

**1. [10 points]** Using information Gain as the criteria for deciding the node to split, construct a decision tree manually with the depth as 2. Using data provided in the Table below, show all the computations/steps of your work for the full credit(i.e. how do you decide which feature to use for performing the first split and the other splits). Draw the final decision tree. (You do not need to show intermediate trees.)

Day	Outlook	Temperature	Humidity	Wind	Played Tennis?
01	Sunny	Hot	High	Weak	No
02	Sunny	Hot	High	Strong	No
03	Overcast	Hot	High	Weak	Yes
04	Rain	Mild	High	Weak	Yes
05	Rain	Cool	Normal	Weak	Yes
06	Rain	Cool	Normal	Strong	No
07	Overcast	Cool	High	Strong	Yes
08	Sunny	Cool	High	Weak	No
09	Sunny	Mild	Normal	Weak	Yes
10	Rain	Hot	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Untitled 1.ods — LibreOffice Calc

File Edit View Insert Format Styles Sheet Data Tools Window Help



Liberation Sans 10 pt B I U A

K19  $\sum$  =

	A	B	C	D	E	F	
1	Day	Outlook	Temperature	Humidity	Wind	Played tennis	
2		1 Sunny	Hot	High	Weak	no	
3		2 Sunny	Hot	High	Strong	no	
4		3 Overcast	Hot	High	Weak	yes	
5		4 Rain	Mild	High	Weak	yes	
6		5 Rain	Cool	Normal	Weak	yes	
7		6 Rain	Cool	Normal	Strong	no	
8		7 Overcast	Cool	High	Strong	yes	
9		8 Sunny	Cool	High	Weak	no	
10		9 Sunny	Mild	Normal	Weak	yes	
11		10 Rain	Hot	Normal	Weak	yes	
12		11 Sunny	Mild	Normal	Strong	yes	
13		12 Overcast	Mild	High	Strong	yes	
14		13 Overcast	Hot	Normal	Weak	yes	
15		14 Rain	Mild	High	Strong	no	
16							
17							

To manually construct a decision tree using **information gain** and **entropy** as the splitting criteria, we will proceed step by step, using the table data you provided. We'll show the entropy calculations and how we decide on the feature to split at each level of the tree, up to depth 2.

### Step 1: Calculate the entropy of the entire dataset

The entropy formula is:

$$H(S) = -p_{yes} \log_2(p_{yes}) - p_{no} \log_2(p_{no})$$

In the dataset, there are 9 "yes" and 5 "no" outcomes for the "Played tennis" feature.

$$p_{yes} = \frac{9}{14}, \quad p_{no} = \frac{5}{14}$$

So, the entropy of the entire dataset  $H(S)$  is:

$$H(S) = - \left( \frac{9}{14} \right) \log_2 \left( \frac{9}{14} \right) - \left( \frac{5}{14} \right) \log_2 \left( \frac{5}{14} \right)$$

$$H(S) = -0.643 \cdot \log_2(0.643) - 0.357 \cdot \log_2(0.357)$$

$$H(S) = -0.643 \cdot (-0.645) - 0.357 \cdot (-1.485) = 0.414 + 0.530 = 0.940$$

The entropy of the entire dataset is approximately 0.940.

## Step 2: Calculate the information gain for each feature

We now calculate the **information gain** for each feature by splitting on that feature and checking how much the entropy decreases.

### Split by "Outlook"

The possible values for "Outlook" are: Sunny, Overcast, Rain.

- **Sunny:** 5 examples (2 yes, 3 no)
- **Overcast:** 4 examples (4 yes, 0 no)
- **Rain:** 5 examples (3 yes, 2 no)

The entropy for each subset:

- $H(\text{Sunny}) = - \left( \frac{2}{5} \right) \log_2 \left( \frac{2}{5} \right) - \left( \frac{3}{5} \right) \log_2 \left( \frac{3}{5} \right)$

$$H(\text{Sunny}) = -0.4 \log_2(0.4) - 0.6 \log_2(0.6) = 0.971$$

- $H(\text{Overcast}) = - \left( \frac{4}{4} \right) \log_2 \left( \frac{4}{4} \right) = 0$  (since all are yes)
- $H(\text{Rain}) = - \left( \frac{3}{5} \right) \log_2 \left( \frac{3}{5} \right) - \left( \frac{2}{5} \right) \log_2 \left( \frac{2}{5} \right)$

$$H(\text{Rain}) = -0.6 \log_2(0.6) - 0.4 \log_2(0.4) = 0.971$$

Now, the weighted average entropy for the split on "Outlook" is:

$$H(\text{Outlook}) = \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 = 0.693$$

