We can start off by just creating a table for the joint distribution of the random variable O for the outcomes and G for the genres. They are defined as

$p(O, G)$	r	d	s	$p(O)$
w	$p(w, r) = \frac{\mathbf{3}}{46}$	$p(w, d) = \frac{\mathbf{8}}{46}$	$p(w, s) = \frac{\mathbf{0}}{46}$	$p(w, G) = \frac{\mathbf{11}}{46}$
n	$p(n, r) = \frac{\mathbf{3}}{46}$	$p(n, d) = \frac{\mathbf{5}}{46}$	$p(n, s) = \frac{\mathbf{4}}{46}$	$p(n, G) = \frac{\mathbf{12}}{46}$
o	$p(o, r) = \frac{\mathbf{10}}{46}$	$p(o, d) = \frac{\mathbf{6}}{46}$	$p(o, s) = \frac{7}{46}$	$p(o, G) = \frac{\mathbf{23}}{46}$
$p(G)$	$p(O, r) = \frac{\mathbf{16}}{46}$	$p(O, d) = \frac{\mathbf{19}}{46}$	$p(O, s) = \frac{\mathbf{11}}{46}$	46

Some notes on this table, in the margins you have the marginal distributions, for simplicity I wrote them as $p(O, r)$ for example isntead of $p(O, G = r)$.

$$P(O) = \sum_{o \in O} p(o, G = r)$$

Formally conditional entropy is defined as follows:

If X and Y are discrete random variables, and $p(x, y)$ and $p(y \mid x)$ are the values of the joint and conditional probability distributions, then we define conditional entropy as

$$H(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y \mid x)$$

Now applying this to our table we have that

- Y is the random variable O and X is the random variable G.
- We have that $p(G) = \frac{16}{46}, \frac{19}{46}, \frac{11}{46}$.
- We have that $O = \{r, d, s\}$ and $G = \{w, n, o\}$.

Starting off we can remember that all entropy is, is the expected codelength for a probability distribution.

$$H(X) = \mathbb{E}[L(x)] = \sum_{x \in X} p(x) L(x) = -\sum_{x \in X} p(x) \log p(x)$$

Joint probability extends this idea to two random variables.

$$H(X, Y) = \mathbb{E}[L(x, y)] = \sum_{x \in X} \sum_{y \in Y} p(x, y) L(x, y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$

Some ways to interpret this

- In both of these cases all we are doing is weighing the codelength of each outcome by the probability of that outcome. So all we are doing is getting the expected number of bits required to encode a certain type of outcome.
- Each instance of a random variable (or pairs of instances for joint entropy) has a probability associated with it $p(x)$ or $p(x, y)$ and a certain number of bits required to encode that outcome $L(x)$ or $L(x, y)$.

Conditional entropy is then the expected number of bits required to encode an outcome of a random variable given that we know the outcome of another random variable. So if we have two random variables like in our instance O and G then we can think of conditional entropy as the expected number of bits required to encode an outcome of O given that we know the outcome of G. In full we can write this as

$$\begin{aligned} H(O \mid G) = \mathbb{E}[L(p(O \mid G))] &= \sum_{g \in G} \sum_{o \in O} p(o, g) L(p(o \mid g)) \\ &= -\sum_{g \in G} \sum_{o \in O} p(o, g) \log p(o \mid g) \\ &= -\sum_{g \in G} \sum_{o \in O} p(g) p(g \mid o) \log p(o \mid g) \\ &= -\sum_{g \in G} p(g) \sum_{o \in O} p(o \mid g) \log p(o \mid g) \\ &= -\sum_{g \in G} p(g) H(O \mid G = g) = -\sum_{g \in G} p(g)\, \mathbb{E}[L(O \mid g)] \end{aligned}$$

So once again we can see that all we are doing is calculating the average number of bits required to encode an outcome (weighted sum) given that we know the outcome of another random variable. And all thats really different here is that $L(O \mid g)$ scales the code length of each outcome by the probability of the outcome we are conditioning on.

It can also be thought of as the joint entropy minus the entropy of the random variable we are conditioning on.

$$H(O \mid G) = H(O, G) - H(G)$$

An important observation to make is that the only difference between conditional and joint entropy is the distribution we use to calculate the code length. In joint entropy we use the joint distribution $p(x, y)$ and in conditional entropy we use the conditional distribution $p(x \mid y)$. For conditional entropy we have for each pair (o, g) the conditional code length is

$$L(o \mid g) = -\log p(o \mid g) = -\log \frac{1}{p(g)} p(o, g)$$

So we can see the only difference between conditional and joint entropy is the factor of $\frac{1}{p(g)}$ in the code length. What this factor does it that it reduces the sample space of the random variable G to the instance where $G = g$, so we only consider within the space of $G = g$.

Bringing this back now to the main point, in a decision tree we want to minimize the conditional entropy of the outcomes (O) given the features (e.g. G).

From the conditional entropy we can then calculate the information gain for each feature. Information gain measures how much knowing the value of G decreases the uncertainty about O. It is the difference between the entropy of the outcomes and the conditional entropy of the outcomes given the feature. Formally we have that

$$I_O(G) = H(O) - H(O \mid G)$$

In practice if the set of instance before the split is S and the split gives us $S_1, S_2, \ldots, S_n$ then the information gain is the entropy of S minus the sum of the entropies of the sets after the split weighted by the size of the set. Formally we have that

$$I_S(G) = H(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} H(S_i)$$

$$\begin{aligned}
I_S(G) &= H(S) - \sum_{g \in G} p(g) H(S_g) \\
&= H(S) - \sum_{g \in G} p(g) H(O \mid G = g) \\
&= H(S) - \sum_{g \in G} p(g)\, \mathbb{E}[L(O \mid g)] \\
&= H(S) - \mathbb{E}[\mathbb{E}[L(O \mid G)]]
\end{aligned}$$

Some clarifying notes here, S in this case is analogous to the probability distribution $p(O) = p(O, G = g)$ and S_i to $p(G) = p(O = i \mid G)$, for example the probability of the outcome o (overlooked) can be calculated as follows

$$\begin{aligned}
\frac{|S_i|}{|S|} &= \frac{23}{(23 + 12 + 11)} = \frac{23}{46} \\
&\Leftrightarrow \\
p(o) &= \frac{\text{probability of outcome = o}}{\text{probability of all other outcomes}} \\
&= \frac{p(o, G)}{p(w, G) + p(o, G) + p(n, G)} = \frac{\sum_{g \in G} p(o, g)}{\sum_{o \in O} \sum_{g \in G} p(o, g)} \\
&= \frac{(23/46)}{(23/46 + 12/46 + 11/46)} = \frac{23}{46}
\end{aligned}$$