Reconstructing Semantic Structures in Technical Documentation with Vector Space Classification

Jan Oevermann

Introduction

We introduce an approach that uses the classification knowledge present in available content components to reconstruct document structures in text extracted from legacy files, such as untagged PDF.

Motivation

Providing granular semantic access to the content of legacy PDF files, for example in information portals.

Method

1. Train vector space model with classified content components
2. Split text extracted from PDF into text chunks of the average component size (see Eq. 1)
3. Classify each text chunk and calculate the classification confidence (see Fig. 2)
4. Set boundaries for semantic document structure on confidence transitions (local minima)

Results

It was possible to reconstruct the basic semantic document structure for two example data sets and multiple untagged PDF files.

Outlook

Results could be further improved by taking neighbouring chunks into account and set minimum lengths for text spans of a given classification.

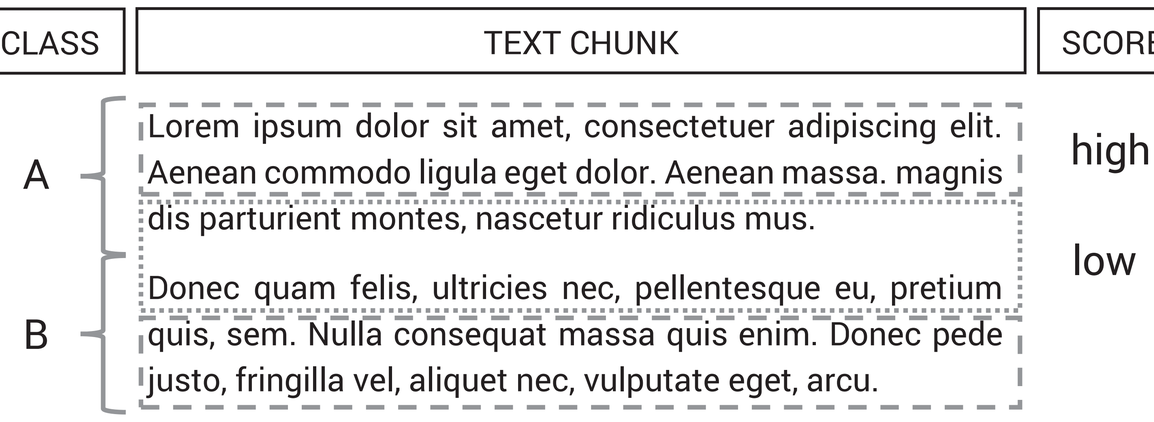

Fig. 1: Text chunk classification: Hypothesis

$$c_i = \{W_{(i-1)*r}, W_{(i-1)*r+1}, ..., W_{(i-1)*r+a}\}$$

Eq. 1: Definition of the i-th text chunk

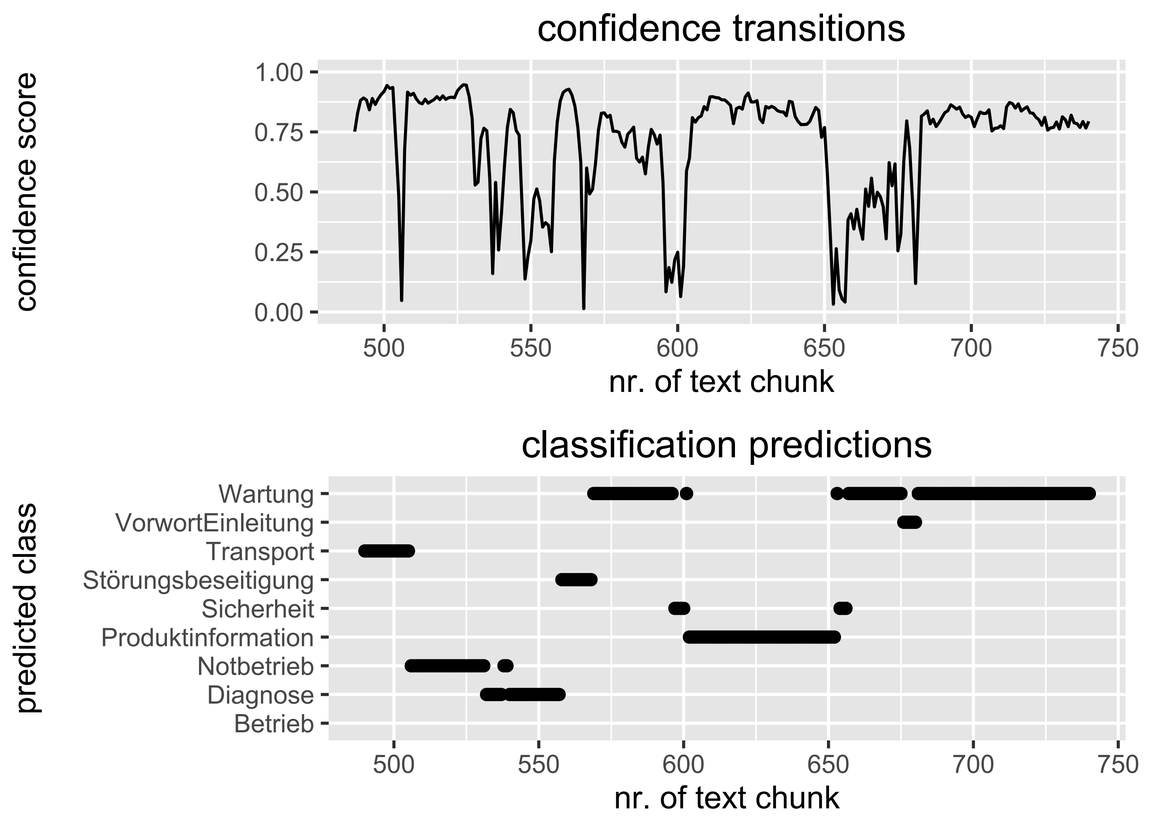

Fig. 2: Results of confidence transitions (Set A)

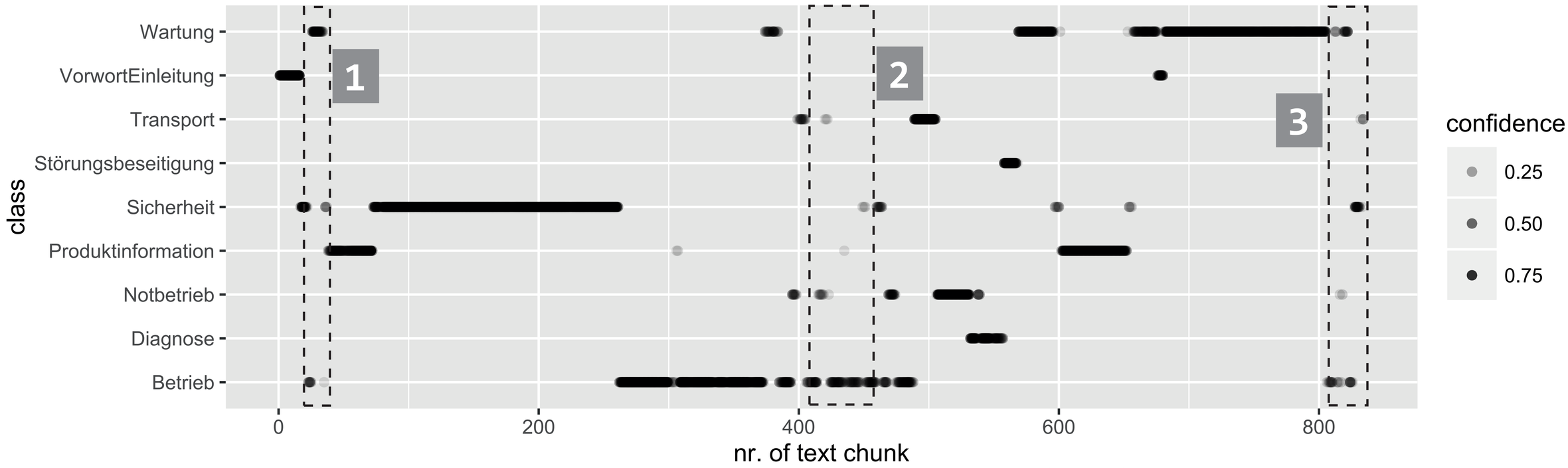

Fig. 3: Predicted classes with annotations (Set A)

Universität Bremen

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Jan Oevermann
jan.oevermann@hs-karlsruhe.de

www.home.hs-karlsruhe.de/~oeja0001