Intelligent Systems

2nd Assignment 2018-2019

The second seminar focuses on natural language processing and text mining techniques. These techniques are used to extract many features from text: from simple, such as word frequency, to more sophisticated, such as sentence structure and writing style.

The assignments are to be completed **individually**. Presentations will take place during the lab practice sessions on the week of **January 14**.

Assignment

The assignment data are organized into three files, found on the course web page. The "essay.zip" file contains 723 essays on the topic of laughter. The "vocabulary.txt" file contains the essays' scores based on wording and figurative language use. The "structure.txt" file contains the essays' scores based on the rhythm and sentences structure. All scores are given on a scale from 1 to 6.

The tasks:

1. Use text mining techniques to extract some interesting features from the given set of essays (e.g. the average number of sentences per document, the average length of sentences, the most popular words, the association between words, terms used together, etc.). Show the extracted features visually.

2. Prepare a data set for training the scoring of essays based on vocabulary. A document-term matrix with tf-idf (term frequency - inverse document frequency) weights can be used as a starting point. Additionally, you can expand the data set by introducing new attributes like the number of different words in the essay, the number of rare words in the essay, etc. The scores in the "vocabulary.txt" file represent the target class values. Train a model for automated essay scoring and evaluate it (e.g., using cross-validation). Prepare a convincing presentation of the obtained results.

3. Prepare a data set for training the scoring of essays based on sentence structure. You can use POS (part-of-speech) tags to define new attributes that describe the average structure of sentences (e.g. the average number of nouns, verbs, adjectives, etc.). The scores in the "structure.txt" file represent the target class values. Train a model for automated essay scoring and evaluate it (e.g., using cross-validation). Prepare a convincing presentation of the obtained results.

4. Implement a function to replace the missing word in a given sentence with the most probable word. In order to determine highly probable replacement words you can use an

additional corpus of documents. A large collection of such documents is available at http://www.gutenberg.org/. The code from the lab exercises contains the most basic approach for achieving this. For your assignment, you should extend that approach (some possible improvements: looking at more than one previous word, looking at words after the missing word, properly preprocessing the text …).

Grading

The final score will be based on the predictive accuracy of trained models, your exposition and justification of the chosen approach, and your interpretation and presentation of the results. At least three of the four problems must be completed and convincingly presented for a positive grade.