Lecture Notes: Bayes Theorem & Conditional Probability

Bayes' Theorem and Conditional Probability form the cornerstone of inferential statistics and probabilistic reasoning. These concepts enable us to update our beliefs based on new data, make predictions, and model complex dependencies between random variables. In this chapter, we will explore the fundamental principles of conditional probability and Bayes' Theorem, along with their applications and examples.

Independence

The concept of independence is crucial in the study of probability and statistics. When dealing with random variables, independence ensures that the realization of one random variable doesn't give any information about the realization of another.

Definition 1. *Let X and Y be two discrete random variables with probability mass functions $p_X(x)$ and $p_Y(y)$. X and Y are said to be **independent** if and only if for all x in the range of X and y in the range of Y, we have:*

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$$

This definition ensures that our understanding of randomness is preserved; knowledge about the outcome of one random process shouldn't provide any information about the outcome of another if they are indeed independent.

Example 1 (Independence with Coin Tosses)**.** *Let's consider a simple experiment involving coin tosses to illustrate the concept of independence between random variables.*

Let:

- *X_1 be a random variable representing the outcome of the first coin toss.*
- *X_2 be a random variable representing the outcome of the second coin toss.*

The coin can land on 'Heads' (H) or 'Tails' (T). For simplicity, let's denote 'Heads' as 1 and 'Tails' as 0. So, the random variables X_1 and X_2 can take values from the set {0, 1}.

By the nature of a fair coin, we have:

$$p_{X_1}(1) = 0.5$$

$$p_{X_2}(0) = 0.5$$

If X_1 and X_2 are independent, then:

$$p_{X_1,X_2}(1, 0) = p_{X_1}(1) \cdot p_{X_2}(0)$$

Plugging in our values:

$$p_{X_1,X_2}(1, 0) = 0.5 \cdot 0.5 = 0.25$$

This implies that the outcomes $X_1 = 1$ and $X_2 = 0$ are independent. Let's further verify the independence by evaluating the probabilities for all four possible outcomes:

$$\begin{aligned} p_{X_1,X_2}(1, 1) &= p_{X_1}(1) \cdot p_{X_2}(1) = 0.5 \cdot 0.5 = 0.25 \\ p_{X_1,X_2}(0, 1) &= p_{X_1}(0) \cdot p_{X_2}(1) = 0.5 \cdot 0.5 = 0.25 \\ p_{X_1,X_2}(0, 0) &= p_{X_1}(0) \cdot p_{X_2}(0) = 0.5 \cdot 0.5 = 0.25 \end{aligned}$$

Since for all combinations of X_1 and X_2, the joint probability mass function equals the product of the marginal probability mass functions, we can conclude that X_1 and X_2 are independent random variables in this coin-tossing experiment.

In practical scenarios, understanding such independence is crucial to ensure that simulations or experiments aren't inadvertently biased. However, there also exist scenarios where random variables can exhibit dependence, influencing each other's outcomes.

Experiment 1. *Let's illustrate dependence by considering two random variables* X_1 *and* X_2 *based on a sample* u *taken from the set* $\{1, 2, \ldots, m\}$.

- $X_1(u) = \frac{u}{m}$
- $X_2(u) = 1 - \frac{u}{m}$

By our definitions, the probabilities are:

- $p_{X_1}(x) = \frac{1}{2}$ *for* $x > \frac{1}{2}$
- $p_{X_2}(y) = \frac{1}{2}$ *for* $y > \frac{1}{2}$

However, if $X_1(u) > \frac{1}{2}$, *then* $X_2(u)$ *will necessarily be less than* $\frac{1}{2}$.
Thus, $p_{X_1,X_2}(x, y)$ *where* $x > \frac{1}{2}$ *and* $y > \frac{1}{2}$ *equals 0.*
This implies:

$$p_{X_1,X_2}(x, y) \neq p_{X_1}(x) \cdot p_{X_2}(y)$$

showing that X_1 *and* X_2 *are dependent random variables.*

It's important to understand the subtleties between independence and dependence as it allows us to model real-world phenomena accurately and make precise inferences about our world.

