



VTT

VesselAI, 2022: Information Extraction from PDFs, User Guide

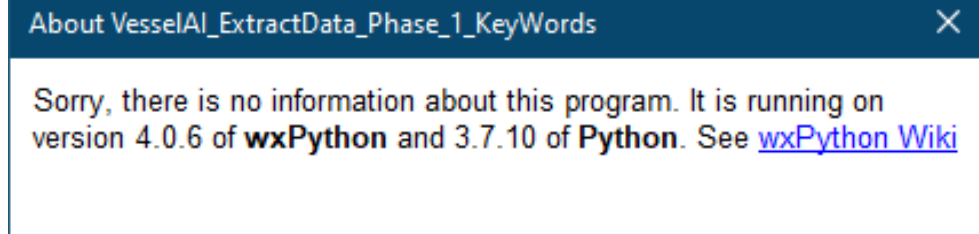
16/12/2022 VTT – beyond the obvious



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957237.

Trying to find marine engine specs & manuals from Internet sources (PDF format)

- We want to test how automatic extraction of information from PDF files would work
- The PDF files describe the performance of ship engines
- This presentation describes 4 Python programs, 3 first of which together implement a workflow for such information extraction from PDF files; NN version is separate
 - VesselAI_ExtractData_1.py
 - VesselAI_ExtractData_2.py
 - VesselAI_ExtractData_3.py
 - VesselAI_ExtractData_NN.py
- The programs were written using Anaconda Python, and wxPython UI library
- Introduction / general information in slides 2-12, the workflow & programs are described in slides 13-19, and installation instructions are in slides 20-22.



