COMP 472: Artificial Intelligence

Bag of Words Model

Solutions

Question 1 Assume the following query:

cheap CD very cheap DVD

(a) Compute the bag of word representation of the query above using term frequency as weight.

	value
cheap	2
CD	1
DVD	1
very	1

(b) Further assume that 2 documents d1 and d2 are represented using the following BOW representation:

	cheap	CD	DVD	very	software	bugs
d1	2	2	0	0	1	5
d2	1	0	1	0	0	4

Compute the cosine similarity between the query & d1, and between the query & d2. Which of the 2 documents is closer to the query.

The formula for the cosine similarity is:

$$cos(\overrightarrow{D}, \overrightarrow{Q}) = \frac{\overrightarrow{D} \cdot \overrightarrow{Q}}{|\overrightarrow{D}| \cdot |\overrightarrow{Q}|} = \frac{\Sigma_{i=1}^{N} d_i q_i}{\sqrt{\Sigma_{i=1}^{N} d_i^2} \sqrt{\Sigma_{i=1}^{N} q_i^2}}$$

Let's calculate the cosine similarity between the query & d1, given that $query = (2, 1, 1, 1, 0, 0)$ *and* $d1 = (2, 2, 0, 0, 1, 5)$.

$$cos(\overrightarrow{query}, \overrightarrow{d1}) = \frac{\Sigma_{i=1}^{N} d_i q_i}{\sqrt{\Sigma_{i=1}^{N} d_i^2} \sqrt{\Sigma_{i=1}^{N} q_i^2}} = \frac{6}{\sqrt{7} \times \sqrt{34}} = \frac{6}{15.4247} = 0.3889$$

Now, it's time to do the same calculation for the query & d2 given that $query = (2, 1, 1, 1, 0, 0)$, *and* $d2 = (1, 0, 1, 0, 0, 4)$.

$$cos(\overrightarrow{query}, \overrightarrow{d2}) = \frac{\Sigma_{i=1}^{N} d_i q_i}{\sqrt{\Sigma_{i=1}^{N} d_i^2} \sqrt{\Sigma_{i=1}^{N} q_i^2}} = \frac{3}{\sqrt{7} \times \sqrt{18}} = \frac{3}{11.25} = 0.2672$$

The Cosine similarity of d1 the query is greater that d2 & the query which means that d1 is closer to the query than d2.

Question 2 Assume that you have the following sentences about the AI class:

d1 : *It was the best of times*
d2 : *It was the worst of times*
d3 : *It was the age of wisdom*
d4 : *It was the best of the best*

(a) Write a Python program that computes the bag of word representation of the sentences above using term frequency as weight.

```
# Scikit Learn
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd

# Create the Document Term Matrix
count_vectorizer = CountVectorizer(stop_words='english')
count_vectorizer = CountVectorizer()
sparse_matrix = count_vectorizer.fit_transform(documents)

# Convert Sparse Matrix to Pandas Dataframe to see the word frequencies.
doc_term_matrix = sparse_matrix.todense()
df = pd.DataFrame(doc_term_matrix,
                  columns=count_vectorizer.get_feature_names(),
                  index=['d1', 'd2', 'd3', 'd4'])
```

```
df
```

	age	best	it	of	the	times	was	wisdom	worst
d1	0	1	1	1	1	1	1	0	0
d2	0	0	1	1	1	1	1	0	1
d3	1	0	1	1	1	0	1	1	0
d4	0	2	1	1	2	0	1	0	0

(b) Compute the cosine similarity between all pairs of documents and print them out.

```
# Compute Cosine Similarity
from sklearn.metrics.pairwise import cosine_similarity
print(cosine_similarity(df, df))
```

```
[[1.         0.83333333 0.66666667 0.86164044]
 [0.83333333 1.         0.66666667 0.61545745]
 [0.66666667 0.66666667 1.         0.61545745]
 [0.86164044 0.61545745 0.61545745 1.        ]]
```

(c) Which 2 documents are closer to each other?
The values of the cosine similarity matrix above, clearly shows that the pair d1 & d4 are the closest documents.

(d) Use your Python code to verify your answer to question 1b above.

```
df
```

	bugs	cd	cheap	dvd	software	very
query	0	1	2	1	0	1
d1	5	2	2	0	1	0
d2	4	0	1	1	0	0

```python
# Compute Cosine Similarity
from sklearn.metrics.pairwise import cosine_similarity
print(cosine_similarity(df, df))
```

```
[[1.         0.38892223 0.26726124]
 [0.38892223 1.         0.88929729]
 [0.26726124 0.88929729 1.        ]]
```

The values of the cosine similarity matrix above shows that d1 is closer to the query than d2.