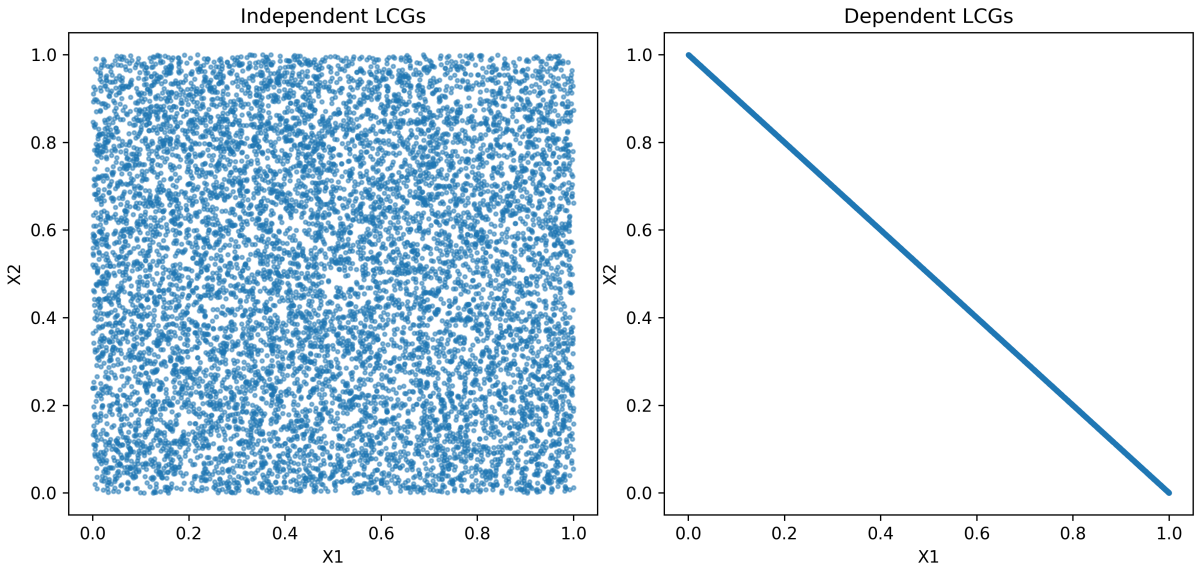

Figure 1: Simulation of dependent and independent LCG.

Conditional Probability

As we explored correlation between random variables in the previous section, we saw that they provide measures of how variables interact or remain unrelated. While these concepts help us understand the relationships between variables, often in the real world, we want to know the probability of one event, given that another event has already occurred. This leads us to the notion of conditional probability. Conditional probability refines our predictions based on new, additional information and serves as a bridge between pure independence and more intricate dependencies.

Definition 2 (Conditional Probability)**.** *Given two random variables X and Y with $p_Y(y) > 0$, the **conditional probability** of $X = x$ given $Y = y$ is defined as:*

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Conditional probability plays a pivotal role. It provides the mathematical framework to assess the likelihood of a random variable X given that another random variable Y has already occurred. Understanding this is particularly essential when X and Y are not independent. In such scenarios, the occurrence of Y can significantly alter the probability landscape for X. A direct consequence of the definition of conditional probability is the Multiplication Theorem. It provides a foundational bridge between joint and conditional probabilities, allowing for systematic computation of joint probabilities $p_{X,Y}(x,y)$.

Theorem 1 (Multiplication Theorem)**.** *Let X and Y be two random variables. The joint probability $p_{X,Y}(x,y)$ can be expressed in terms of conditional probabilities as:*

$$p_{X,Y}(x,y) = p_{X|Y}(x|y) \cdot p_Y(y) = p_{Y|X}(y|x) \cdot p_X(x)$$

Essentially, the Multiplication Theorem states that if we have knowledge of the probability of one random variable given another, we can ascertain the joint probability $p_{X,Y}(x,y)$.

For many random variables $X_1, X_2, \ldots, X_n$, the joint probability can be expressed as:

$$p_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = \prod_{i=1}^{n} p_{X_i|X_1,\ldots,X_{i-1}}(x_i|x_1,\ldots,x_{i-1})$$

This represents the product of the conditional probabilities of each random variable occurring given the occurrence of all previous random variables.

Marginal Probabilities

Definition 3 (Marginal Probability)**.** *Given a joint probability distribution $p_{X,Y}(x,y)$, the marginal probability $p_X(x)$ of any outcome x for the random variable X is obtained by summing the joint probabilities over all possible outcomes y for Y. Mathematically, the marginal probability $p_X(x)$ is given by:*

$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

where the sum is over all possible outcomes of the random variable Y.