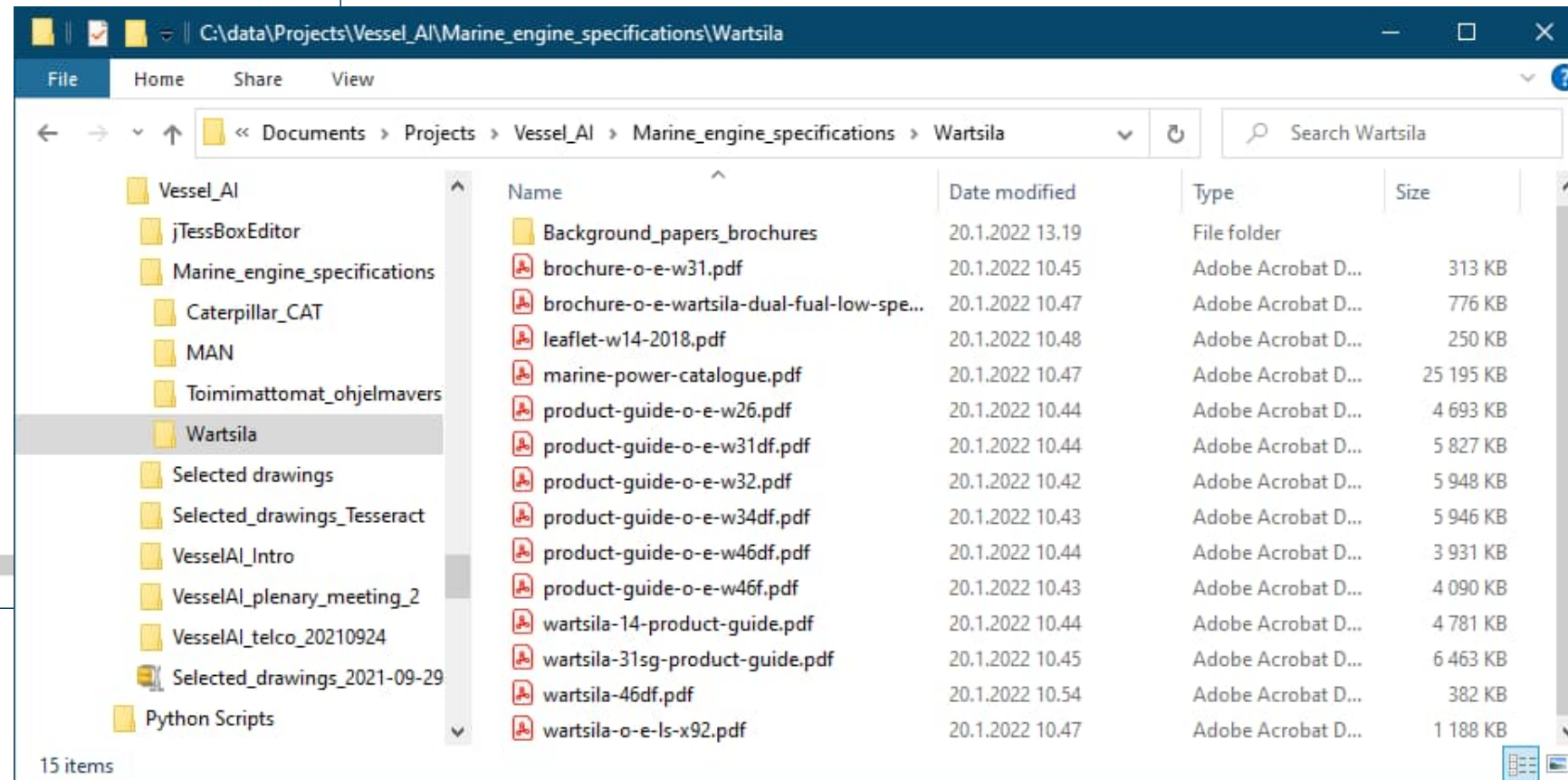
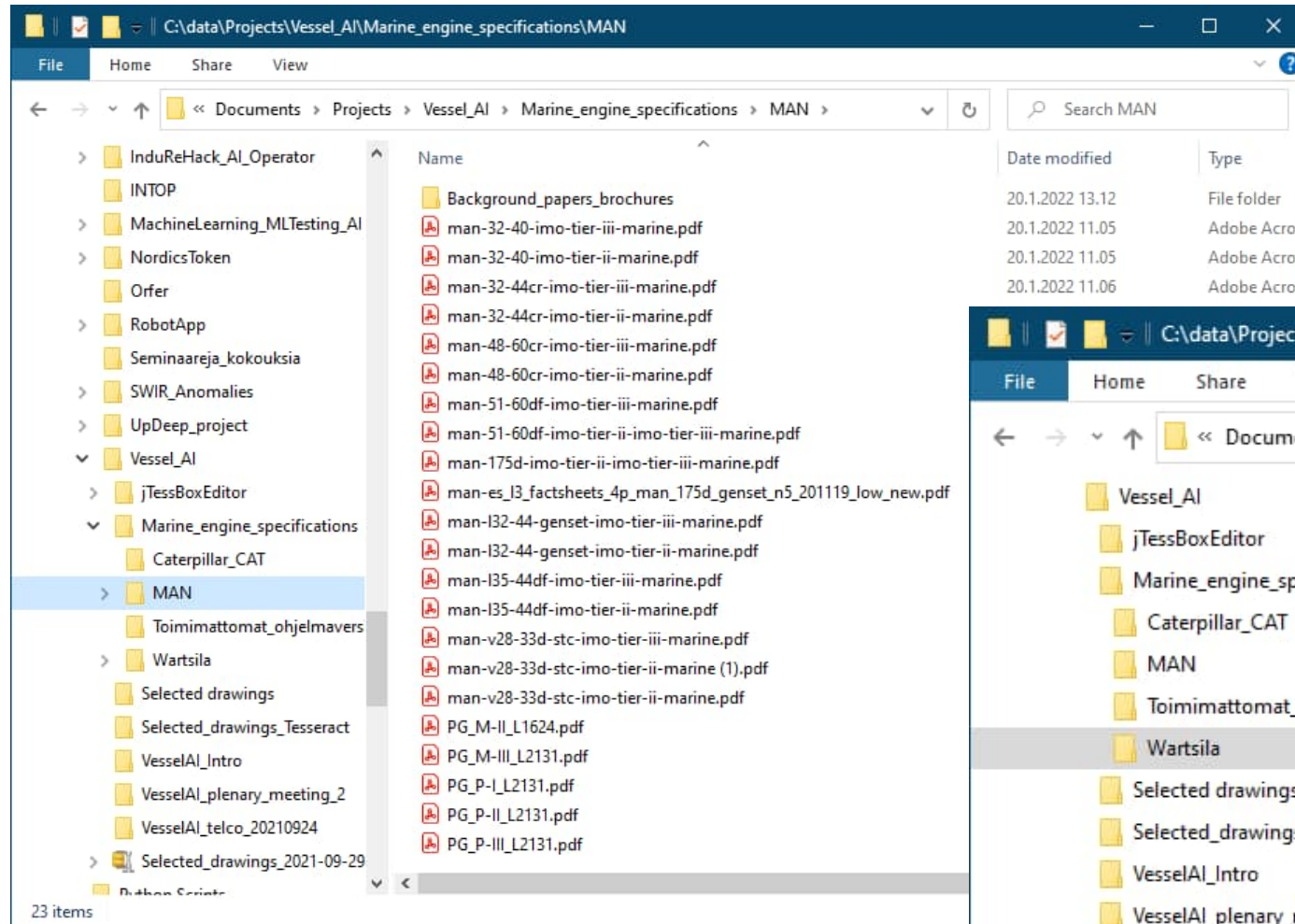
PDFs from largest ship engine manufacturers

- It was decided that we start with the 3 largest marine engine companies: Wärtsilä, MAN, and Caterpillar
- We downloaded all “technical looking” PDFs that consider marine engines
 - From Wärtsilä we downloaded 14 PDFs
 - From MAN we downloaded 22 PDFs
 - Alas, for Caterpillar / CAT there were no PDF specifications that could be freely downloaded...
- To make sure that our code is versatile enough, we later searched for PDF files in WinGD and HHI (Hyundai Heavy Industries) web pages also
 - Google search: engine specification “pdf” site:www.wingd.com
 - Google search: engine specification “pdf” site:english.hhi.co.kr
 - We downloaded 27 files from WinGD, and 11 files from HHI
- This was our test material!

