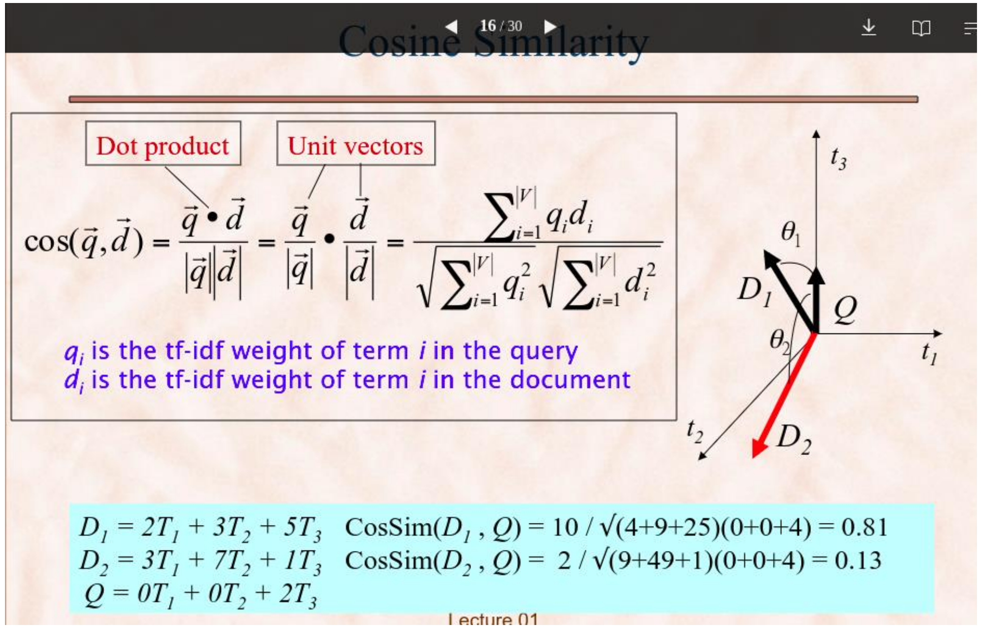
Assignement 2

Hypothesis

Find the query : "killed people garden" into the 200 scripts

CALCULS



16 / 30

Cosine Similarity

Dot product

Unit vectors

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|} \bullet \frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

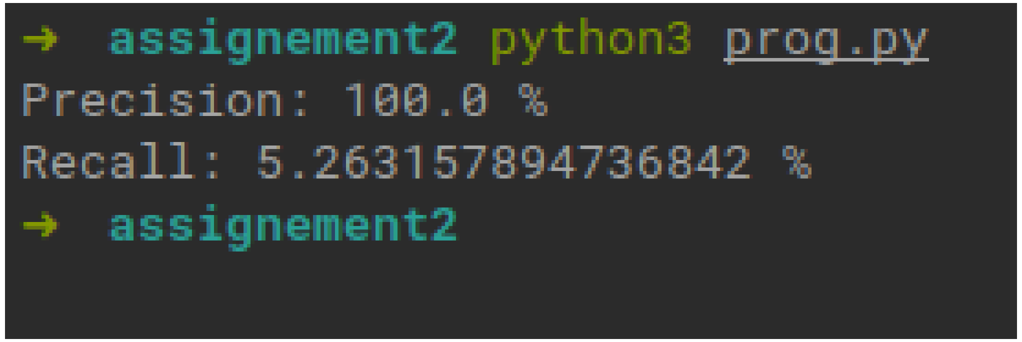
q_i is the tf-idf weight of term i in the query
d_i is the tf-idf weight of term i in the document

$D_1 = 2T_1 + 3T_2 + 5T_3$ CosSim(D_1 , Q) = 10 / √(4+9+25)(0+0+4) = 0.81
$D_2 = 3T_1 + 7T_2 + 1T_3$ CosSim(D_2 , Q) = 2 / √(9+49+1)(0+0+4) = 0.13
$Q = 0T_1 + 0T_2 + 2T_3$

Lecture 01

- I did the same base as the first assignement.
- I calculated the TF so: 1 + log(the number of occurence of the word)
- Then I calculated the IDF: log2(total documents / documents where the word is)
- I multiplied IDF x TF to have the weight: so each IDF per word for the word with TF value
- I calculated the distance of the document by doing the calcul on the class pdf . So we can score the document.

CALCULS



```
➜  assignement2 python3 prog.py
Precision: 100.0 %
Recall: 5.263157894736842 %
➜  assignement2
```

- I calculated the precision and the recall:
- Precision = (tp / (tp +fp)) *100 (convert in %)
- Recall = (tp/ (tp + fn)) * 100 (convert in %)
- So as retrieval documents I set at the value of 10.
- For the precision it means on 10 documents, all the words of the query were found.
- For the recall, it means on 200 scripts we took ~5,26% of the total documents to search the query