The connection between marginal and conditional probabilities can be understood through the law of total probability. The marginal probability $p_X(x)$ can be expressed in terms of conditional probabilities as follows:

$$p_X(x) = \sum_y p_{X|Y}(x|y) \cdot p_Y(y)$$

This relationship demonstrates that the marginal probability of an outcome for a random variable can be obtained by considering all the ways that outcome can occur, weighted by the probability of each of those ways.

Example 2 (Health Screening)**.** *Consider a health camp where participants are screened for two health conditions based on their age group:*

- *Participants under 40 are screened for Vitamin D deficiency.*
- *Participants 40 and above are screened for High Blood Pressure (HBP).*

From past records, the camp organizers know the following:

- *60% of the participants are under 40.*
- *40% are 40 and above.*
- *Among those under 40, 20% are found to have Vitamin D deficiency.*
- *Among those 40 and above, 30% are found to have HBP.*

Let X represent a random variable for a participant having a health condition and Y be the random variable for a participant being under 40. We're interested in $p_X(x)$, the marginal probability of a participant having a health condition.

First, we compute the conditional probabilities:

$$p_{X|Y}(x|y) = 0.20$$

$$p_{X|Y^c}(x|y^c) = 0.30$$

Using the provided probabilities for $p_Y(y)$ and $p_{Y^c}(y^c)$, we can apply the law of total probability:

$$p_X(x) = p_{X|Y}(x|y) \cdot p_Y(y) + p_{X|Y^c}(x|y^c) \cdot p_{Y^c}(y^c)$$
$$p_X(x) = 0.20 \cdot 0.60 + 0.30 \cdot 0.40 = 0.24$$

Thus, the marginal probability of a participant having a health condition, either Vitamin D deficiency or HBP, at this health camp is 24%.

Bayes' Theorem

Suppose you are trying to determine if a piece of fruit picked from a bag is an apple. Your initial belief (prior probability) might be based on the overall percentage of apples in the bag. However, when you touch the fruit and feel it's round and smooth, you can update your belief based on this new evidence. Bayes' theorem provides a way to combine these sources of information.

The interplay of events in a probabilistic framework is not always straightforward. Often, we have evidence or observations and seek to update our understanding of a particular event's probability based on this new information. Bayes' theorem offers a mathematical means to achieve this. It allows us to reverse conditional probabilities, turning our perspective from the probability of observing evidence given an event to the probability of the event given the observed evidence.

Theorem 2 (Bayes' Theorem)**.** *Given two random variables X and Y with $p_Y(y) \neq 0$, the conditional probability $p_{X|Y}(x|y)$ is:*

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x) \cdot p_X(x)}{p_Y(y)}$$

Where: $p_{X|Y}(x|y)$ is the posterior probability, $p_{Y|X}(y|x)$ is the likelihood, $p_X(x)$ is the prior probability, and $p_Y(y)$ is the evidence.

Bayes' theorem provides a way to compute a posterior probability. It relates the likelihood of observing Y given X, the prior probability of X, and the total probability of observing Y.

Starting from the Multiplication Theorem:

$$p_{X,Y}(x, y) = p_{Y|X}(y|x) \cdot p_X(x)$$

Since $p_{X,Y}(x,y) = p_{Y,X}(y,x)$, we also have:

$$p_{Y,X}(y,x) = p_{X|Y}(x|y) \cdot p_Y(y)$$

Equating the two gives:

$$p_{X|Y}(x|y) \cdot p_Y(y) = p_{Y|X}(y|x) \cdot p_X(x)$$

Rearranging, we arrive at Bayes' theorem:

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x) \cdot p_X(x)}{p_Y(y)}$$

Example 3. *Imagine there's a rare disease, and there's a test for it. The disease affects 1% of the population, and the test is 99% accurate. If you test positive, what's the chance you actually have the disease?*

Using Bayes' theorem:

Let X be the random variable representing the presence of the disease. Let Y be the random variable representing the test result.

We want to find: $p_{X|Y}(x|+)$

Given:

- $p_X(x) = 0.01$ *(1% of the population has the disease)*
- $p_{Y|X}(+|x) = 0.99$ *(The test is 99% accurate)*
- $p_Y(+)$ *is the total probability of testing positive.*

$$p_{X|Y}(x|+) = \frac{p_{Y|X}(+|x) \cdot p_X(x)}{p_Y(+)}$$

To find $p_Y(+)$, consider:

$$p_Y(+) = p_{Y|X}(+|x) \cdot p_X(x) + p_{Y|X^c}(+|x^c) \cdot p_{X^c}(x^c)$$

$$p_Y(+) = (0.99)(0.01) + (0.01)(0.99) = 0.0198$$

Plugging in the numbers:

$$p_{X|Y}(x|+) = \frac{(0.99)(0.01)}{0.0198} \approx 0.5$$

So, even with a 99% accurate test, if you test positive, there's only a 50% chance you actually have the disease!