PDF specifications from Wärtsilä and MAN

- These are the Wärtsilä and MAN PDF files that we considered:

- 22 PDFs from MAN, 14 from Wärtsilä...
- 27 files from WinGD, 11 files from HHI



PDF files NOT included!

- There are 73 PDF files in total, containing 18 019 pages
- These files are (or at least were in spring 2022) publicly available
- However, due to copyright issues we cannot distribute them
- Lists of these files are presented in the next slides
- You must try to find them from the vendors' web sites and download them yourself
- Google searches for MAN, Wärtsilä, HHI, WinGD PDF files (**or**: “engine manual”...):
 - engine specification “pdf” site: www.man-es.com/marine/
 - engine specification “pdf” site: www.wartsila.com
 - engine specification “pdf” site: english.hhi.co.kr
 - engine specification “pdf” site: www.wingd.com
- Save the found files into subfolders ‘MAN’, ‘Wartsila’, ‘Hyundai_HHI’, and ‘WinGD’
- However, the Python programs contain hard-wired lists of the files to be processed. If you cannot find all these listed files, you must modify these lists
- But the NN (neural network) program version also contains the indices of documents and pages that the human expert considered relevant (“Ground Truth”). These index lists must be modified also, which is tedious...

PDF files NOT included! MAN files to find

- MAN/man-175d-imo-tier-ii-imo-tier-iii-marine.pdf
- MAN/man-32-40-imo-tier-ii-marine.pdf
- MAN/man-32-40-imo-tier-iii-marine.pdf
- MAN/man-32-44cr-imo-tier-ii-marine.pdf
- MAN/man-32-44cr-imo-tier-iii-marine.pdf
- MAN/man-48-60cr-imo-tier-ii-marine.pdf
- MAN/man-48-60cr-imo-tier-iii-marine.pdf
- MAN/man-51-60df-imo-tier-ii-imo-tier-iii-marine.pdf
- MAN/man-51-60df-imo-tier-iii-marine.pdf
- MAN/man-es_l3_factsheets_4p_man_175d_genset_n5_201119_low_new.pdf
- MAN/man-l32-44-genset-imo-tier-ii-marine.pdf
- MAN/man-l32-44-genset-imo-tier-iii-marine.pdf
- MAN/man-l35-44df-imo-tier-ii-marine.pdf
- MAN/man-l35-44df-imo-tier-iii-marine.pdf
- MAN/man-v28-33d-stc-imo-tier-ii-marine.pdf
- MAN/man-v28-33d-stc-imo-tier-iii-marine.pdf
- MAN/PG_M-III_L2131.pdf
- MAN/PG_M-II_L1624.pdf
- MAN/PG_P-III_L2131.pdf
- MAN/PG_P-II_L2131.pdf
- MAN/PG_P-I_L2131.pdf

PDF files NOT included! Wärtsilä files to find

- Wartsila/brochure-o-e-w31.pdf
- Wartsila/brochure-o-e-wartsila-dual-fual-low-speed.pdf
- Wartsila/leaflet-w14-2018.pdf
- Wartsila/marine-power-catalogue.pdf
- Wartsila/product-guide-o-e-w26.pdf
- Wartsila/product-guide-o-e-w31df.pdf
- Wartsila/product-guide-o-e-w32.pdf
- Wartsila/product-guide-o-e-w34df.pdf
- Wartsila/product-guide-o-e-w46df.pdf
- Wartsila/product-guide-o-e-w46f.pdf
- Wartsila/wartsila-14-product-guide.pdf
- Wartsila/wartsila-31sg-product-guide.pdf
- Wartsila/wartsila-46df.pdf
- Wartsila/wartsila-o-e-ls-x92.pdf

PDF files NOT included! HHI files to find

- Hyundai_HHI/2016_HHI_en.pdf
- Hyundai_HHI/2017_HHI_en.pdf
- Hyundai_HHI/2018_HHI_en.pdf
- Hyundai_HHI/2019_HHI_en.pdf
- Hyundai_HHI/2020_KSOE_IR_EN.pdf
- Hyundai_HHI/2021_KSOE_IR_EN.pdf
- Hyundai_HHI/HHI_EMD_brochure2017_1.pdf
- Hyundai_HHI/HHI_EMD_brochure2017_2.pdf
- Hyundai_HHI/HHI_EMD_brochure2017_3.pdf
- Hyundai_HHI/HHI_EMD_brochure2019_1.pdf
- Hyundai_HHI/special_NSD2020.pdf