Information Gain for "Outlook":

$$IG(\text{Outlook}) = H(S) - H(\text{Outlook}) = 0.940 - 0.693 = 0.247$$

Split by "Temperature"

The possible values for "Temperature" are: Hot, Mild, Cool.

- **Hot:** 4 examples (2 yes, 2 no)
- **Mild:** 6 examples (4 yes, 2 no)
- **Cool:** 4 examples (3 yes, 1 no)

The entropy for each subset:

- $H(\text{Hot}) = - \left(\frac{2}{4}\right) \log_2 \left(\frac{2}{4}\right) - \left(\frac{2}{4}\right) \log_2 \left(\frac{2}{4}\right) = 1.0$
- $H(\text{Mild}) = - \left(\frac{4}{6}\right) \log_2 \left(\frac{4}{6}\right) - \left(\frac{2}{6}\right) \log_2 \left(\frac{2}{6}\right) = 0.918$
- $H(\text{Cool}) = - \left(\frac{3}{4}\right) \log_2 \left(\frac{3}{4}\right) - \left(\frac{1}{4}\right) \log_2 \left(\frac{1}{4}\right) = 0.811$

Weighted average entropy for "Temperature":

$$H(\text{Temperature}) = \frac{4}{14} \times 1.0 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.811 = 0.911$$

Information Gain for "Temperature":

$$IG(\text{Temperature}) = 0.940 - 0.911 = 0.029$$

Split by "Humidity"

The possible values for "Humidity" are: High, Normal.

- **High:** 7 examples (3 yes, 4 no)
- **Normal:** 7 examples (6 yes, 1 no)

The entropy for each subset:

- $H(\text{High}) = -\left(\frac{3}{7}\right) \log_2 \left(\frac{3}{7}\right) - \left(\frac{4}{7}\right) \log_2 \left(\frac{4}{7}\right) = 0.985$
- $H(\text{Normal}) = -\left(\frac{6}{7}\right) \log_2 \left(\frac{6}{7}\right) - \left(\frac{1}{7}\right) \log_2 \left(\frac{1}{7}\right) = 0.592$

Weighted average entropy for "Humidity":

$$H(\text{Humidity}) = \frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.592 = 0.789$$

Information Gain for "Humidity":

$$IG(\text{Humidity}) = 0.940 - 0.789 = 0.151$$

Split by "Wind"

The possible values for "Wind" are: Weak, Strong.

- **Weak:** 8 examples (6 yes, 2 no)
- **Strong:** 6 examples (3 yes, 3 no)

The entropy for each subset:

- $H(\text{Weak}) = -\left(\frac{6}{8}\right) \log_2 \left(\frac{6}{8}\right) - \left(\frac{2}{8}\right) \log_2 \left(\frac{2}{8}\right) = 0.811$
- $H(\text{Strong}) = -\left(\frac{3}{6}\right) \log_2 \left(\frac{3}{6}\right) - \left(\frac{3}{6}\right) \log_2 \left(\frac{3}{6}\right) = 1.0$

Weighted average entropy for "Wind":

$$H(\text{Wind}) = \frac{8}{14} \times 0.811 + \frac{6}{14} \times 1.0 = 0.892$$

Information Gain for "Wind":

$$IG(\text{Wind}) = 0.940 - 0.892 = 0.048$$

### Step 3: Choose the best feature to split

Step 3: Choose the Feature with the Highest Information Gain

The information gains for each feature are:

- Outlook: 0.247
- Temperature: 0.029
- Humidity: 0.151
- Wind: 0.048

Since **Outlook** has the highest information gain (**0.247**), we choose **Outlook** as the first feature to split on.

### Conclusion:

- The first split should be on **Outlook** because it provides the highest information gain.

The feature with the highest information gain is "Outlook" with an information gain of **0.247**.



