

# Project in IKT 407 – Webmining

## **Group 6**

Thomas Jakobsen and Thomas Skardal

# Our task

Create a plugin for NLTK which determines the readability of a text.

# What is readability?

- A numeric value that indicates how difficult it is to read and understand a text.
- There are several tests for calculating this.
  - Automated Readability Index (ARI)
  - Flesch
  - SMOG
  - ...
- These tests use statistics from the text such as number of words, sentences, syllables and complex words.

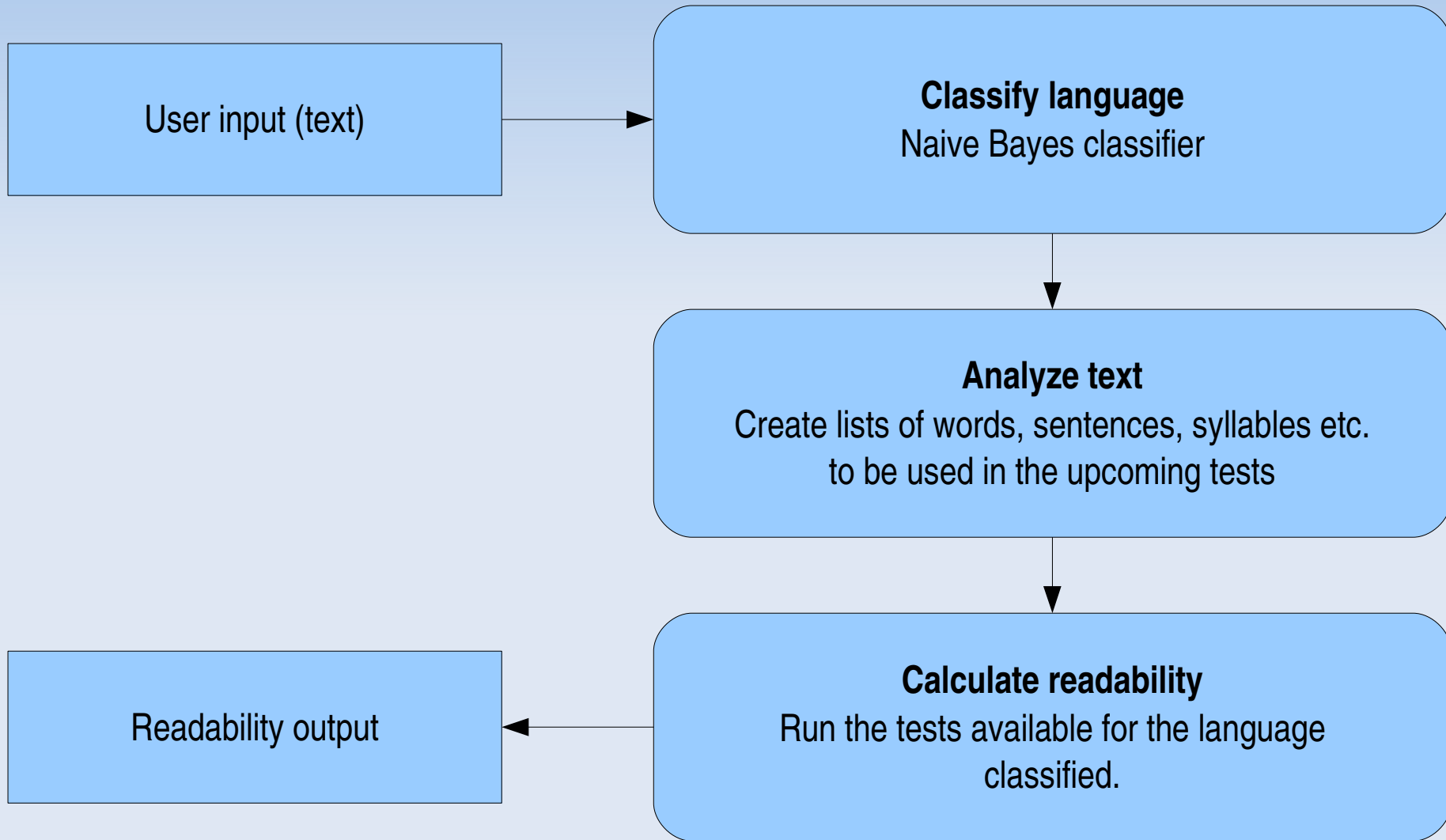
# Natural Language ToolKit

- A collection of open source Python modules, data and documentation.
- Used for development and research in Natural Language Processing.
- Already contains functionality that will be helpful for our task, such as different tokenizers.

# Language Classification

- Not all readability tests are language neutral.
- We will implement a classifier that can determine the language of a text.
- We will only train this classifier for Norwegian and English.
- This will be a naive bayes classifier.

# Flow



# The next steps

- We need to be able to count the words, sentences, syllables etc. of a text.
- Implement the different tests, with focus on English at first.
- Create and train the language classifier
- Determine which tests that works best with which languages.