PDF files NOT included! WinGD files to find

- WinGD/16-06-pres-kit_x92.pdf
- WinGD/cimac2016_120_kyrtatos_paper_thedevelopmentofthefirstmodern2strokeengine.pdf
- WinGD/cimac2016_173_brueckl_paper_virtualdesign-and-simulation-in2strokeengine.pdf
- WinGD/cimac2016_233_ott_paper_x-df.pdf
- WinGD/dominikschneiter_tieriii-programme.pdf
- WinGD/Fuel-Flexible-Injection-System-How-to-Handle-a-Fuel-Spectrum-from-Diesel_CIMAC2019_paper_404_Andreas-Schmid.pdf
- WinGD/marcspahni_generation-x-engines.pdf
- WinGD/MIM_WinGD-RT-flex48T-D.pdf
- WinGD/MIM_WinGD-X72.pdf
- WinGD/MM_WinGD-RT-flex58T-D.pdf
- WinGD/MM_WinGD-X35-B_2021-09.pdf
- WinGD/MM_WinGD-X52.pdf
- WinGD/MM_WinGD-X72.pdf
- WinGD/MM_WinGD-X72DF.pdf
- WinGD/MM_WinGD-X82-B.pdf
- WinGD/Motorship-May-2018-VP-R-D-Leader-Brief.pdf
- WinGD/OM_WinGD-X62_2021-09.pdf
- WinGD/OM_WinGD-X72DF_2021-09.pdf
- WinGD/OM_WinGD-X82-B.pdf
- WinGD/OM_WinGD_RT-flex50DF.pdf
- WinGD/WinGD-12X92DF-Development-of-the-Most-Powerful-Otto-Engine-Ever_CIMAC2019_paper_425_Dominik-Schneiter.pdf
- WinGD/wingd-paper_engine_selection_for_very_large_container_vessels_201609.pdf
- WinGD/WinGD-WiDE-Brochure.pdf
- WinGD/WinGD_Engine-Booklet_2021.pdf
- WinGD/wingd_moving-inlet-ports-concept-for-optimization-of-2-stroke-uni-flow-engines_patrick-rebecchi.pdf
- WinGD/X-DF-Engines-by-WinGD.pdf
- WinGD/X-DF-FAQ.pdf

Wärtsilä PDF files: Relevant tables

- Then a few of the Wärtsilä PDF files were inspected
- The first one that contains relevant numerical information was ‘product-guide-o-e-w26.pdf’
- It contains many tables of the same format
- The page that contains the first such table is shown here:
- The four last lines give **fuel consumption** at 100% / 85 % / 75 % / 50 % engine load [g/kWh]
- The four first lines in group “Exhaust gas system” give **exhaust gas flow** at 100% / 85 % / 75 % / 50 % engine load [kg/s]
- There was no **charge air flow**; however, the first line in group “Combustion air system” gives “Flow of air at 100 % load” [kg/s]; we will use these numbers instead
- Header data above is crucial; we want to know the engine model etc. to which the numbers apply!

3. Technical Data

Wärtsilä 26 Product Guide

3.2 IMO Tier 2

3.2.1 Wärtsilä 6L26

Table 3-3

Wärtsilä 6L26		AE/DE IMO Tier 2	AE/DE IMO Tier 2	ME IMO Tier 2	ME IMO Tier 2
Cylinder output	kW/cyl	325	340	325	340
Engine speed	rpm	900	1000	900	1000
Engine output	kW	1950	2040	1950	2040
Mean effective pressure	MPa	2.55	2.4	2.55	2.4
Combustion air system (Note 1)					
Flow of air at 100% load	kg/s	3.7	4.1	3.9	4.1
Temperature at turbocharger intake, max.	°C	45	45	45	45
Air temperature after air cooler, nom. (TE601)	°C	55	55	55	55
Exhaust gas system (Note 2)					
Flow at 100% load	kg/s	3.8	4.2	4.0	4.2
Flow at 85% load	kg/s	3.3	3.7	3.4	3.5
Flow 75% load	kg/s	3.0	3.4	3.0	3.1
Flow 50% load	kg/s	2.1	2.3	1.9	2.2
Temp. after turbo, 100% load (TE517)	°C	343	324	318	324
Temp. after turbo, 85% load (TE517)	°C	343	319	327	328
Temp. after turbo, 75% load (TE517)	°C	349	321	339	341
Temp. after turbo, 50% load (TE517)	°C	367	344	402	388
Backpressure, max.	kPa	3.0	3.0	3.0	3.0
Exhaust gas pipe diameter, min	mm	500	500	500	500
Calculated exhaust diameter for 35 m/s	mm	492	507	493	507
Heat balance (Note 3)					
Jacket water, HT circuit	kW	348	372	336	372
Charge air, LT-circuit	kW	611	723	689	723
Lubricating oil, LT-circuit	kW	288	306	282	306
Radiation	kW	92	97	92	97
Fuel system (Note 4)					
Pressure before injection pumps (PT101)	kPa	700±50	700±50	700±50	700±50
Engine driven pump capacity at 12 cSt (MDF only)	m³/h	2.9	3.2	2.9	3.2
Fuel flow to engine (without engine driven pump), approx.	m³/h	1.6	1.8	1.7	1.8
HFO viscosity before engine	cSt	16...24	16...24	16...24	16...24
HFO temperature before engine, max. (TE 101)	°C	140	140	140	140
MDF viscosity, min	cSt	2.0	2.0	2.0	2.0
MDF temperature before engine, max. (TE 101)	°C	45	45	45	45
Fuel consumption at 100% load	g/kWh	190.6	194.4	191.5	194.4
Fuel consumption at 85% load	g/kWh	189.6	193.4	188.7	191.5
Fuel consumption at 75% load	g/kWh	192.0	195.3	190.6	193.4
Fuel consumption at 50% load	g/kWh	202.3	207.0	196.6	201.3