Bayes' theorem is foundational for the fields of Bayesian statistics and machine learning. It provides a mechanism to update our beliefs in light of new evidence, making it central to numerous applications, from medical diagnostics to recommendation systems. The theorem reminds us of the importance of prior knowledge and illustrates how, in a world filled with data, we can use this data to make more informed decisions and predictions.

Applications of the Bayes' Theorem

Understanding the significance of Bayes' theorem extends beyond just statistics and machine learning. It's a philosophy of learning from data, constantly updating our beliefs as we gather more information. In subsequent sections, we can delve into how this philosophy is applied in various scientific disciplines.

Bayes' theorem is foundational in modern statistics and machine learning, having a profound influence on the way we model and interpret data. This section will explore its significance and applications.

4.0.1 Natural Language Processing (NLP)

Bayes' theorem forms the foundation of many NLP applications:

- **Spam Filters**: Bayesian classifiers can determine if an email is spam or not based on the frequency of certain words.
- **Sentiment Analysis**: Using Bayes' theorem, algorithms can determine the sentiment of a given text (positive, negative, neutral) by analyzing the words used.

Example: Email Filtering

Email services use advanced algorithms to filter out spam emails, ensuring that users mostly receive legitimate messages in their inbox. One popular approach employs Bayesian methods to classify an email as spam or not based on its content.

Let's consider a general scenario:

- $p_S(\text{True})$ be the probability that the filter correctly classifies a spam email as spam (HitRate).
- $p_{F|S^c}(\text{True}|\text{False})$ be the probability that a legitimate email is incorrectly classified as spam (FalseAlarm).
- $p_S(\text{True})$ be the rate at which spam emails are received among all emails (SpamRate).

Given that an email has been flagged as spam by the filter, what's the probability that it truly is a spam email?

Using Bayes' theorem, we express this as:

$$p_{S|F}(\text{True}|\text{True}) = \frac{p_{F|S}(\text{True}|\text{True}) \times p_S(\text{True})}{p_F(\text{True})}$$

Where:

$$\begin{aligned} p_F(\text{True}) = {} & p_S(\text{True}) \times p_{F|S}(\text{True}|\text{True}) \\ & + (1 - p_S(\text{True})) \times p_{F|S^c}(\text{True}|\text{False}) \end{aligned}$$

4.0.2 Finance and Economics

Bayes' theorem finds applications in finance, especially in the area of risk management and investment strategies. Bayesian models can help in predicting the likelihood of certain economic conditions based on the current and past data.

- **Portfolio Management**: Investors can update their beliefs about the expected returns of assets based on new market information.
- **Risk Assessment**: Bayesian models can evaluate the risk of investments or loans by considering both historical data and expert judgment.

Example: Portfolio Management

Let:

- $p_{\text{D}}(D)$ be the prior density of the random variable D that represents a 5% decline, estimated to be 0.7.
- $p_{\text{N}|\text{D}}(N|D)$ be the density of N given D, which represents the probability that the news predicts a 10% decline given that there will be an actual 5% decline. Estimated to be 0.8.

- $p_{\mathrm{N}}(N)$ be the marginal density of N, which accounts for the times the news correctly and incorrectly predicts a 10% decline.

Using Bayes' theorem:

$$p_{\mathrm{D|N}}(D|N) = \frac{p_{\mathrm{N|D}}(N|D) \times p_{\mathrm{D}}(D)}{p_{\mathrm{N}}(N)}$$

Where $p_{\mathrm{N}}(N)$ can be calculated as:

$$p_{\mathrm{N}}(N) = p_{\mathrm{D}}(D) \times p_{\mathrm{N|D}}(N|D) + (1 - p_{\mathrm{D}}(D)) \times p_{\mathrm{N}|\neg\mathrm{D}}(N|\neg D)$$

Example: Sentiment Analysis in Product Reviews

Suppose we are working on sentiment analysis for product reviews and want to assess the sentiment behind certain reviews.

