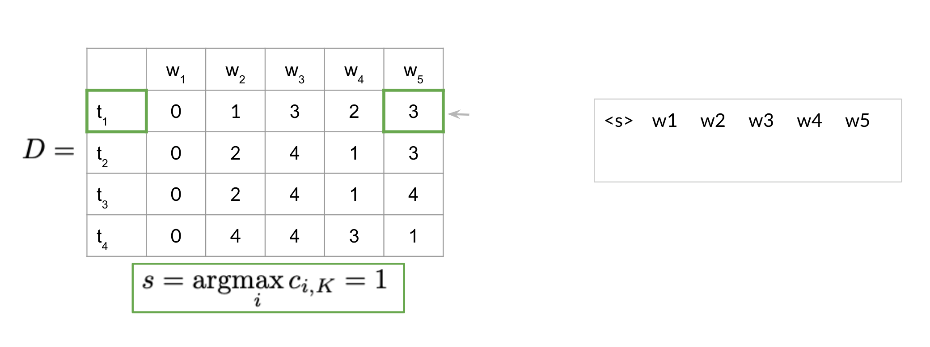
## Week 1 - Auto-Correct and Minimum Edit Distance

1. The minimum edit distance between the words ‘deep’ and ‘creepy’ is:
   1. 4 (You need to replace *d* for *c*, which counts for 2, insert *r* and insert *y*.)
2. What is NOT a valid example of an edit string operation?
   1. SWITCH a letter ‘Lusca’ --> ‘Lucas’ (Switching a letter is a valid operation ONLY when switching adjacent letters. In this case, there were two switches: *s* and *c*, after *s* and *a*.)
3. Autocorrect is only applicable when dealing with misspelled words.
   1. False. (Autocorrect can be used for words that do not make any sense for a particular sentence. For example, ‘Happy birthday deer friends’ is a correctly-spelled sentence, but the word ‘deer’ makes no sense – it should be ‘dear’.)
4. Given the corpus: “I am happy because I am doing quizzes.” and the following sentence: “I sm very good at solving quizzes.”, what is true?
   1. There is a unique correction for the misspelled word “sm”. (The correction would be the word “am”.)
5. What is true about the probabilistic model:
   1. Replacing a character costs more than deleting a character. (Replacing a word costs 2 whereas deleting it costs 1.)
   2. If is the number of times a word appears in a corpus and is the corpus size, then the probability of the word in the corpus is .
   3. The sentence “Happy birthday deer friends” would not have any word corrected. (Since the probabilistic model just looks at misspelled words, the above sentence would not be corrected.)
6. If we build a distance matrix for the following case: Source: Pie --> Target: Bye, what is the value of ?
   1. 5
7. If is the distance matrix for two words of the same size and is the matrix size for the minimum edit distance algorithm, then:
   1. (The first line will always have increasing values as we move to the right, because it is the cost from editing the null string.)
   2. The algorithm avoids use of brute force by implementing a dynamic programming approach. (Using previously computed cells to compute the next one is a dynamic programming method.)
8. What is NOT true about minimum edit distance?
   1. It is used to check if a word is misspelled.
9. The minimum edit distance calculation is more computationally expensive for a large corpus.
   1. False. (The minimum edit distance depends only on the editing cost and the two words that are being considered, not on any corpus or vocabulary.)
10. For the corpus “Autocorrect is a powerful tool and it is used on our computers”, the value for is:
    1. .

## Week 2 – Part of Speech Tagging

1. The transition matrix A allows you to:
   1. Compute the probability of going from a part of speech tag to another part of speech tag.
2. The emission matrix B allows you to:
   1. Compute the probability of going from a part of speech tag to a word.
3. The column sum of the emission matrix has to be equal to 1.
   1. False.
4. The row sum of the transition matrix has to be equal to 1.
   1. True.
5. Why is smoothing usually applied?
   1. Applying smoothing, for the minority of cases, allows us to increase the probabilities in the transition and emission matrices and this allows us to have non zero probabilities.
   2. Applying smoothing, for the majority of cases, allows us to decrease the probabilities in the transition and emission matrices and this allows us to have non zero probabilities.
6. Given the following D matrix, what would be the sequence of tags for the words on the right?



1. We had been multiplying the raw probabilities but in reality we take the log of those probabilities. Why?
   1. Because probabilities are bounded between 0 and 1 and as a result, the numbers could be too small and go toward 0.
2. What are useful applications for part of speech tagging?
   1. Speech recognition.
   2. Coreference resolution.
   3. Named entity recognition.

## Week 3 - Autocomplete

1. Corpus: “In every place of great resort the monster was the fashion. They sang of it in the cafes, ridiculed it in the papers, and represented it on the stage. ” (Jules Verne, Twenty Thousand Leagues under the Sea)

In the context of our corpus, what is the probability of word “papers” following the phrase “it in the”.