PDF files: Relevant tables

- After much testing we settled for the following rules:
- We first search for these keywords (can be inside () or []):
 - g/kWh
 - kJ/kWh
 - kg/s
 - kg/kWh
- But with two upper keywords the following other words must also be found in the same page:
 - fuel **AND** consumption
 - **OR** sfoc **OR** bsfc
- And with two lower keywords the following other words must also be found in the same page:
 - charge air **OR** combustion air **OR** exhaust gas
 - **AND** flow
- This produced good enough results!

3.2 IMO Tier 2

3.2.1 Wärtsilä 6L26

Table 3-3

Wärtsilä 6L26		AE/DE IMO Tier 2	AE/DE IMO Tier 2	ME IMO Tier 2	ME IMO Tier 2
Cylinder output	kW/cyl	325	340	325	340
Engine speed	rpm	900	1000	900	1000
Engine output	kW	1950	2040	1950	2040
Mean effective pressure	MPa	2.55	2.4	2.55	2.4
Combustion air system (Note 1)					
Flow of air at 100% load	kg/s	3.7	4.1	3.9	4.1
Temperature at turbocharger intake, max.	°C	45	45	45	45
Air temperature after air cooler, nom. (TE601)	°C	55	55	55	55
Exhaust gas system (Note 2)					
Flow at 100% load	kg/s	3.8	4.2	4.0	4.2
Flow at 85% load	kg/s	3.3	3.7	3.4	3.5
Flow 75% load	kg/s	3.0	3.4	3.0	3.1
Flow 50% load	kg/s	2.1	2.3	1.9	2.2
Temp. after turbo, 100% load (TE517)	°C	343	324	318	324
Temp. after turbo, 85% load (TE517)	°C	343	319	327	328
Temp. after turbo, 75% load (TE517)	°C	349	321	339	341
Temp. after turbo, 50% load (TE517)	°C	367	344	402	388
Backpressure, max.	kPa	3.0	3.0	3.0	3.0
Exhaust gas pipe diameter, min	mm	500	500	500	500
Calculated exhaust diameter for 35 m/s	mm	492	507	493	507
Heat balance (Note 3)					
Jacket water, HT circuit	kW	348	372	336	372
Charge air, LT-circuit	kW	611	723	689	723
Lubricating oil, LT-circuit	kW	288	306	282	306
Radiation	kW	92	97	92	97
Fuel system (Note 4)					
Pressure before injection pumps (PT101)	kPa	700±50	700±50	700±50	700±50
Engine driven pump capacity at 12 cSt (MDF only)	m³/h	2.9	3.2	2.9	3.2
Fuel flow to engine (without engine driven pump), approx.	m³/h	1.6	1.8	1.7	1.8
HFO viscosity before engine	cSt	16...24	16...24	16...24	16...24
HFO temperature before engine, max. (TE 101)	°C	140	140	140	140
MDF viscosity, min	cSt	2.0	2.0	2.0	2.0
MDF temperature before engine, max. (TE 101)	°C	45	45	45	45
Fuel consumption at 100% load	g/kWh	190.6	194.4	191.5	194.4
Fuel consumption at 85% load	g/kWh	189.6	193.4	188.7	191.5
Fuel consumption at 75% load	g/kWh	192.0	195.3	190.6	193.4
Fuel consumption at 50% load	g/kWh	202.3	207.0	196.6	201.3

