Challenge: Patent Analysis

One of the text genres we frequently work with is *patents*. In this challenge, we are asking you to sketch an approach for extracting information from patents, and we are asking you to implement an initial prototype for the extraction.

Concept for Measurement Extraction

For patents related to BASF's business we are interested in extracting detailed technical information. One case of interest are *measurements and their values*. Consider, for instance, the patent US8022010B2 (https://patents.google.com/patent/US8022010B2/en). In Example 1 there is the sentence "The resulting BaCO3 had a crystallite size of between about 20 and 40 nm". We would like to extract and store the information that 'crystallite size' of 'BACO3' was measured in the unit 'nm', and the value was between 20 and 40.

For the first part of the challenge, please sketch how you would approach this measurement extraction task (including arriving at a task definition, annotation, model architecture and building, evaluation).

Prototype Implementation

Although a successful solution for the measurement extraction task requires careful modelling, annotation, model development and evaluation, we can gain valuable insights into the task by developing an initial prototype. Hence, for second part of the challenge, please implement a prototype approach that extracts measurement information from patents. Since large language models such as GPT-3.5 often yield decent performance with comparably low effort for such tasks, please use large language models as an underlying technology for the prototype.

The subtasks are as follows:

1. Download one ZIP archive containing granted patent full text data (without images) from https://bulkdata.uspto.gov/.
2. Read in the contained patents.
3. Devise and implement an approach based on a large language model (LLM) of your choice to extract measurements from the patents. The measurements

should be returned in a structured format (such as JSON). You do not need to process all patents but can restrict yourself to a set of a few tens or a few hundreds of patents.

4. Based on an inspection of the model's results on a held-out set, discuss the results/output of your models (e.g., what are typical errors, where do you see points for improvement).
5. Implement an improvement and compare and discuss the differences in the results.

Further information:

- We define an LLM as a large autoregressive model, such as GPT-3.5/4, the Anthropic models, LLaMa, MPT, …
- You are free to use a local installation of a LLM. If you want to use OpenAI's models, please get in touch with us to receive an API key.
- Patent XML format documentation (Word file)
- IPC Classification might be useful for filtering patents.

Solution Format

Please submit the following:

- concept: one-page sketch
- implementation: private GitHub project (implementation in Python)
- overall: short presentation (5 to 8 slides)

For questions regarding the challenge, please contact Sebastian Martschat (sebastian.martschat@basf.com). For obtaining an OpenAI API key, please contact Michael Schuhmacher (michael.schuhmacher@basf.com).