1. Given the following conditional probabilities, approximate the probability of the following sentence with bigrams: “Mary likes cats”:

; ; ; ;  
; ;

1. Given the following conditional probabilities, approximate the probability of the following sentence with bigrams:

; ; ; ;  
; ;

1. Given the logarithm of the following conditional probabilities, approximate the log probability of the following sentence with bigrams:

; ; ;

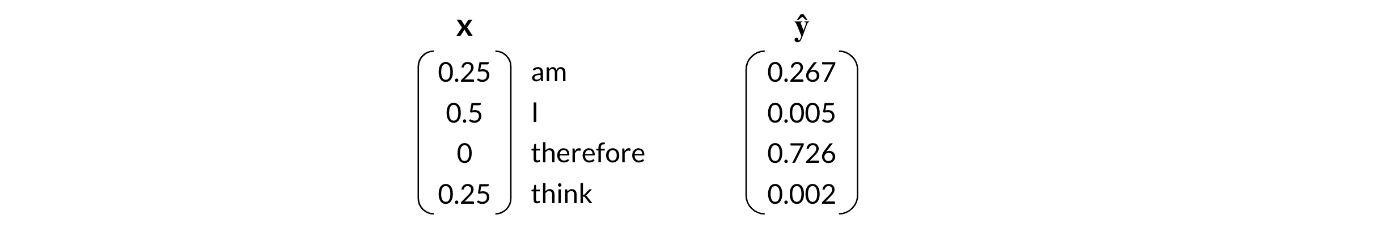
1. Given the same conditional probabilities, what is the model’s perplexity assuming the test set is
2. Given the following training corpus and a minimum word frequency of 2, what would the vocabulary for the corpus pre-processed with look like?

1. Given the following corpus, what is the estimates probability of the word “can” following the word “I” using the bigram model and adding k-smoothing where .

1. What are applications of n-gram language models?
   1. Speech recognition
   2. Auto-complete
   3. Auto-correct
   4. Augmentative communication
2. The higher the perplexity score, the more our corpus will make sense.
   1. False.
3. The perplexity score increases as we increase the number of tokens.
   1. False.

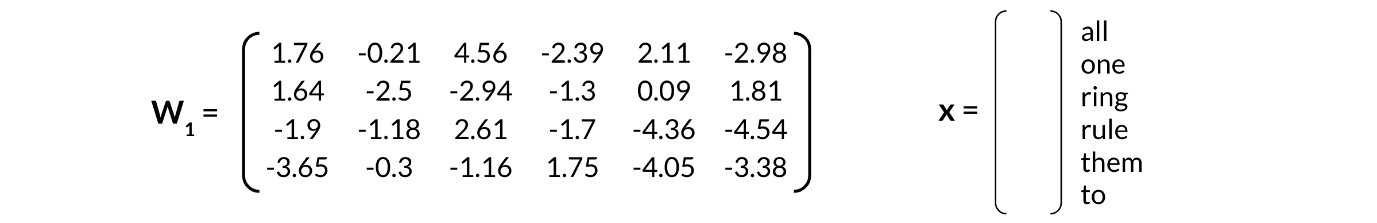
## Week 4 – Word Embeddings

1. Which word representation is most likely to correspond to a word embedding representation in a general-purpose vocabulary. i.e. which one is most likely to capture meaning and important information about the words?
   1. Car -> (0.1, 1); Caravan -> (-0.1, 0.9). (This is a vector representation where similar terms are closer together. (In a vehicle-specific vocabulary where cars and caravans are seen as dissimilar, Car -> (1, 0.1); Caravan -> (-1, -0.9) could have been a possible representation.)
2. The meaning of the words, as carried by the word embeddings, depends on the embedding approach.
   1. True. (The specifics of the task are what will ultimately define the meaning of the individual words, e.g. assuming that words are surrounded by the same kings of words have similar meaning.)
3. You need to train a deep neural network to learn word embeddings.
   1. False.
4. Consider the corpus "A robot may not injure a human being or, through inaction, allow a human being to come to harm." and assume you are preparing data to train a CBOW model. Ignoring punctuation, for a context half-size of 3, what are the context words of the centre word "inaction"?
   1. “being or through allow a human” (The context words are 3 words to both the left and the right of the centre word.)
5. The continuous bag-of-words model learns to predict context words given a centre word.
   1. False. (It’s the reverse: The CBOW model learns to predict a centre word give context words. The continuous skip-gram model learns to predict context words given a centre word.)
6. You are designing a neural network for a CBOW model that will be trained on a corpus with a vocabulary of ,8000 words. If you want it to learn 400-dimensional word embedding vectors, what should be the sizes of the input, hidden, and output layers?
   1. 8,000 (input layer), 400 (hidden layer), 8,000 (output layer)
7. If you are designing a neural network for a CBOW model that will be trained on a corpus of 8000 words, and if you want it to learn 400-dimensional word embedding vectors, what should be the size of , the weighting matrix between the input layer and hidden layer, if it is fed training examples in batches of 16 examples represented by an 8,000 row by 16 column matrix?  
   (Hint: If is the input matrix, is the matrix for the hidden layer and the bias matrix, then ).
   1. 400 rows by 8,000 columns. (The size of does not depend on batch size.)
8. Given the input vector x below, a trained continuous bag-of-words model outputs the vector below. What is the word predicted by the model?



* 1. Therefore. (Values in are interpreted as the probabilities of each word being the centre word. As “therefore” corresponds to the element in with the highest value, it’s the model’s prediction.)

1. The following weighting matrix has been learned after training a CBOW model. You are also given word-to-row mapping for the input column vectors.

  
What is the word embedding vector for “ring”?

1. Select the correct statements:
   1. Extrinsic evaluation evaluates actual usefulness of embeddings, is time consuming and is more difficult to troubleshoot.
   2. You can perform intrinsic evaluation by using a clustering algorithm to group similar word embedding vectors, and determining if the clusters capture related words.
   3. To evaluate word embeddings with extrinsic evaluation, you use the word embeddings to perform an external task, which is typically the real-world task that you initially needed the word embeddings for. Then, use the performance metric of this task as a proxy for the quality of the word embeddings.