PDF files: Learning Program

- Another version of the original Python program was written
- Instead of selecting keywords (e.g. “kJ/kWh”, “fuel”, “consumption”; see previous slide) we select a few of our data pages as “Model Pages”
- For example, some Model Pages could be related to fuel consumption, some to charge air / exhaust gas / heat balance
- Then the program computes the relative frequencies of all the words in each Model Page
- And when a candidate page is processed, a similar word frequency histogram is computed, and compared to the frequency histograms of all the Model Pages
- Words that can be converted to floating-point numbers are excluded from the diff. sums
- If the best difference is small enough (e.g. ≤ 0.5), i.e. that candidate page matches at least one of the Model Pages, then that candidate page is accepted

- **New versions of the Python programs discussed previously were written, so that together the new programs create a whole workflow for information extraction:**
 1. We extend the first version of our program (using keywords) so that the user enters some keywords (e.g. fuel consumption). Then the program finds those *.pdf file **pages** that match these keywords. Optionally the user can eliminate some (irrelevant) pages. The remaining pages go to phase 2.
 2. We cluster the found pages into groups (i.e. clusters); group center pages will be the proposed Model Pages. Optionally the user can eliminate some (irrelevant) Model Pages. The remaining pages go to phase 3.
 3. We extend the second version of our program (learning version) to use these Model Pages, and then to find more of similar pages within the PDF files. Besides, for each model page the user can define rules for lines where the relevant data values are, thus reducing the amount of irrelevant information even further.
- **And now we have a toolchain that can extract information from a large number of PDF files: Return desired parameter values for desired marine engine types.**

Implementation of the workflow (1/4)

VTT

- Still using our set of downloaded PDF files
- Can be later extended to web crawlers...
- Three separate Python programs to realize the workflow:
 - VesselAI_ExtractData_1.py
 - VesselAI_ExtractData_2.py
 - VesselAI_ExtractData_3.py
- In Phase 1 the user enters keywords and other words, and the program finds the matching files/pages; links to these files/pages are saved
- The Phase 2 program reads these files, and tries to group into clusters (4 algorithms); central page in each cluster is saved as a Model Page
- In Phase 3 the user can edit this set of Model Pages; and then the program tries to find more of similar pages, as well as extract relevant lines

The screenshot displays the VesselAI_ExtractData_Phase_1 application interface. It features a top navigation bar with 'File' and 'Help' menus. The main workspace is divided into several functional areas:

- Data Sets:** A section for managing data sets, including a list of fuel types (Fuel SFOC, Fuel BPC, Charge air, Combust air, Exhaust gas) and a 'Selected Data Set' dropdown.
- Keywords:** A section for defining keywords, with a text input field containing 'g/kWh kW/h' and a 'Keywords' dropdown.
- Name is an ID for the Data Set:** A section for defining the name of the data set, with a text input field.
- Data Sets List:** A table listing data sets with columns for 'Data Set', 'Name', 'Keywords', and 'Other words'.
- Search for text in all data:** A section for searching through the data, with a search bar and buttons for 'Search', 'Current Page', 'Open PDF', 'Copy Link', 'Open HTML', and 'Refresh Page'.
- Clusters:** A section for managing clusters, with a table listing clusters and buttons for 'Rename', 'Delete', and 'Execute'.

The interface is designed for data extraction and clustering from PDF files, providing a structured way to organize and analyze the data.

Implementation of the workflow (2/4): Phase 1

- The user enters a number of Data Sets, each contains Name (=ID for DS), Keywords, Other words; the list of defined Data Sets is shown
- The Data Set definitions are saved for next use. At first use: reasonable default values
- The algorithm of this program is executed with button 'Refresh', and the found files/pages are shown in a drop-down, and the selected page can be seen as an image and as text
- User can exclude some of the result pages from results
- Phase 2 program is started with button 'Next >>'

The screenshot displays the VesselAI ExtractData_Phase_1 application window. The interface is divided into several sections:

- Data Sets:** A list of predefined data sets including Fuel consump, Fuel SFOC, Fuel BSFC, Charge air, Combust air, and Exhaust gas. A dropdown menu is open for 'Fuel consump'.
- Selected Data Set:** 'Fuel consump' is selected.
- Name:** 'Fuel consump'.
- Keywords:** 'g/kWh kJ/kWh'.
- Other words:** 'fuel consumption'.
- Buttons:** 'Add/Mod', 'Delete', 'Refresh', 'Next >>', 'Exclude All', 'Include All', 'Search', 'Open PDF', 'Copy Link', 'Open Html', 'Refresh Png'.
- File Path:** 'file:///C:/data/Projects/Vessel_AI/Marine_engine_specifications/MAN/man-175d-imo-tier-ii-imo-tier-iii-marine.pdf#page=79'.
- Search Results:** A list of found files/pages, including '3 Technical data and engine performance', '3.1 Performance data - Mechanical propulsion applications, IMO Tier II', and '3.1.1 MAN 20V175D-ML, 220 kW/cyl., 2,000 rpm, IMO Tier II'.
- Technical data and engine performance:** A detailed table showing engine specifications for the MAN 20V175D-ML engine. The table includes columns for Units, ISO, and Limit conditions. The data is organized into sections for Engine output, Engine speed, Specific fuel oil consumption, Total fuel oil consumption, Lube oil consumption, and Reference conditions.