Let:

- $p_{\mathrm{Pos}}(\mathrm{Pos}) = 0.6$ be the prior density that a randomly selected review is positive.
- $p_{\mathrm{Neg}}(\mathrm{Neg}) = 0.3$ be the prior density that a randomly selected review is negative.
- $p_{\mathrm{Neut}}(\mathrm{Neut}) = 0.1$ be the prior density that a randomly selected review is neutral.
- $p_{\mathrm{LA|Pos}}(L, A|\mathrm{Pos}) = 0.7$ be the density that the words "love" and "average" appear in a positive review.
- $p_{\mathrm{LA|Neg}}(L, A|\mathrm{Neg}) = 0.2$ be the density that the words "love" and "average" appear in a negative review.
- $p_{\mathrm{LA|Neut}}(L, A|\mathrm{Neut}) = 0.1$ be the density that the words "love" and "average" appear in a neutral review.

First, we calculate $p_{\mathrm{LA}}(L, A)$ as follows:

$$p_{\mathrm{LA}}(L, A) = p_{\mathrm{Pos}}(\mathrm{Pos}) \times p_{\mathrm{LA|Pos}}(L, A|\mathrm{Pos}) + p_{\mathrm{Neg}}(\mathrm{Neg}) \times p_{\mathrm{LA|Neg}}(L, A|\mathrm{Neg}) + p_{\mathrm{Neut}}(\mathrm{Neut}) \times p_{\mathrm{LA|Neut}}(L, A|\mathrm{Neut})$$

$$= 0.6 \times 0.7 + 0.3 \times 0.2 + 0.1 \times 0.1 = 0.42 + 0.06 + 0.01 = 0.49$$

Using Bayes' theorem, the posterior density $p_{\mathrm{Pos|LA}}(\mathrm{Pos}|L, A)$ can be calculated as:

$$p_{\mathrm{Pos|LA}}(\mathrm{Pos}|L, A) = \frac{p_{\mathrm{LA|Pos}}(L, A|\mathrm{Pos}) \times p_{\mathrm{Pos}}(\mathrm{Pos})}{p_{\mathrm{LA}}(L, A)}$$

$$= \frac{0.7 \times 0.6}{0.49} \approx 0.8571$$

This indicates that given the words "love" and "average" appear in a review, there is approximately an 85.71% chance that the review is positive.

Automated text generations

The pursuit of automated text generation has been a focal point in computational linguistics, aiming to create coherent and contextually relevant text based on a set of rules, patterns, or examples. The motivation behind this study is to explore the realm of possibilities within text generation, focusing on generating stylistically consistent and coherent text using conditional probability concepts.

The model commences text generation by calculating the conditional probability of the occurrence of a word given the preceding sequence of words. For example, if the depth is set to 2, the model calculates the probability that a word follows a pair of two preceding words:

$$P(\text{word}_2|\text{word}_1) \tag{1}$$

The model utilizes these probabilities to make informed decisions on the next word in the sequence, aiming to produce coherent and stylistically similar text to the input.

Part 1: The robot just born

Imagine we start programing a robot, and we want him to learn how to generate text. We first provide the robot with a first text, this constitutes all information he knows about the world.

> *Once upon a time in a land far, far away, there lived a kind and gentle king. The king had a beautiful daughter, the princess, who was admired by everyone in the kingdom. The princess loved to wander in the woods and sit by the river, a favorite spot of hers, and dream for hours.*
>
> *One day while the princess was sitting by the river, a frog hopped onto her lap. The princess was startled but felt pity for the frog. The frog said, "I am not really a frog but a handsome prince who has been cursed by a wicked witch. If you kiss me, I will turn back into a prince.*
>
> *The princess leaned down and gave the frog a gentle kiss. In an instant, the frog transformed into the prince, and they lived happily ever after.*

So, for the robot to learn, a first approach is to look at all two-word combinations, we can plot that in a heat map