## Step 4: Second Split (Depth 2)

We now split based on the "Outlook" feature:

- **Overcast:** All outcomes are "yes", so this becomes a leaf node.
- **Sunny** and **Rain** require further splitting. We now compute the information gain for these subsets based on the remaining features.

To decide which feature to split on within the "**Sunny**" subset, we need to calculate the **information gain** for each of the remaining features: **Temperature**, **Humidity**, and **Wind**, and choose the feature with the highest information gain.

### Step 1: Subset of data where Outlook is Sunny

From the table, when **Outlook** is "Sunny", we have the following data:

Day	Outlook	Temperature	Humidity	Wind	Played tennis
1	Sunny	Hot	High	Weak	no
2	Sunny	Hot	High	Strong	no
8	Sunny	Cool	High	Weak	yes
9	Sunny	Mild	Normal	Weak	yes
11	Sunny	Mild	High	Strong	no

### In this subset:

- Total examples = 5
- Positive examples (yes) = 2
- Negative examples (no) = 3

## Step 2: Calculate the entropy of the "Sunny" subset

The entropy for the subset where **Outlook** is "Sunny" is:

$$H(\text{Sunny}) = - \left( \frac{2}{5} \right) \log_2 \left( \frac{2}{5} \right) - \left( \frac{3}{5} \right) \log_2 \left( \frac{3}{5} \right)$$

$$H(\text{Sunny}) = -0.4 \log_2(0.4) - 0.6 \log_2(0.6) = 0.971$$

## Step 3: Calculate information gain for each remaining feature

We will now calculate the information gain for **Temperature**, **Humidity**, and **Wind** within the "Sunny" subset.

### Split by **Temperature**:

The possible values for **Temperature** are: Hot, Mild, Cool.

- **Hot**: 2 examples (0 yes, 2 no)
- **Mild**: 2 examples (1 yes, 1 no)
- **Cool**: 1 example (1 yes, 0 no)

The entropy for each subset:

- $H(\text{Hot}) = - \left(\frac{0}{2}\right) \log_2 \left(\frac{0}{2}\right) - \left(\frac{2}{2}\right) \log_2 \left(\frac{2}{2}\right) = 0$  (since both are "no")
- $H(\text{Mild}) = - \left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) = 1.0$
- $H(\text{Cool}) = - \left(\frac{1}{1}\right) \log_2 \left(\frac{1}{1}\right) = 0$  (since it's all "yes")

The weighted average entropy for the split on **Temperature** is:

$$H(\text{Temperature}) = \frac{2}{5} \times 0 + \frac{2}{5} \times 1.0 + \frac{1}{5} \times 0 = 0.4$$

**Information Gain for Temperature:**

$$IG(\text{Temperature}) = H(\text{Sunny}) - H(\text{Temperature}) = 0.971 - 0.4 = 0.571$$

**Split by Humidity:**

The possible values for **Humidity** are: High, Normal.

- **High:** 3 examples (1 yes, 2 no)
- **Normal:** 2 examples (1 yes, 0 no)

The entropy for each subset:

- $H(\text{High}) = -\left(\frac{1}{3}\right) \log_2 \left(\frac{1}{3}\right) - \left(\frac{2}{3}\right) \log_2 \left(\frac{2}{3}\right) = 0.918$
- $H(\text{Normal}) = -\left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) = 0$

The weighted average entropy for the split on **Humidity** is:

$$H(\text{Humidity}) = \frac{3}{5} \times 0.918 + \frac{2}{5} \times 0 = 0.550$$

**Information Gain for Humidity:**

$$IG(\text{Humidity}) = H(\text{Sunny}) - H(\text{Humidity}) = 0.971 - 0.550 = 0.421$$

**Split by Wind:**

The possible values for **Wind** are: Weak, Strong.