Units	ISO	Limit conditions ¹⁾
Air temperature	25	45
Seawater inlet temperature	18	32
Air pressure ²⁾	1,000	
Exhaust back pressure ³⁾	50	
Relative humidity	30	60

Units	ISO	Limit conditions ¹⁾
Air temperature	25	45
Seawater inlet temperature	18	32
Air pressure ²⁾	1,000	
Exhaust back pressure ³⁾	50	
Relative humidity	30	60

Implementation of the workflow (3/4): Phase 2

- The program tries to group Phase 1 result pages into clusters
- The central page of each cluster becomes a Model Page for Phase 3
- At first two algorithms for page clustering were available: **K-means** and **DBSCAN**; both from scikit learn (a.k.a. sklearn) program library:
 - https://scikit-learn.org/stable/auto_examples/text/plot_document_clustering.html#sphx-glr-auto-examples-text-plot-document-clustering-py
 - (“Clustering text documents using k-means”)
 - <https://programmer.group/61755e726f9b3.html>
 - (“Self defined distance measurement method of clustering algorithm in scikit learn Library”)
 - Later **AffinityPropagation** and **SpectralClustering** algorithms from scikit learn were also added!
- A lengthy Precomputation step is necessary before the 4 algorithms can be tried, but clustering itself is fast

The screenshot displays the 'VesselAI_ExtractData_Phase_2_Clusters' application. The interface includes a file selection dropdown, a clustering algorithm dropdown set to 'K-Means', and a 'Number of Clusters' input set to 6. Below this, a list of clusters is displayed, including 'Cluster 0 (28 pages)', 'Cluster 1 (168 pages)', 'Cluster 2 (55 pages)', 'Cluster 3 (20 pages)', 'Cluster 4 (85 pages)', and 'Cluster 5 (150 pages)'. A 'Selected Model Page of Cluster' dropdown is also present. The main area displays the content of the selected cluster, showing technical data and engine performance for a MAN 20V175D-M engine. The right sidebar shows a preview of the selected cluster's content, including a table of engine performance data.

Units	100 %	85 %	75 %	50 %	25 %	10 %	
Engine output	kW	4,400	3,740	3,300	2,200	1,100	440
Engine speed (RPM-curve)	rpm	2,000	1,885	1,817	1,587	1,260	928
Specific fuel of consumption ¹⁾	g/kWh	199.0	195.0	195.0	190.0	201.0	305.0
Total fuel of consumption ²⁾	l/h	1,047.0	876.0	799.0	500.0	265.0	161.0
Lube oil consumption ³⁾	g/kWh	0.11	-	-	-	-	-

Implementation of the workflow (4/4): Phase 3

- The user selects a number Phase 2 result pages as Model Pages, each contains Name (=ID for MP) and rules for lines of relevant data; the list of defined Model Pages is shown
- The contents of this list comes from Phase 2 (no data: reasonable default contents)
- For each Model Page, the user must define the lines where relevant data values are found
- The algorithm of this program is executed with button 'Refresh', and the found files/pages are shown in a drop-down, and the selected page can be seen as an image and as text. Data in relevant lines shown separately
- User can exclude some of the result pages from results
- Result data (=relevant lines) are shown with button 'Results'. New Notepad window is opened

The screenshot displays the VesselAI ExtractData Phase 2 software interface. The main window is titled 'VesselAI ExtractData Phase 2' and contains several sections for configuring data extraction. At the top, a file path is shown: 'file:///C:/data/Projects/Vessel_AI/Marine_engine_specifications/MAN/man-175d-imo-tier-ii-imo-tier-ii-marine.pdf#page=79'. Below this, a 'Model Pages' section lists several pages: 'Fuel oil 1', 'Fuel oil 2', 'Air flow 3', 'Fuel oil 4', 'Air flow 5', and 'SFOC 6'. Each page has a 'Modify' button. A 'Selected Model Page' section shows 'Fuel oil 1' and 'Doc 0, Pg 79'. A 'Lines' section has a text input for 'Specific fuel oil consumption' and a 'Line 1' dropdown. A 'Text in Line 1' section shows the selected page content. A 'Max Hist Diff' section has a slider and 'Refresh' and 'Results' buttons. A 'Search for text in all data' section has a search bar and buttons for 'Current Page', 'Open PDF', 'Copy Link', 'Open HTML', and 'Refresh Png'. The bottom section shows a list of result pages, including '3.1.1 MAN 20V175D-ML, 220 kW/cyl., 2,000 rpm, IMO Tier II'. The right side of the interface shows a preview of the selected page, which is a technical data table for the MAN 20V175D-ML engine.

