RQ 1.1: Why do we remove common stop words and lemmatise the text? Give an example of a scenario where, in addition to common stopwords, there are also application-specific stop words.

```
if we dont remove stop words. they might always be the most frequent words over
trully disiring informations. And without lemmatization, the inflections of one
same word will portion the frequency and make one same lemma less frequent than it
should be.
example of scenario: In the slides of our lectures, words like "thanks for today"
should be removed because it do not provide actual information about the slide
contents.
```

RQ 1.2: In Problem 2, what do the dimensions of the matrix X correspond to? What information from the original data is preserved, what information is lost in this matrix?

```
the rows are the numbers of input texts, the columns are the dimention of tfidf
vector, which is the number of all terms in the document.
the matrix preserved every text's terms(if we have the column names) and their
tfidf. But the order of terms, stopwords, different infelctions of lemma will be
lost, we cant retrive the origin text from the matrix.
```

RQ 1.3: What does it mean that a term has a high/low idf value? Based on this, how can we use idf to automatically identify stop words? Why do you think is idf not used as a term weighting on its own, but always in connection with term frequency (tf–idf)?

```
lower/higher idf means the term appears in more/less documents among all.
stop words will have the lowest idf among all terms, we can identify them by that.
words with low idf might not provide desiring informations, for exemple stop
words. so it's a good idea to use tf-idf to identidy words which are frequeny and
only frequent in corresponding documents. its high tf implys its important to that
certain document and its high idf means is not a word of generic, not-interested
meaning which is commonly used in all documents.
```