# Homework Unit 2 - IR: Inverted Index, Positional Index, Basic NLP / NLTK

**To be presented:**

**Basics of NLTK:**

1. Download the text you selected (in **plain text** format) from https://www.gutenberg.org/
    a. **Update: please send the id of the book you selected from gutenberg to me, and I will insert it into the "Students_list" spreadsheet or make a comment**
    b. **Email of Gerhard Wohlgenannt: gwohlg@corp.ifmo.ru**
2. Apply word and sentence **tokenization**
3. Convert to a nltk Text (text = nltk.Text(tokens))
4. Use NLTK FreqDist to print and plot the most common words in your book
5. Compare the frequency to "Moby Dick" (text1) book in NLTK,
   `from nltk.book import text1`
   What are the differences in the 50 most frequent words?
6. Repeat step 5, but first remove all stopwords, and apply lemmatization to the list of tokens

**IR simple search:**

7. Split your assigned gutenberg book into paragraphs -- we will treat these paragraphs as single documents in the remainder of the task
    a. You can just say every block of 20 (or 50..) sentences from sentence tokenization is one paragraph (==document)

8. Now create an **positional index**, with the paragraphs being the documents (implement the positional index yourself!).

9. Implement simple search for 2 word phrase queries ( eg: "Arnold Schwarzenegger")

10. Do some example searches to show that the positional index works

11. Bonus (optional for extra points): Apply POS (part-of-speech) tagging to the text, and search only for nouns and adjectives in the text.
    https://www.nltk.org/book/ch05.html

12. ~~Add the phrase query feature to your IR system. The system should be able to process a query like~~
    ~~`man AND "strong will today", man OR ""strong will today""`~~