Intelligent Data Mining - Exercise 3

Michael Debono Mrđen

8 November 2017

1 Assignment 1: Jaccard Similarity

a. Compute the Jaccard similarities of each pair of the following three sets.

- $\text{SIM}(\{1,2,3,4\},\{2,3,5,7\})$

  $= |\{1,2,3,4\} \cap \{2,3,5,7\}| / |\{1,2,3,4\} \cup \{2,3,5,7\}| = 2/6 = 1/3$

- $\text{SIM}(\{2,3,5,7\},\{2,4,6\})$

  $= |\{2,3,5,7\} \cap \{2,4,6\}| / |\{2,3,5,7\} \cup \{2,4,6\}| = 1/6$

- $\text{SIM}(\{1,2,3,4\},\{2,4,6\})$

  $= |\{1,2,3,4\} \cap \{2,4,6\}| / |\{1,2,3,4\} \cup \{2,4,6\}| = 2/5$

b. Compute the Jaccard bag similarity of each pair of the following three bags.

- $\text{SIM}_{\text{bag}}(\{1,1,1,2\},\{1,1,2,2,3\})$

  $= |\{1,1,1,2\} \cap \{1,1,2,2,3\}| / |\{1,1,1,2\} \cup \{1,1,2,2,3\}| = 3/9 = 1/3$

- $\text{SIM}_{\text{bag}}(\{1,1,2,2,3\},\{1,2,3,4\})$

  $= |\{1,1,2,2,3\} \cap \{1,2,3,4\}| / |\{1,1,2,2,3\} \cup \{1,2,3,4\}| = 3/9 = 1/3$

- $\text{SIM}_{\text{bag}}(\{1,1,1,2\},\{1,2,3,4\})$

  $= |\{1,1,1,2\} \cap \{1,2,3,4\}| / |\{1,1,1,2\} \cup \{1,2,3,4\}| = 2/8 = 1/4$

2 Assignment 2: Shingling

a. What are the first ten 3-shingles in the first sentence of Section 3.2?

"`The most effective way to represent documents as sets, for the purpose of identifying lexically similar documents is to construct from the document the set of short strings that appear within it.`"

1. “The”
2. “he ”
3. “e m”
4. “ mo”
5. “mos”
6. “ost”
7. “st ”
8. “t e”
9. “ ef”
10. “eff”

b. If we use the stop-word-based shingles of Section 3.2.4, and we take the stop words to be all the words of three or fewer letters, then what are the shingles in the first sentence of Section 3.2?

“*The* most effective *way to* represent documents *as* sets, *for the* purpose *of* identifying lexically similar documents *is to* construct from *the* document *the set of* short strings that appear within *it*.”

1. “The most effective”
2. “way to represent”
3. “to represent documents”
4. “as sets, for”
5. “for the purpose”
6. “of identifying lexically”
7. “is to construct”
8. “to construct from”
9. “the document the”
10. “the set of”
11. “set of short”
12. “of short strings”

3 Assignment 3: Jaccard similarity with Mahout

This assignment is done separately.