# Intelligent Data Management - Exercise 3

Name: Triet Ho Anh Doan
Date: November 8, 2017

## Assignment 1: Jaccard similarity

a. Compute the Jaccard similarities of each pair of the following three sets: $A = \{1, 2, 3, 4\}, B = \{2, 3, 5, 7\}, C = \{2, 4, 6\}$

- $SIM(A, B) = |A \cap B| / |A \cup B| = 2/6 = 1/3$
- $SIM(A, C) = |A \cap C| / |A \cup C| = 2/5$
- $SIM(B, C) = |B \cap C| / |B \cup C| = 1/6$

b. Compute the Jaccard bag similarity of each pair of the following three bags: $A = \{1, 1, 1, 2\}, B = \{1, 1, 2, 2, 3\}, C = \{1, 2, 3, 4\}$

- $SIM(A, B) = 3/9 = 1/3$
- $SIM(A, C) = 2/8 = 1/4$
- $SIM(B, C) = 3/9 = 1/3$

## Assignment 2: Shingling

The first sentence of section 3.2 is:

> 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.

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

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

1. The most effective
2. way to represent
3. to represent documents
4. as sets, for
5. for the purpose

6. the purpose of
7. of identifying lexically
8. is to construct
9. to construct from
10. the document the
11. the set of
12. set of short
13. of short strings