CS-E4850 Computer Vision, Answers to Exercise Round 7

Adam Ilyas 725819

November 8, 2018

Exercise 1. Comparing bags-of-words with tf-idf weighting.

Assume that we have an indexed collection of documents containing the five terms of the following table, where the second row indicates the percentage of documents in which each term appears.

term	cat	dog	mammals	mouse	pet
% of documents	5	20	2	10	60

Now, given the query Q={mouse, cat, pet, mammals}, compute the similarity between Q and the following documents D1, D2, D3, by using the cosine similarity measure and tf-idf weights (i.e. term frequency - inverse document frequency) for the bag-of-words histogram representations of the documents and the query.

- D1={Cat is a pet, dog is a pet, and mouse may be a pet too.}
- D2={Cat, dog and mouse are all mammals.}
- D3={Cat and dog get along well, but cat may eat a mouse.}

Ignore other words except the five terms. You may proceed with the following steps:

a) Compute and report the inverse document frequency (idf) for each of the five terms. Use the logarithm with base 2. (idf is the logarithm on slide 69 of Lecture 6.)

$$idf = \log \frac{N}{n_i} = \log \frac{\text{total no of doc in database}}{\text{no of doc containing word i}}$$

Inverse Document Frequency:

'cat': 4.321928094887363,

'dog': 2.321928094887362,

'mammals': 5.643856189774724,

'mouse': 3.321928094887362,

'pet': 0.7369655941662062

b) Compute the term frequencies for the query and each document.

Term Frequency:

Query

'cat': 0.25, 'mouse': 0.25, 'mammals': 0.25, 'pet': 0.25

Document 1

'is': 0.133, 'and': 0.067, 'too': 0.067, 'may': 0.067, 'dog': 0.067, 'pet': 0.2, 'a': 0.2 , 'mouse': 0.067, 'be': 0.067, 'cat': 0.067

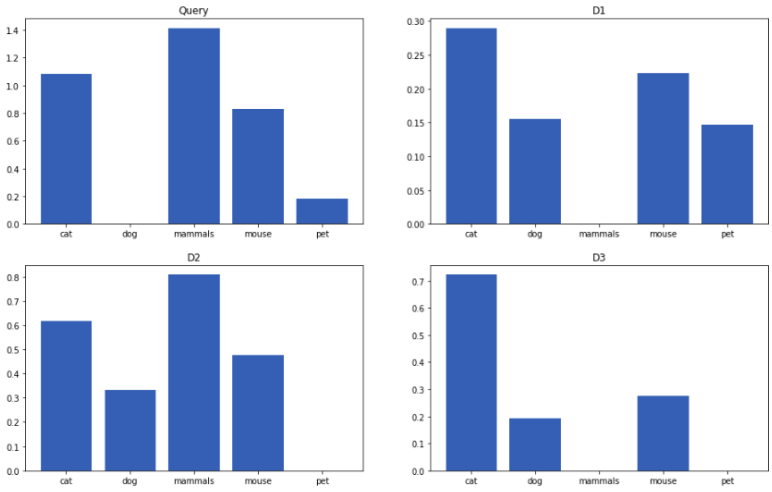
Document 2

'and': 0.143, 'dog': 0.143, 'mouse': 0.143, 'all': 0.143, 'mammals': 0.143, 'are': 0.143, 'cat': 0.143

Document 3

'and': 0.083, 'well': 0.083, 'may': 0.083, 'dog': 0.083, 'mouse': 0.083, 'a': 0.083, 'eat': 0.083, 'get': 0.083, 'but': 0.083, 'along': 0.083, 'cat': 0.167

c) Form the tf-idf weighted word occurrence histograms for the query and documents.



d) Evaluate the cosine similarity between the query and each document (i.e. normalized scalar product between the weighted occurrence histograms as shown on slide 45).

$$sim(d_j, q) = \frac{\langle d_j, q \rangle}{||d_j|| ||q||} = \frac{\sum_{i=1}^{V} d_j(i) \times q(i)}{\sqrt{\sum_{i=1}^{V} d_j(i)^2} \sqrt{\sum_{i=1}^{V} q(i)^2}}$$

similarity(Query, Document 1): 0.6291036970635482

similarity(Query, Document 2): 0.9546948111493485

similarity(Query, Document 3): 0.6430077230767424

e) Report the relative ranking of the documents. (You should get similarities 0.95, 0.64, and 0.63, but you need to determine which corresponds to which document.)

Document 2: 0.95, Document 3: 0.64, Document 1: 0.63.

Document 2 is most similar. Document 1 is least similar.

Exercise 2. Precision and recall. (pen & paper problem)

There is a database of 10000 images and a user, who is only interested in images which contain a car. It is known that there are 500 such images in the database. An automatic image retrieval system retrieves 300 car images and 50 other images from the database. Determine and report the precision and recall of the retrieval system in this particular case.

precision = #relevant / #returned
is number of correct images / number of images returned,

recall = #relevant / #total relevant
is number of correct images / number of total relevant images in db

Ans. precision = 300/350 = 0.857, recall = 300/500 = 0.600