Assignment 3 – Information Retrieval

**Task 1:**

1. Tokenizing describes the process of the recognition of words in sequence of characters. Different problems occur in small words/abbreviations, hyphenated forms, numbers or periods. The preprocessing has to be the same for the docs and the query.

Examples: “E-bay is a good internet platform to sell t-shirts of EA.”

“The I.B.M. top 10 courses can be accessed with a nokia 3250 and is especially important for german-speaking students of CAU.”

“C++ can’t be as easy as JavaScript.”

1. . Pros: index gets smaller, stop words occur in high frequency in english language -> improve response time and effectiveness

Contra: some queries might be only identified with highest relevance if stop words aren’t filtered: (Example: The Doors), removal of stop words can put words together that aren’t relevant for the query (Example: Aftab likes the wine of the Zugspitze.” -> query: wine Zugspitze matches.

1. Yes -> The, And, In, For, Is

**Task 2:**

1. Stemming rule 1: Suffix ‘s -> works for “dogs->dog / horses->horse”, doesn’t work for “ups->up”/ ”windows-> window”

Stemming rule 2: Suffix -ing -> works for “shooting->shoot / climbing->climb”, doesn’t work for “heading->head”

1. Pros: Index gets smaller, improve effectiveness

Contra: False positives, false negatives, meaning of original words gets changed in some cases.

1. Stemming can cause same stem but different meaning. Therefore it can improve recall (lower FN) but lower precision (higher FP).

**Task 3:**

1. The two main options for computing n-grams are the bi- (2-word sequence) and trigrams (3-word sequence).
2. Pros: better query matching, meaningful phrases, improving recall

Contra: Index gets very big. Document containing 1,000 words would contain 3,990 instances of word n-grams of length 2 ≤ n ≤ 5 (lecture).

**Task 4:**

1. Named Entity Types (NER) can improve query matching in information retrieval (especially in question answering or search). Imagine the query: “Who is the president of the USA?” The word “who” implicates that the query is looking for a person for which the NER can use its classification to better match the query.
2. **KEINE AHNUNG, Lasst euch was einfallen**