totally different sentences. This was also demonstrated by the exercise in the class, where a photo of a white porche was shown to us and all we could utter is the word “car” even though “automobile” is just as correct. However, the model performs well on questions that have answers a little un-corelated to one another. Like the question no. 19th, “fast” has which needs an antonym here has words like “huge” and “broken” which are not closely related neither are they related to “fast”.

SAT Analogy Questions

1. Why did you pick the items you picked? How well did model do on them?
   1. The way I checked similarity in this case was to subtract the answer from the question value for a pair of words and then comparing it with the other pair using np.dot() functionality which generates a dot product. I experimented with a lot of other functionalities like absolute difference, vector addition then comparing with dot product, Euclidean distance, cosine similarity and the approach I chose gave me the max number of correct outputs i.e. 3/7. To experiment, I picked items which had some mixtures of synonym, antonym, and other word associations in them to really check how well the model works. The model worked well on the items that had a clear winner and the words did not have very complicated associations and relations.
2. What kind of items does the model do well on? What kind of items does the model fail on? Why does it fail on those items?
   1. The model did well on as expected synonym, antonym questions because they were the simplest to perform and had a clear distinction in them, for example the below query. The words “warm” and “hot” are synonyms of each other and are closely related, the word that is the closest to “hilarious” is “amusing” based on the calculations that my function performed.



The model failed on items that had very complex word correlations or in some cases even If it was a simple enough relation, it was not able to answer correctly. Like the example question, “[about, bout, \_\_\_\_\_\_, mend]” Options - amend near tear dismiss amend Model Answer – near. We as humans can understand that “bout” is some word made from removing the “a” or the first letter of the word “about” but as word vectors these words are just not related.

Questions on LSA

1. LSA uses 300 dimensions for vector representation. Imagine you repeat this assignment but only use first 50 dimensions of vector representations. How do you think this would affect the results on the different tasks?
   1. It would computationally definitely be very effective, but there might be some information that is lost when dimensionality reduction occurs on a dataset. It can often lead to better performance because the data that was removed was just noise or redundant or even irrelevant data to figuring out the relation between word vectors. Or the model can just exhibit overfitting in these cases. In other cases, removing critical information can lead to the model answering at a very low accuracy. The key here is to experiment and identify which 50 dimensions can provide the correct output while still maintaining the accuracy levels of the overall experiment.
2. Do you think humans have a vector representation model for words? No, why not? Yes, then how do humans learn these vector representations?
   1. I do think humans have a vector representation model that is how we learn by association and correlation. Although it can be said of most cases like what animal is to cat or dog, or bark is to dog but other relations like “bout” is to “about” are learned and only a human cognition can sense these complex problems. LSA models do capture some complex concepts and vector representation is very useful in many cases and humans do in certain cases exhibit learning like a vector model but it is not for every case. Like the case for question where we had to find the synonym for “large”, it can be “massive” or “huge”, it can completely depend on the context for us whereas for an LSA model, it might not. For example, the “car” example from class, although all students in unison said “car” which is a result of typicality effect but there might be many minds that think of the color of the car or the brand or the model. But for a word vector model, it might not be the case.