Readme Data Endopaths

Author: Nicolai Wolpert (nicolai.wolpert@capgemini.com)

Date: October 2024

Link to git: https://gitlab.engine.capgemini.com/ti/france/data-and-ai/dimedia/endopath_new

This is the list of all the data that are important or might be useful, or that are used in the pipeline. There are numerous other data versions whose purpose is not clear and that can be ignored.

Name	Explanation	Path
Raw		
dossier-gyneco-23-03-2022.csv	Raw, unprocessed data with medical texts. Opening this file requires a password which can be found in ‘Raw’ > ‘Nouveau dossier’ > ‘PASSWORD_1.txt’	Raw
Recueil (1).csv	Data with demographical informations and symptoms, '1' specifying presence, '0' absence of symptom. Used in the Machine Learning scripts to predict endometriosis.	Raw
Recueil_final.csv	Like 'Receuil (1), but with tab that explains the meaning of the different columns. Not used in any script or notebook.	Raw
Processed		
all_words.txt	All words that appear in the text. Created in ‘create_correct_dict_nlp.ipynb’	DATA_PROCESSED /Correction_mots
comparaison_gynéco_receuil.csv	This file was created manually. For each feature, the present/absent values are noted for the original recueil data on the one hand and the infos extracted manually from the gyneco files on the other. This is to compare the two and resolve conflicting informations. Lines that are marked in red present cases where the two files seem to give conflicting information (to be reviewed by a medical expert).	DATA_PROCESSED
data_gynéco_manual_extraction_raw.csv	This was generated by going through the gynéco files and	DATA_PROCESSED

	noting all the informations manually. It is not the final file version used in the code because it contains some comments etc.	
data_gynéco_manual_extraction.csv	Notes all the informations (present/absent) for the features as they have been found in the gynéco files by manual inspection. In difference to 'data_gynéco_manual_extraction_raw', the column names have the same format as in the recueil data here, and everything is cleanly formatted. Not finalized yet as of October '24. Copy missing columns in question from 'comparison_gynéco_recueil.csv'.	DATA_PROCESSED
data_synth_priorité_gynéco.csv	Data synthesized between the gynéco and recueil data. In cases of conflicting information between the two (the cases marked in red in 'comparison_gynéco_recueil.csv'), prioritizes the gynéco data.	DATA_PROCESSED
data_synth_priorité_receuil.csv	Data synthesized between the gynéco and recueil data. In cases of conflicting information between the two (the cases marked in red in 'comparison_gynéco_recueil.csv'), prioritizes the recueil data.	DATA_PROCESSED
dictionnaire_correction.json	Dictionary of misspelled words and their corrections for later use. Created in 'create_correct_dict_nlp.ipynb'	DATA_PROCESSED /Correction_mots
donnees_entree_nlp_sans_endo.csv	Data with medical texts, preprocessed and corrected, without the target columns referring to endometriosis. Used for NLP scripts.	DATA_PROCESSED

The NLP/ML scripts in many cases have variables in the beginning of the script by which you can choose which type of data you want to enter into the model (e.g. either the original recueil data or the informations extracted manually from gynéco files). The documentation on Endopaths normally mentions in the top right of the slide which data have been used.