- **Weak:** 3 examples (2 yes, 1 no)
- **Strong:** 2 examples (0 yes, 2 no)

The entropy for each subset:

- $H(\text{Weak}) = -\left(\frac{2}{3}\right) \log_2 \left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) \log_2 \left(\frac{1}{3}\right) = 0.918$
- $H(\text{Strong}) = -\left(\frac{0}{2}\right) \log_2 \left(\frac{0}{2}\right) - \left(\frac{2}{2}\right) \log_2 \left(\frac{2}{2}\right) = 0$

The weighted average entropy for the split on **Wind** is:

$$H(\text{Wind}) = \frac{3}{5} \times 0.918 + \frac{2}{5} \times 0 = 0.550$$

Information Gain for Wind:

$$IG(\text{Wind}) = H(\text{Sunny}) - H(\text{Wind}) = 0.971 - 0.550 = 0.421$$

**Step 4: Choose the best feature to split**

**The information gains are:**

- **Temperature:** 0.571
- **Humidity:** 0.421
- **Wind:** 0.421

Since **Temperature** has the highest information gain (0.571), we will split the "**Sunny**" subset based on **Temperature**.

Once we decide to split the **Sunny** subset on **Temperature**, the next step is to determine the outcome for each branch (i.e., decide what the leaf nodes will be). The process for determining the leaf nodes is based on the distribution of the target variable ("Played tennis") within each branch of the **Temperature** split.

## Sunny Subset

The **Sunny** subset contains the following data:

Day	Outlook	Temperature	Humidity	Wind	Played tennis
1	Sunny	Hot	High	Weak	no
2	Sunny	Hot	High	Strong	no
8	Sunny	Cool	High	Weak	yes
9	Sunny	Mild	Normal	Weak	yes
11	Sunny	Mild	High	Strong	no

From this subset, we split on **Temperature**, which has three possible values: **Hot**, **Mild**, and **Cool**.

## Step 1: Evaluate Each Temperature Branch

- **Hot:**
  - Days: 1, 2
  - Outcome: "no", "no"
  - Since all outcomes for **Hot** are "no", this branch will be a **leaf node** labeled "no".
- **Cool:**
  - Day: 8
  - Outcome: "yes"
  - Since all outcomes for **Cool** are "yes", this branch will be a **leaf node** labeled "yes".
- **Mild:**
  - Days: 9, 11
  - Outcome: "yes", "no"
  - Here, we have a mix of "yes" and "no" outcomes (one "yes" and one "no").
  - Since the tree depth is limited to 2, we cannot split further. In this case, the common approach is to label the leaf node with the majority class. In the **Mild** branch, there is 1 "yes" and 1 "no", so there is no majority class.

## Step 2: Resolve Ties in the Mild Branch

Since the **Mild** branch has equal occurrences of "yes" and "no", you can handle this tie in several ways:

1. **Use the parent node's majority class:** In this case, the parent node (Sunny subset) has 3 "no" outcomes and 2 "yes" outcomes, so you could label the **Mild** branch as "no" (following the majority class from the Sunny subset).
2. **Randomly assign a label:** This approach would assign either "yes" or "no" arbitrarily.
3. **Consider external factors:** If more information is available, you might use that to resolve the tie.

For this example, we will follow the **first approach** and label the **Mild** branch as "**no**".



To decide what to split on for the **Rainy** branch, we follow the same process of calculating **information gain** for each of the remaining features—**Temperature**, **Humidity**, and **Wind**—using only the examples where the **Outlook** is "Rain".

## Rainy Subset

From the dataset, the examples where **Outlook** is "Rain" are:

Day	Outlook	Temperature	Humidity	Wind	Played tennis
4	Rain	Mild	High	Weak	yes
5	Rain	Mild	Normal	Weak	yes
6	Rain	Cool	Normal	Strong	no
10	Rain	Mild	Normal	Weak	yes
14	Rain	Mild	High	Strong	no

## In the **Rainy** subset:

- Total examples = 5
- Positive examples (yes) = 3
- Negative examples (no) = 2

## Step 1: Calculate the entropy of the Rainy subset

The entropy for the **Rainy** subset is:

$$H(\text{Rain}) = - \left( \frac{3}{5} \right) \log_2 \left( \frac{3}{5} \right) - \left( \frac{2}{5} \right) \log_2 \left( \frac{2}{5} \right)$$

$$H(\text{Rain}) = -0.6 \log_2(0.6) - 0.4 \log_2(0.4) = 0.971$$

## Step 2: Calculate the information gain for each remaining feature

We now calculate the information gain for **Temperature**, **Humidity**, and **Wind** in the Rainy subset.

