PART A:

1. Euclidean distance

$$\text{dist}_{\text{euclid}}(d_1, d_2) = \sqrt{\sum_{i=1}^{n} (w_{1,i} - w_{2,i})^2}$$

$$\text{sim}(d_1, d_2) = \frac{1}{1 + \text{dist}(d_1, d_2)}$$

2. Dot product

$$sim(d_1, d_2) = \vec{d_1} \cdot \vec{d_2} = \sum_{i=1}^{n} w_{1,i} \cdot w_{2,i}$$

3. Cosine similarity

$$\text{sim}_{\cos}(d_1, d_2) = \cos\varphi = \frac{\vec{d_1} \cdot \vec{d_2}}{|\vec{d_1}||\vec{d_2}|} = \frac{\sum_{i=1}^{n} w_{1,i} \cdot w_{2,i}}{\sqrt{\left(\sum_{i=1}^{n} w_{1,i}^2\right) \cdot \left(\sum_{i=1}^{n} w_{2,i}^2\right)}}$$

- We construct a tf-idf matrix that consists of weights of all the terms across the document collection which is then used to calculate similarities.
- Using the above formulas, we calculate the similarities between document1 ("Today is sunny.") and document2("She is a sunny girl").

OUTPUT:

```
------PART A------
TF-IDF matrix:

    [D1       D2        D3        D4        D5        D6]

a = [0.0, 0.7781512503836436, 0.0, 0.0, 0.0, 0.0, 0.0]
always = [0.0, 0.0, 0.0, 0.0, 0.0, 0.7781512503836436, 0.0]
be = [0.0, 0.0, 1.5563025007672873, 0.0, 0.0, 0.0, 0.0]
berlin = [0.0, 0.0, 0.0, 0.3010299956639812, 0.3010299956639812,
0.3010299956639812, 0.0, 0.0, 0.0]
exciting = [0.0, 0.0, 0.0, 0.0, 0.0, 0.7781512503836436, 0.0]
girl = [0.0, 0.7781512503836436, 0.0, 0.0, 0.0, 0.0, 0.0]
in = [0.0, 0.0, 0.0, 0.7781512503836436, 0.0, 0.0, 0.0]
is = [0.17609125905568124, 0.17609125905568124, 0.0, 0.17609125905568124,
0.0, 0.17609125905568124, 0.0, 0.0, 0.0, 0.0]
not = [0.0, 0.0, 0.7781512503836436, 0.0, 0.0, 0.0, 0.0]
or = [0.0, 0.0, 0.7781512503836436, 0.0, 0.0, 0.0, 0.0]
she = [0.0, 0.47712125471966244, 0.0, 0.47712125471966244, 0.0, 0.0, 0.0,
0.0]
sunny = [0.3010299956639812, 0.3010299956639812, 0.0, 0.0,
0.3010299956639812, 0.0, 0.0, 0.0, 0.0]
to = [0.0, 0.0, 1.5563025007672873, 0.0, 0.0, 0.0, 0.0]
today = [0.47712125471966244, 0.0, 0.0, 0.47712125471966244, 0.0, 0.0, 0.0,
0.0]

Vector representation of Document 1: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.17609125905568124, 0.0, 0.0, 0.0, 0.3010299956639812, 0.0,
0.47712125471966244]
```

```
Vector representation of Document 2: [0.7781512503836436, 0.0, 0.0, 0.0,
0.0, 0.7781512503836436, 0.0, 0.17609125905568124, 0.0, 0.0,
0.47712125471966244, 0.3010299956639812, 0.0, 0.0]

Similarity between Document 1 and Document 2 using Euclidean distance =
0.4365166571371468
Similarity between Document 1 and Document 2 using Dot product =
0.12162718980527158
Similarity between Document 1 and Document 2 using cosine similarity =
0.16475679254915546
```

PART B

- For the given query ("She is a sunny girl."), we use Vector Space model and BM25 model to calculate the scores to find relevant documents.
- BM25 takes into account both term frequency (TF) and document length normalization to determine the relevance of a document to a given query.
- We can see that in both the cases, document1 has the highest score though different model generated different score.

OUTPUT:

```
------PART B------

query : 'She is a sunny girl.'

Scores using Vector Space Model:
4.136239 :  She is a sunny girl
1.6720572 : Today is sunny
1.4238253 :  She is in Berlin today
1.1028148 :  Sunny Berlin
0.66823614 :  Berlin is always exciting

Scores using BM25 Model :
4.8487716 :  She is a sunny girl
1.3601658 :  She is in Berlin today
1.2818048 : Today is sunny
0.88044095 :  Sunny Berlin
0.44918302 :  Berlin is always exciting
```