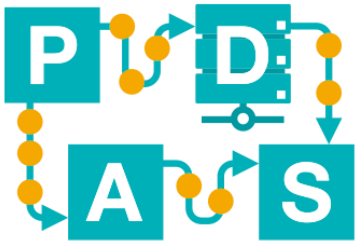
Text mining – Instruction

Miriam Wagner

IDS-I11



Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Querying systems

Exercise 1

D1: 'Cats are the only pet of the felines family, while dogs are canids.'

D2: 'Cats are the third-most popular pet in the US.'

D3: 'Dogs have been selected for centuries as pet animals.'

D4: 'Usually dogs are not aggressive to other dogs outside their territory.'

Give for the four documents (D1 – D4) and three queries the tf-idf scores.

Q1: 'dog'

Q2: 'dog pet'

Q3: 'dog cat'

Thereby, assume that plural normalizes on singular (pets = pet). No further preprocessing necessary.

Skip-grams (k-skip n-grams)

A skip-gram is an n-gram where it is allowed to "skip" some words.

Skip-grams where constructed for a certain skip distance k allow a total k or less skips for n-gram.

A 3-skip-gram includes: 3 skips, 2 skips, 2 skip and 0 skip

N-gram

Exercise 1

For the following sentence: “Insurgents killed in ongoing fighting”, create:

1. **2-grams:**
2. **2-skip-2-grams:**
3. **3-grams:**
4. **2-skip-3-grams:**