

## Assignment 9

### Task 1 - Document Similarity

- (a) Sentence 1: {impeachment: 1, -: 1, pelosi: 1, prepares: 1, to: 2, send: 1, articles: 1, senate: 1}  
 $(v(1) = (1 \ 1 \ 1 \ 1 \ 2 \ 1 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0))$   
Sentence 2: {head: 1, of: 1, hr: 1, watch: 1, denied: 1, entry: 1, to: 1, hong: 1, kong: 1}  
 $(v(2) = (0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0))$   
Sentence 3: {senate: 1, awaits: 1, articles: 1, of: 1, impeachment: 1}  
 $(v(3) = (1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1))$

(b)  $s(v(1), v(2)) = \frac{2 \cdot 1}{\sqrt{11} \cdot \sqrt{9}} = \frac{2}{9.95} = 1.809$

$$s(v(1), v(3)) = \frac{1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1}{\sqrt{11} \cdot \sqrt{5}} = 4.045$$

$$s(v(2), v(3)) = \frac{1 \cdot 1}{\sqrt{9} \cdot \sqrt{5}} = 0.745$$

The third sentence has a higher similarity to the first one than the second one.

- (c) One shortcoming is that this would result in sparse vectors which still need to be stored. The second disadvantage is, that this does not take the importance of words into account. It weights all words the same, but stop-words or less important ones could be neglected.

### Task 2 - Keyword Extraction