### 1. Information Gain for "Temperature"

The possible values for **Temperature** are: Mild, Cool.

- **Mild**: 4 examples (3 yes, 1 no)
- **Cool**: 1 example (0 yes, 1 no)

### The entropy for each subset:

- $H(\text{Mild}) = - \left( \frac{3}{4} \right) \log_2 \left( \frac{3}{4} \right) - \left( \frac{1}{4} \right) \log_2 \left( \frac{1}{4} \right) = 0.811$
- $H(\text{Cool}) = 0$  (since all examples are "no")

The weighted average entropy for the split on **Temperature** is:

$$H(\text{Temperature}) = \frac{4}{5} \times 0.811 + \frac{1}{5} \times 0 = 0.649$$

**Information Gain for Temperature:**

$$IG(\text{Temperature}) = H(\text{Rain}) - H(\text{Temperature}) = 0.971 - 0.649 = 0.322$$

## 2. Information Gain for "Humidity"

The possible values for **Humidity** are: High, Normal.

- **High:** 2 examples (0 yes, 2 no)
- **Normal:** 3 examples (3 yes, 0 no)

## The entropy for each subset:

- $H(\text{High}) = 0$  (since both examples are "no")
- $H(\text{Normal}) = 0$  (since all examples are "yes")

The weighted average entropy for the split on **Humidity** is:

$$H(\text{Humidity}) = \frac{2}{5} \times 0 + \frac{3}{5} \times 0 = 0$$

**Information Gain for Humidity:**

$$IG(\text{Humidity}) = H(\text{Rain}) - H(\text{Humidity}) = 0.971 - 0 = 0.971$$

### 3. Information Gain for "Wind"

The possible values for **Wind** are: Weak, Strong.

- **Weak:** 3 examples (3 yes, 0 no)
- **Strong:** 2 examples (0 yes, 2 no)

### The entropy for each subset:

- $H(\text{Weak}) = 0$  (since all examples are "yes")
- $H(\text{Strong}) = 0$  (since both examples are "no")

The weighted average entropy for the split on **Wind** is:

$$H(\text{Wind}) = \frac{3}{5} \times 0 + \frac{2}{5} \times 0 = 0$$

Information Gain for Wind:

$$IG(\text{Wind}) = H(\text{Rain}) - H(\text{Wind}) = 0.971 - 0 = 0.971$$

Step 3: Choose the feature with the highest information gain

The information gains for each feature in the Rainy subset are:

- Temperature: 0.322
- Humidity: 0.971
- Wind: 0.971


Both **Humidity** and **Wind** have the highest information gain of **0.971**. In cases where multiple features have the same information gain, you can choose any of them to split on. For simplicity, we'll split on **Wind**.

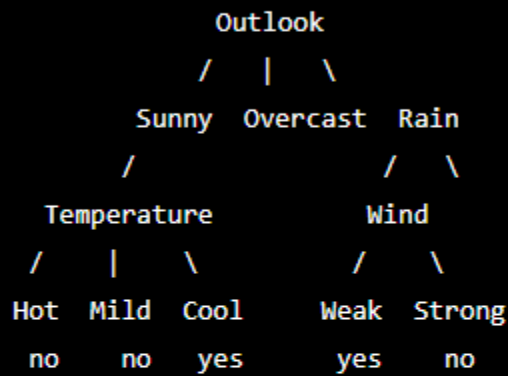
## Step 4: Determine the leaf nodes after splitting on Wind

After splitting on **Wind**, we look at the outcomes for each branch:

- **Weak:**
  - Days: 4, 5, 10
  - Outcomes: "yes", "yes", "yes"
  - Since all outcomes for **Weak** are "yes", this branch becomes a **leaf node** labeled "yes".
- **Strong:**
  - Days: 6, 14
  - Outcomes: "no", "no"
  - Since all outcomes for **Strong** are "no", this branch becomes a **leaf node** labeled "no".

