NLP Methods for Insurance Document Comparison

Auteur : Schoffeniels, Adrien
Promoteur(s) : Ittoo, Ashwin; 12800
Faculté : Faculté des Sciences appliquées
Diplôme : Master en ingénieur civil en informatique, à finalité spécialisée en "intelligent systems"
Année académique : 2020-2021
URI/URL : http://hdl.handle.net/2268.2/13271

ABSTRACT

University of Liège
School of Engineering and Computer Science

NLP Methods For Insurance Document Comparison

Adrien Schoffeniels

Supervised by Pr. Ashwin Ittoo

Academic year 2020-2021

This work aims to study the different steps of a process that would allow to compare 2 different versions of a document. This process is decomposed into 4 parts : text extraction, text segmentation, text matching and text comparison, which have been the subject of research and experiments. Especially, one show that comparing the sections of the documents rather than the complete documents improve the quality of the comparison.

The text matching task, which is the part studied in more depth, is a variant of the classification task, with the difference that there are no general categories from which we try to classify. Instead, each document has a unique set of classes, corresponding to each section, that can not be known in advance. This has many implications, mainly the fact that traditional classifiers cannot be used, as one cannot create training data for this task.

Different natural language processing (NLP) methods have been compared on the text matching task. For this purpose, a small dataset of pairs of documents with their matching has been built, and metrics inspired from the confusion matrix for the classification task has been designed, to be able to assess the performances of the different models. The models compared are term frequency (TF), TF-IDF, Word2vec combined with the Word Mover's distance, Doc2vec, BERT and RoBERTa. The different experiments show that more complex models are not suited for this matching task, and that it is preferable to use simple statistical models. Further work may investigate the performances of Latent Semantic Analysis (LSA) for this matching task.