IT-309
EXPERIMENT 8 REPORT

Indian Institute of Information Technology, Vadodara

Group-1

Contents

0.1 Experiment -8 : Latent Semantic Indexing

0.1.1 Theory

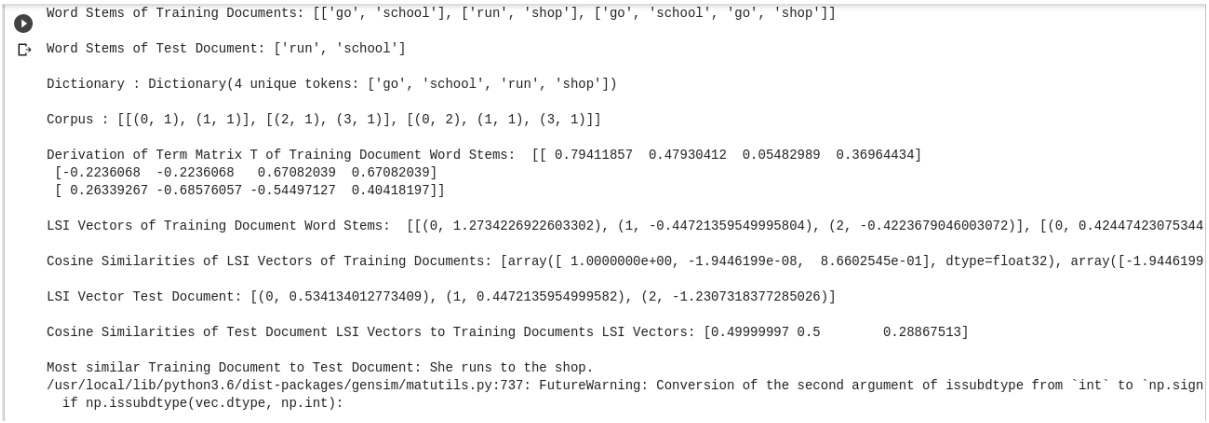
- Latent Semantic Indexing is a process in which The term-document matrix is approximated to a lower rank so that the similarity scores between documents are more precise.
- This is done to account for two discrepancies in calculating the similarity scores - Synonyms and Polysemy.
- Synonyms are words that have the same meaning. This underestimates the true similarity that a user would perceive.
- Polysemy is when a single word has two different meanings as per context. this overestimates the similarity.

0.1.2 Experiment

- The code has been given three training "documents" which are three sentences. and a test document which is a single sentence.
- Input(Training):
  - *"She goes to school"*
  - *"She runs to the shop"*
  - *"I go to school and he goes to the shop"*
- Input(Testing):
  - *"He runs to school"*
- The output given are:
  - Word stems
  - dictionary
  - Corpus
  - Term matrix
  - The LSI Vectors of the training documents

  - the cosine similarities
  - LSI vector of test document
  - Cosine similarities of training and testing document LSI vectors
  - Most similar training document to testing document.

- Sorting the keys according to the values in descending order gives us the rank



```
Word Stems of Training Documents: [['go', 'school'], ['run', 'shop'], ['go', 'school', 'go', 'shop']]

Word Stems of Test Document: ['run', 'school']

Dictionary : Dictionary(4 unique tokens: ['go', 'school', 'run', 'shop'])

Corpus : [[(0, 1), (1, 1)], [(2, 1), (3, 1)], [(0, 2), (1, 1), (3, 1)]]

Derivation of Term Matrix T of Training Document Word Stems:  [[ 0.79411857  0.47930412  0.05482989  0.36964434]
 [-0.2236068  -0.2236068   0.67082039  0.67082039]
 [ 0.26339267 -0.68576057 -0.54497127  0.40418197]]

LSI Vectors of Training Document Word Stems:  [[(0, 1.2734226922603302), (1, -0.44721359549995804), (2, -0.4223679046003072)], [(0, 0.42447423075344

Cosine Similarities of LSI Vectors of Training Documents: [array([ 1.0000000e+00, -1.9446199e-08,  8.6602545e-01], dtype=float32), array([-1.9446199

LSI Vector Test Document: [(0, 0.534134012773409), (1, 0.4472135954999582), (2, -1.2307318377285026)]

Cosine Similarities of Test Document LSI Vectors to Training Documents LSI Vectors: [0.49999997 0.5        0.28867513]

Most similar Training Document to Test Document: She runs to the shop.
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: FutureWarning: Conversion of the second argument of issubdtype from `int` to `np.sign
  if np.issubdtype(vec.dtype, np.int):
```