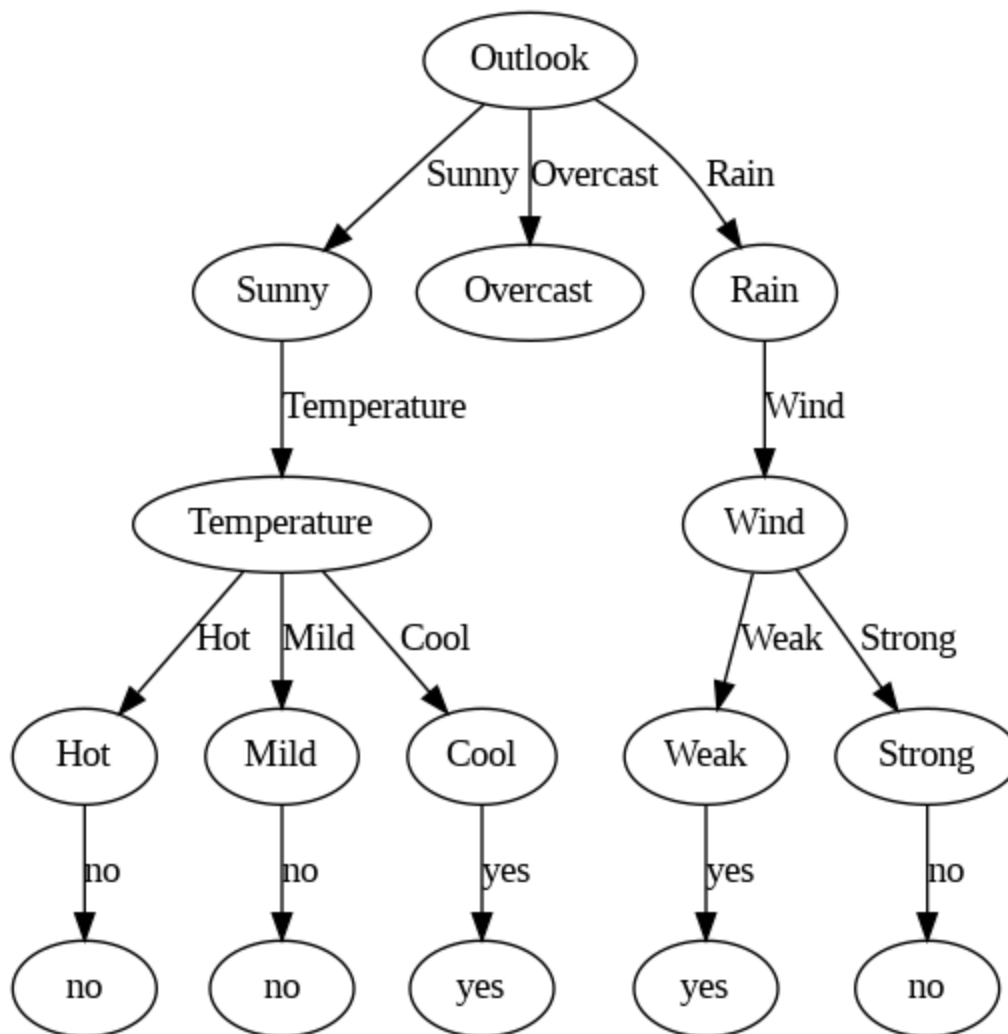
## Final Decision Tree (Depth 2):

 Copy code



## Conclusion:

- For the **Sunny** branch, we split on **Temperature**, and the leaf nodes are determined by the majority class (or tie-breaking rule for Mild).
- For the **Rainy** branch, we split on **Wind**, and the leaf nodes are determined by the uniform class in each branch (all "yes" for Weak and all "no" for Strong).





# Cosine Similarity

**2. [8 points]** Calculate the cosine similarity between two documents. Cosine similarity is defined as the dot product of the vectors divided by their magnitude.

$$\text{cosine\_similarity}(A, B) = \cos \theta = \frac{A \cdot B}{\|A\| \|B\|}$$

where  $\theta$  is the angle between the two vectors  $A$  and  $B$  and  $A \cdot B$  is the dot product of the two vectors  $A$  and  $B$ , the dot product can be calculated as  $A \cdot B = \sum_{i=1}^n A_i B_i = [A_1 B_1 + A_2 B_2 + \dots + A_n B_n]$ .  $A_1, A_2, \dots, A_n$  and  $B_1, B_2, \dots, B_n$  are the components of the vectors  $A$  and  $B$ .  $\|A\|$  represents the magnitude of the vectors and can be calculated as  $\|A\| = \sqrt{A_1^2 + A_2^2 + \dots + A_n^2}$ .

