Dataset of Space Systems Corpora (Thesis Data)

Audrey Berquand (audrey.berquand@protonmail.com)

The dataset contains several corpora used to train the methods developed in the thesis "*Text Mining and Natural Language Processing for the Early Stages of Space Mission Design*" by Audrey Berquand, supervised by Dr. Annalisa Riccardi, University of Strathclyde.

The dataset is divided into three sub-folders:

- Processed Corpus: the main text sources (Publications, Books and Wikipedia pages) processed with a domain-specific Natural Language Processing pipeline.
- SpaceTransformers Corpus: Input for the SpaceTransformers family of models: parsed sentences from the publications' abstracts, the Wikipedia pages, and the books. The tokenization (Word Piece or BPE) is done within the further pre-training process.
- Additional Text Sources: Parsed ECSS and mission requirements.

Detailed information on each corpus and on the processing pipeline are found in: A. Berquand, "*Text Mining and Natural Language Processing for the Early Stages of Space Mission Design*", PhD Thesis, University of Strathclyde, Glasgow, UK, 2021.

1. Processed Corpus:

The main dataset includes **parsed and processed** data extracted from Wikipedia pages, books and publications. The publications corpus includes 4,991 articles published in the *Acta Astronautica*, the *Advance in Space Research* (ASR), and the *Aerospace Science and Technology* (AST) journals, accessed through the University of Strathclyde's subscription to Elsevier. The Wikipedia corpus includes 242 webpages. The books corpus contains 39 books related to space mission design, manually selected, and whose content is made available through the University of Strathclyde's subscription to Springer. The content of these articles, webpages and books were transformed and processed to generate new data for text mining purposes.

This collection of documents is **processed** with a domain-specific Natural Language Processing pipeline based on the Python NLTK library and including ECSS standards for acronyms expansion and multi-words recognition. This corpus is used for the generation of a domain-specific lexicon [1], and the training of a Topic Modelling spaceLDA approach [2] and a word embedding word2vec model [1].

2. SpaceTransformers Corpus:

This dataset includes 217,042 sentences extracted from 242 Wikipedia pages, 39 books, and 4,991 publications abstracts. These sentences are used as training and testing datasets (80/20) for the SpaceTransformers family of models published in [3].

3. Additional Text Sources

The collection of 27,016 ECSS requirements was extracted from 126 active standards. These standards were kindly provided by Mirtcheva, S. and Valera, S. (ESA). The ECSS requirements were used to train a doc2vec model in [4] and as case study corpus in [3].

The mission requirements are extracted from two ESA documents, publicly available, the SMOS mission System Requirement Document [5] and MarcoPolo-R's Mission Requirement Document [6]. They are used as case study corpus in [2].

References:

[1] A. Berquand, Y. Moshfeghi, and A. Riccardi, “Space mission design ontology: extraction of domain-specific entities and concepts similarity analysis," in AIAA Scitech 2020 Forum, Orlando, Florida, USA, 2020.

[2] A. Berquand, Y. Moshfeghi, and A. Riccardi, “SpaceLDA: Topic distributions aggregation from a heterogeneous corpus for space systems," Engineering Applications of Artificial Intelligence, vol. 102, no. April 2021, p. 104273, 2021. [Online]. Available: doi.org/10.1016/j.engappai.2021.104273

[3] A. Berquand, P. Darm and A. Riccardi, "SpaceTransformers: Language Modeling for Space Systems," in *IEEE Access*, vol.9, pp. 133111-133122, 2021, doi.org/10.1109/ACCESS.2021.3115659

[4] A. Berquand and A. Riccardi, “From Engineering Models to Knowledge Graph: Delivering New Insights Into Model," in Proceedings of the 9th International Conference on Systems & Concurrent Engineering for Space Applications (SECESA), 2020.

[5] ESA, “SMOS Systems Requirements Document," Tech. Rep., 2005.

[6] ESA, “MarcoPolo-R Mission Requirements Document," Tech. Rep., 2012.