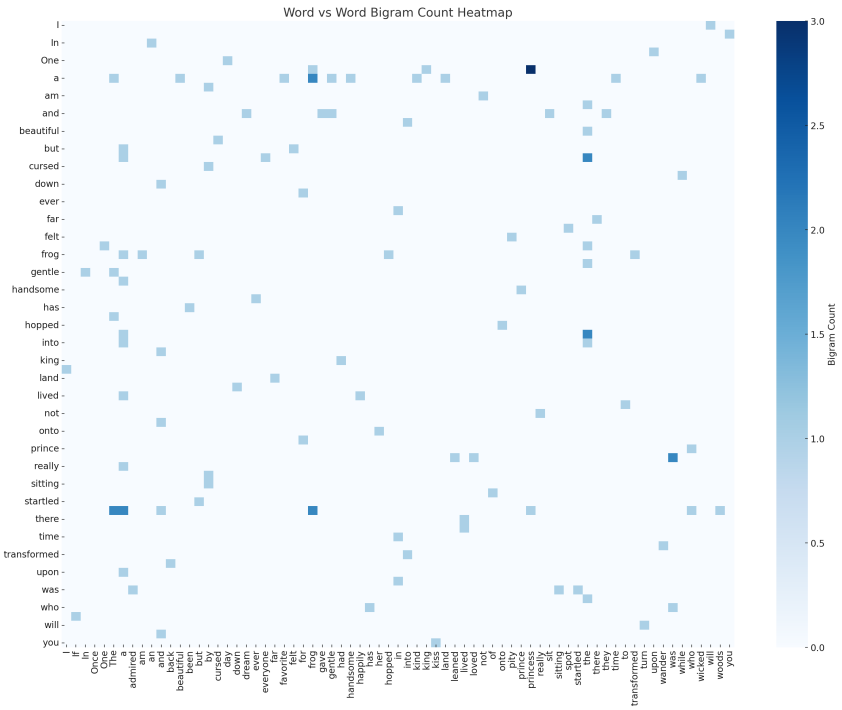

Figure 2: Heat map illustrating the frequency of two-word combinations in the story.

Now the robot knows something. Part of what he knows is described in a network in Figure bellow.

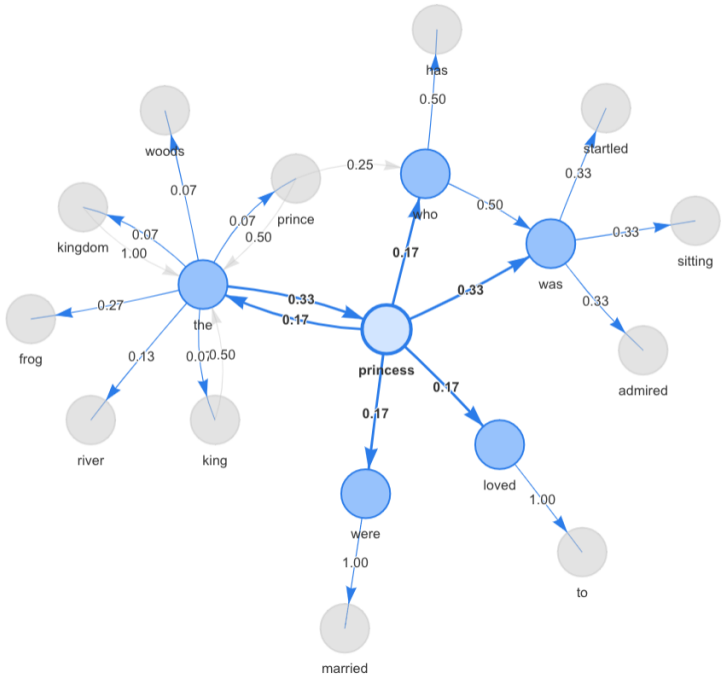

Figure 3: Textual Network of the Story

For the leaning process of the robot, we tokenized the story and represented it as a directed graph where each unique word is a node, and each directed edge represents the transition from one word to the next, with edge weights representing the frequency of transition.

The network has the probabilities of going from one word to another, we can use this information to generate text.

Part 2: The robot learns from whole books now, and get smarter

The model commences text generation by calculating the conditional probability of the occurrence of a word given the preceding sequence of words. Now, we can feed the robot with more data: Books.

Moreover, with enough data we can also meke it more inteligent by making the model slightly more complex. For example, if the depth is set to 2, the model calculates the probability that a word follows a pair of two preceding words:

$$P(\text{word}_3|\text{word}_1, \text{word}_2) \tag{2}$$

The model utilizes these probabilities to make informed decisions on the next word in the sequence, aiming to produce coherent and stylistically similar text to the input.

The writer function operates using a list of words from the input text and generates text based on conditional probabilities. Here is a simplified overview of the algorithm:

1. Load the text data from a source file.
2. Define the parameters:
    - *n.words:* The desired length of the generated text.
    - *depth:* The number of preceding words to be considered for generating the next word.
    - *seed:* The starting sequence for the generated text.

3. If no seed is provided, select a starting point randomly from the input list of words.
4. For each new word to be generated, calculate the conditional probability based on the preceding sequence of words and select the next word accordingly.
5. Continue the process until the generated text reaches the specified length, i.e., *n.words*.