For this assignment, you would need to first calculate the document vectors manually from the document pairs and then calculate the cosine similarity. The document pairs are:

i). *"Today is a nice day. I want to go for a walk."* and *"The man is taking his dog for a walk as the weather is nice today"*. The set of words for this problem is [today, nice, day, want, go, walk, weather, man, dog, take] Consider 'taking' to be the same word as 'take' for constructing the document vector since they are from the same root word.

To calculate the **cosine similarity** between the two documents, follow these steps:

## Step 1: Create the document vectors

First, we need to create vectors for the two documents based on the words in the set:

**Word set:** [today, nice, day, want, go, walk, weather, man, dog, take]

Each document will be represented by a vector, where each element corresponds to the frequency of the respective word from the word set in the document.

**Document 1:** "Today is a nice day. I want to go for a walk."

1. today: 1
2. nice: 1
3. day: 1
4. want: 1
5. go: 1
6. walk: 1
7. weather: 0
8. man: 0
9. dog: 0
10. take: 0 (taking and take are considered the same, but no occurrence here)

So, the vector for Document 1 is:

$$A = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]$$

Document 2: "The man is taking his dog for a walk as the weather is nice today."

1. today: 1
2. nice: 1
3. day: 0
4. want: 0
5. go: 0
6. walk: 1
7. weather: 1
8. man: 1
9. dog: 1
10. take: 1 (since "taking" is considered as "take")

So, the vector for Document 2 is:

$$B = [1, 1, 0, 0, 0, 1, 1, 1, 1, 1]$$

## Step 2: Calculate the dot product $A \cdot B$

The dot product is calculated by multiplying corresponding elements of the two vectors and summing the results:

$$\begin{aligned} A \cdot B &= (1 \times 1) + (1 \times 1) + (1 \times 0) + (1 \times 0) + (1 \times 0) \\ &\quad + (1 \times 1) + (0 \times 1) + (0 \times 1) + (0 \times 1) + (0 \times 1) \end{aligned}$$

$$A \cdot B = 1 + 1 + 0 + 0 + 0 + 1 + 0 + 0 + 0 + 0 = 3$$

### Step 3: Calculate the magnitude of vectors $A$ and $B$

The magnitude of a vector  $A$  is calculated as:

$$||A|| = \sqrt{A_1^2 + A_2^2 + \cdots + A_n^2}$$

For  $A = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]$ :

$$||A|| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2} = \sqrt{6} = 2.449$$

For  $B = [1, 1, 0, 0, 0, 1, 1, 1, 1, 1]$ :

$$||B|| = \sqrt{1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{7} = 2.646$$

### Step 4: Calculate cosine similarity

Now, we use the formula for cosine similarity:

$$\text{Cosine Similarity} = \frac{A \cdot B}{||A|| \times ||B||}$$

Substitute the values:

$$\text{Cosine Similarity} = \frac{3}{2.449 \times 2.646} = \frac{3}{6.478} = 0.463$$

## Final Result:

The **cosine similarity** between the two documents is approximately **0.463**.

ii). *"The weather is dull today"* and *"How are you doing? Are you going to the college"* Consider the set of words to be [today, weather, dull, go, college, do]. Consider 'going' to be the same word as 'go' for constructing the document vector since they are from the same root word. Similarly, consider 'doing' to be the same word as 'do' for constructing the document vector.

To calculate the cosine similarity between these two documents, we'll first follow the same process as before: construct the document vectors, calculate the dot product, and finally compute the cosine similarity.

