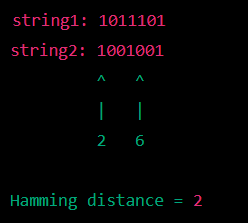
# String similarity metrics

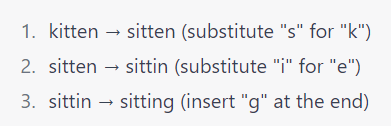
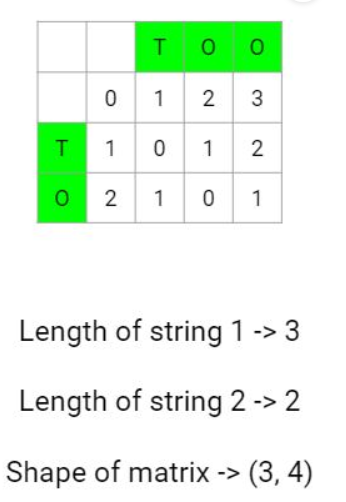
## Hamming distance

* Number of positions at which two strings differ.
* String must have same length.
* Add padding to with shorter if length is different.
* Can be used in spell-checking, plagiarism detection, and DNA sequence analysis



## Levenshtein distance or Minimum edit distance

Number of edit required for that will transform one string to other.

Answer is 3. Answer is 1.

Check if the characters are same

# fill the element with the minimum of (diagonal value, first corresponding value + 1, second corresponding value + 1)

If characters are not same

# fill the element with the minimum of corresponding values + 1

## Jaccard similarity

This measures the size of the intersection of two sets of characters divided by the size of the union of the two sets.

## Cosine similarity

This measures the cosine of the angle between two vectors in a multi-dimensional space, where each string is represented as a vector of character counts.

## Longest common subsequence

This measures the length of the longest sequence of characters that appears in both strings in the same order, but not necessarily consecutively.

Lecture 3:

Manhattan distance better then euclidean distance dealing with outlier

MD = abs difference btn axis

ED = squre root of difference

# Probablistic model:

* N-gram models are a type of Markov mode : probability of a sequence of words based on the probability of the previous n-1 words
* Latent Dirichlet Allocation (LDA) - Generative probabilistic mode, Topic modelling
* Probabilistic context-free grammar (PCFG) : Rule of grammer depends upon the grammer

## Markov Assumption:

the probability of a word depends only on the previous n-1 words

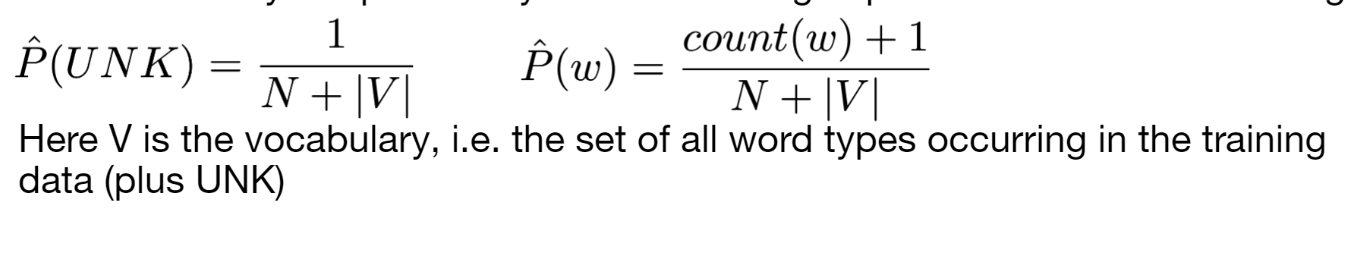
## Problem with unigram model:

* Lack of context:
* Out-of-vocabulary (OOV) words
* Overfitting
* High dimensionality

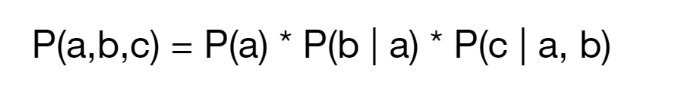
Still, it is very useful in the text classification problem.

