Concept Study: using Large Language Models for automated management of Functional Safety standards document

Title: Toward the [Document AI](#) application on the [Functional Safety](#) Standards Documents

Administration

- Suitable for both graduate and undergraduate students
- Supervised by Dr. Anastasia Braginsky (RedHat Research) and Dr. Daniel Weingaertner (RedHat Germany)
- Please use this [form](#) for registration (to be later approved by Anastasia)
- Feel free refer to anastas@cs.technion.ac.il with any questions

Overview

Functional Safety standard's documents (ISO, CENELEC, DIN) are written in a particular technical writing style, aiming at being comprehensible and human readable, while avoiding imprecisions, double meanings or multiple interpretations. These standards are also written for specific industry fields (medical, automotive, railway) in order to cope with the particularities of that field.

These standards documents are often tailored for specific industries, although keeping a similar [document structure](#), as they share a common origin and subject.

A product (e.g. an operating system), developed to be used in a specific industry field, has to comply with the specific standard for that field. In order to be used in a different field, the product usually has to undergo a certification by another (hopefully related) standard. Therefore, it would be helpful to compare these standards documents, e.g. by comparing their requirements and objectives.

Project Description

The problem consists in identifying and extracting the structural elements of a functional safety standards document (i.e. objectives, requirements, glossary) in order to facilitate the (automatic) comparison of such documents across different fields.
The automatic management of the standard documents enables developers of a product, intended to be used in diverse fields, to verify the suitability of the product to the Functional Safety requirements on these fields. The idea is to use the result of this project as an input to tools such as the Standards Atlas, in order to link related or interdependent elements and to better navigate through the standards.

Given that Large Language Models (LLMs) are known to be good tools for textual information extraction, they should be investigated as means to automatically define and extract structural elements of the standard texts. There are a couple of pre-existing projects dealing with similar problems: Layout Parser, LayoutML, SecCerts, Google's Document AI.

Research questions

1. What are the important structural elements in functional safety standards documents?
    a. Can we automatically identify them, or do we need a predefined list?
2. How can we extract structural information from standards documents? Applying the similar to Layout Parser, LayoutML, SecCerts techniques.
3. How can we associate the content to the corresponding structural elements?

Possible outcomes or products
1. Analysis of the identified structural elements present in the functional safety standards documents
2. Development/adaptation of software to extract structural elements of the functional safety PDFs, classifying them and associating them to the content of that element
3. Integration of this software into the Standards Atlas project

What needs to be done?

- **Objectives:**
  - Identify existing software and/or techniques to automatically extract structure from documents
  - Apply structure extraction techniques to functional safety standards documents
- **Phases of Work:**
  - Collect diverse samples of standard documents to be used
  - Literature review of structure identification/extraction in documents
  - Use the pre-existing (with possible tuning) software to extract structure of a given standards PDF and provide it in the StandardsAtlas format (YAML)
  - Access the quality of the resulting structure
  - Associate document content relating to each identified structural element
- **Metrics for Evaluating and Monitoring:**
  - Quality of literature review regarding document structure extraction
  - Evaluation of the identified structural elements in the standard documents
  - Ability to use the generated structure in the Standards Atlasa

Learning goals

Students will learn about Functional Safety subject and usage of the LLMs for documents analyzing. LLM usage can be explicit (hands-on manipulations) or implicit via different tools already embedding LLMs.

Knowledge/skills preferable to the students:
- **Systems:** Doorstop, StandardsAtlas
- **Programming languages:** Python, YAML
- **Expertise:** ML, AI, previous work with LLMs - preferrable