## Documents:

1. **Document 1:** "The weather is dull today"
2. **Document 2:** "How are you doing? Are you going to the college"

## Step 1: Create the document vectors

**Word set:** [today, weather, dull, go, college, do]

## Step 1: Create the document vectors

Word set: [today, weather, dull, go, college, do]

Document 1: "The weather is dull today"

1. today: 1
2. weather: 1
3. dull: 1
4. go: 0
5. college: 0
6. do: 0 (since "doing" and "do" are considered the same word)

So, the vector for Document 1 is:

$$A = [1, 1, 1, 0, 0, 0]$$

Document 2: "How are you doing? Are you going to the college"

1. today: 0
2. weather: 0
3. dull: 0
4. go: 1 (since "going" is considered as "go")
5. college: 1
6. do: 1 (since "doing" and "do" are considered the same word)

So, the vector for Document 2 is:

$$B = [0, 0, 0, 1, 1, 1]$$

## Step 2: Calculate the dot product $A \cdot B$

The dot product is calculated by multiplying corresponding elements of the two vectors and summing the results:

$$A \cdot B = (1 \times 0) + (1 \times 0) + (1 \times 0) + (0 \times 1) + (0 \times 1) + (0 \times 1)$$

$$A \cdot B = 0 + 0 + 0 + 0 + 0 + 0 = 0$$

## Step 3: Calculate the magnitude of vectors $A$ and $B$

For  $A = [1, 1, 1, 0, 0, 0]$ :

$$||A|| = \sqrt{1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2} = \sqrt{3} = 1.732$$

For  $B = [0, 0, 0, 1, 1, 1]$ :

$$||B|| = \sqrt{0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2} = \sqrt{3} = 1.732$$

#### Step 4: Calculate cosine similarity

Now, we use the formula for cosine similarity:

$$\text{Cosine Similarity} = \frac{A \cdot B}{||A|| \times ||B||}$$

Substitute the values:

$$\text{Cosine Similarity} = \frac{0}{1.732 \times 1.732} = \frac{0}{3} = 0$$

#### Final Result:

The **cosine similarity** between the two documents is **0**.

Since the two documents have no common words from the set, the dot product is 0, and thus the cosine similarity is 0, meaning the documents are orthogonal or dissimilar in terms of the word set provided.



To calculate the **Simple Matching Coefficient (SMC)** and **Jaccard Coefficient** between two binary variables  $P$  and  $Q$ , let's first define how these metrics work.

## 1. Simple Matching Coefficient (SMC)

The **Simple Matching Coefficient** measures the proportion of matches (both 1s and 0s) between two binary variables.

Formula:

$$\text{SMC} = \frac{(a + d)}{(a + b + c + d)}$$

Where:

- $a$ : The number of times both  $P$  and  $Q$  are 1 (True Positive)
- $d$ : The number of times both  $P$  and  $Q$  are 0 (True Negative)
- $b$ : The number of times  $P$  is 1 and  $Q$  is 0 (False Negative)
- $c$ : The number of times  $P$  is 0 and  $Q$  is 1 (False Positive)

## 2. Jaccard Coefficient

The **Jaccard Coefficient** measures the similarity between two sets, focusing only on the 1s.

Formula:

$$\text{Jaccard} = \frac{a}{(a + b + c)}$$

Where:

- $a$ : The number of times both  $P$  and  $Q$  are 1 (True Positive)
- $b$ : The number of times  $P$  is 1 and  $Q$  is 0 (False Negative)
- $c$ : The number of times  $P$  is 0 and  $Q$  is 1 (False Positive)

Values from  $P = (1, 1, 1, 1, 0, 1)$  and  $Q = (1, 0, 0, 1, 1, 0)$ :

$P$	$Q$	Matches (1-1 and 0-0)
1	1	$a = 1$
1	0	$b = 1$
1	0	$b = 1$
1	1	$a = 1$
0	1	$c = 1$
1	0	$b = 1$

Counts:

- $a = 2$  (both P and Q are 1)
- $b = 3$  (P is 1, Q is 0)
- $c = 1$  (P is 0, Q is 1)
- $d = 0$  (both P and Q are 0)

### Simple Matching Coefficient (SMC):

$$\text{SMC} = \frac{(a + d)}{(a + b + c + d)} = \frac{(2 + 0)}{(2 + 3 + 1 + 0)} = \frac{2}{6} = 0.333$$

### Jaccard Coefficient:

$$\text{Jaccard} = \frac{a}{(a + b + c)} = \frac{2}{(2 + 3 + 1)} = \frac{2}{6} = 0.333$$

### Conclusion:

- Simple Matching Coefficient (SMC) = 0.333
- Jaccard Coefficient = 0.333

Both coefficients result in the same value in this case because there are no cases where both  $P$  and  $Q$  are 0, thus no true negatives.