## Laplace smoothing

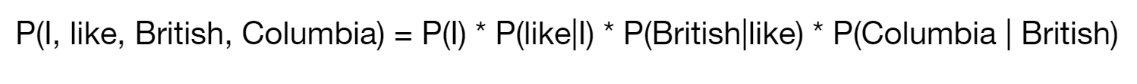
It is known as add-k smoothing. it help to give probablity to unseen data of training set.



Chain of probablity:



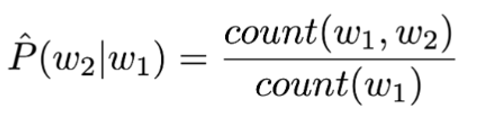
Based on the markov(unigram) :



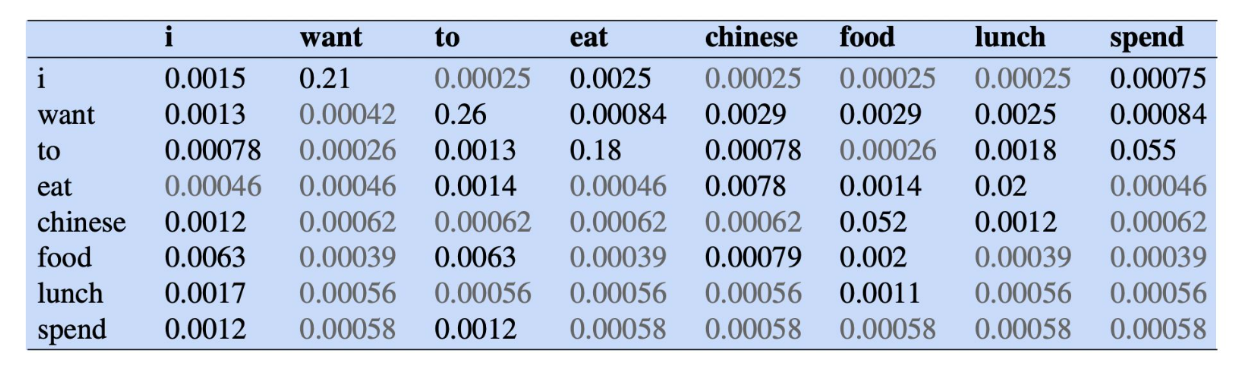
For biggram we need two word after like in 2nd prob.

## Bi gram model

It is also called as Markov chain.

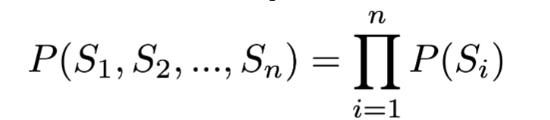


Normalized bi gram model:



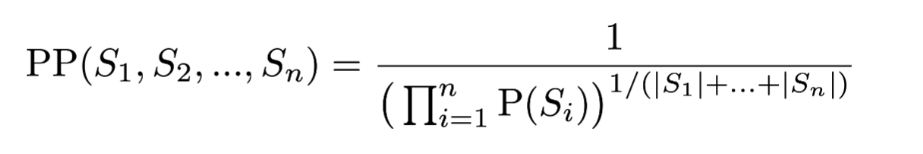
Good way to evaluate the model is test on the unseen data..

Likelihood of a collection of sentences:

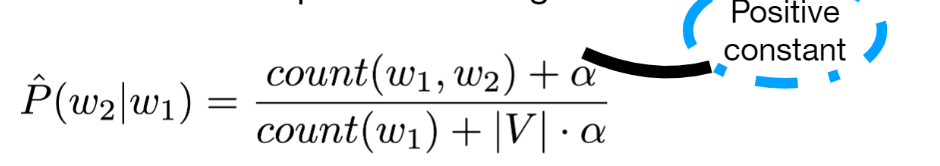


## Perplexity

* For perplexity we measure the normalized inverse above likelihood.
* Perplexity is a measure of how well a probability model is able to predict a given set of observations.
* It is evaluation matrix for the probablity model.



## Adding alpha smoothing



By using hmm with this approach there would be thousands of parameters. So to handle this stituation we can **manage with three state** :

* self.emissions, the list of word types, like “cat”, “dog”
* self.**states**, the list of possible POS tags. “noun”, “verb”
* **Emission distribution** : word is associated with the states like “noun” is probablity distribution for “cat’ and dog may higher then “the”
* 1 x size\_of\_state\_set
* P(w | state = i)
* **Transition probability** : probability of moving from one state to another
* From “noun” to “verb”
* P(state = j | state = i)
* size\_of\_state\_set x size\_of\_state\_set’

With the help of these four HMM to try to find the next word.

## **Generating from an HMM:**

* Generating new observations using new observation using all three states which are similar to used in training.
* It is new training part but we can say prediction part.

### Process steps

The process of generating from an HMM typically involves the following steps:

1. Start in a randomly chosen initial state.
2. Use the transition probabilities to determine the next state.
3. Use the emission probabilities of the current state to determine the next observation.
4. Repeat steps 2 and 3 for a desired number of steps or until a stopping criterion is reached.

Success of right next word : it is based on the **probabilities** learned from the training set.

* This can be achieved by using the **Viterbi Algorithm.**
* P(I,am,Sam; 1,2,3) = P(1|START) \* P(I|1) \* P(2|1) \* P(am|2) \* P(3|2) \* P(Sam|3)

not the same as P(I am Sam)

## **Inference Tasks**

Two states (Rainy and Sunny) and two observations (Umbrella and No Umbrella)

### **Evaluation**

Given a sequence of observations and the HMM parameters, this task calculates the likelihood of the observations being generated by the HMM.

Given observations [Umbrella, No Umbrella, No Umbrella, Umbrella, Umbrella, No Umbrella, No Umbrella].

### **Decoding**

Given a sequence of observations and the HMM parameters, this task finds the most likely sequence of states that generated the observations. This is also known as the Viterbi algorithm.

we have a sequence of observations for the week that is [Umbrella, No Umbrella, No Umbrella, Umbrella, Umbrella, No Umbrella, No Umbrella]. Given the HMM parameters, the most likely sequence of states that generated the observations is [Rainy, Sunny, Sunny, Rainy, Rainy, Sunny, Sunny].

## Hidden and visible states

### Case 1: Inferring Hidden States in an Unsupervised Manner

The states are completely unknown: In this case, the states are not given and we need to infer them from the data in an **unsupervised** manner. This can be done by using techniques such as clustering or Expectation-Maximization (EM) algorithm to group similar observations together and assign each group to a state.

### Case 2: Known Hidden States

The states aren't actually hidden: In this case, the states can be POS tags that come from an annotated corpus. This means that the states are already known and we don't have to infer them from the data.

It is straightforward to estimate model parameters: With the known states, we can use the annotated corpus to estimate the model parameters such as the initial state probabilities, the emission probabilities, and the transition probabilities.

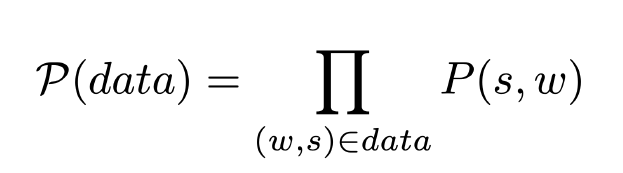
In the 2nd case, it's a **supervised** learning problem, thus it's easier to estimate the model parameters comparing to the 1st case.

## **Training with observable states**

When we train with observable states (e.g. POS tags), the objective is to find model parameters which maximize the joint probability of the states and

emissions in our training corpus:

Where data is a collection of sentences w = w1 ... wk and state sequences s = s1 ... sk



This is called maximum likelihood estimation (MLE)