Units	100 %	85 %	75 %	50 %	25 %	10 %	
Engine output	kW	4,400	3,740	3,300	2,200	1,100	440
Engine speed (FPP-curve)	rpm	2,000	1,895	1,817	1,567	1,260	928
Specific fuel oil consumption ¹⁾	g/kWh	199.0	196.0	195.0	190.0	201.0	305.0
Total fuel oil consumption ²⁾	l/h	1,047.0	876.0	769.0	500.0	265.0	161.0
Lube oil consumption ³⁾	g/kWh	0.11	-	-	-	-	-

¹⁾ Tolerance +5 %.
²⁾ Based on ISO reference conditions (according to ISO 15500:2002; ISO 3046:2002) and a lower calorific value of 42 700 kJ/kg and engine equipped with attached lube oil pump(s), fuel oil pump(s), HT- and LT cooling water pump(s).
³⁾ Total fuel oil consumption (l/h) calculated based on above stated specific fuel oil consumption (g/kWh) and a density of 837 kg/m³.
⁴⁾ See accordingly section 'Lube oil consumption, Page 220'.
Table 25: Marine mechanical propulsion light duty, 220 kW/cyl., 2,000 rpm, IMO Tier II

Reference conditions	Units	ISO	Limit conditions ¹⁾
Air temperature	°C	25	45
Seawater inlet temperature	°C	18	32
Air pressure ²⁾	mbar	-	1,000
Exhaust back pressure ³⁾	mbar	-	50
Relative humidity	%	80	90

¹⁾ Please contact MAN Energy Solutions if project specific limit conditions might be exceeded.
²⁾ Intake air depression up to 30 mbar allowed.
³⁾ Reference value for the difference pressure of exhaust gas line (plant) at MCR for IMO Tier II variant.
Table 26: Reference conditions – MAN 175D IMO Tier II

- Please note that when you open the window of the next phase (button 'Next >>'), a new window is opened, but the original window remains open behind this new window; however, because the new window uses the same Python/wxPython environment as the original one, the original window remains inactive until the new window is closed
- Lengthy operations (e.g. 'Precomp' in Phase 2, 'Refresh' in Phases 1 and 3) are executed in the main thread, so the UI is frozen until the operation is completed. Progress is shown in the bottom subwindow, but this window is not always updated properly; especially if you move the focus into another window, the UI becomes frozen until the computation ends
- The programs have been tested in Windows 11, Anaconda Python; in other environments (e.g. Linux) there may be problems...

Neural network program

- Ground truth of relevant pages by human expert: 690 pages were deemed relevant
- NN program: many settings were tried, results of best one
 - Precision = 0.998, Recall = 0.979, Accuracy = 0.999
- NN produced much better results than either Phase 1 or Phase 3
- However, for new data (e.g. new marine engine manufacturer) both NN and Phase 3 are likely to fail
- But Phase 1 could still work: SI units (our keywords) and technical terms (our other words) are the same

The screenshot displays the VesselAI/ExtractData_NN application interface. The main window is titled 'VesselAI/ExtractData_NN' and contains a 'File' menu and a 'Help' button. The interface is divided into several sections:

- Program mode:** A dropdown menu set to 'Train and Evaluate ALL samples'.
- How to process numbers?** A dropdown menu set to 'Remove numbers'.
- How to score words?** A dropdown menu set to 'binary'.
- Convert all words to lowercase?** A checkbox labeled 'Turn to Lowercase' is checked.
- Remove Punctuation from all words?** A checkbox labeled 'Remove Punctuations' is checked.
- Remove English stop words from Vocabulary?** A checkbox labeled 'Remove Stop Words' is checked.
- Eliminate short words from Vocabulary?** A checkbox labeled 'Minimum word length' is set to 2.
- Eliminate infrequent words from Vocabulary?** A checkbox labeled 'Minimum occurrence' is set to 2.
- Duplicate positive samples N times?** A checkbox labeled 'Duplicate x N' is set to 10.
- Neural Network structure?** A checkbox labeled 'Use 2 Layers' is checked. The 'Neurons in Layer 1' is set to 50 and 'Neurons in Layer 2' is set to 50.

The 'Train and evaluate Neural Network' section shows a file path: 'file:///C:/Users/terak/Docs/Projects/Vessel_AI/Marine_engine_specifications/MAN/man-175d-imo-tier-ii-imo-tier-iii-marine.pdf#page=79'. Below this, there is a search bar and buttons for 'Search', 'Current Page', 'Open PDF', 'Copy Link', 'Open HTML', and 'Refresh Png'.

A list of documents is displayed, including:

- WIND/MW_WinGD-X72DF.pdf
- WIND/MW_WinGD-X82-B.pdf
- WIND/Motorship-May-2018-VF-R-D-Leader-Brief.pdf
- WIND/OM_WinGD-X62_2021-09.pdf
- Doc:62, Page:509
- WIND/OM_WinGD-X72DF_2021-09.pdf
- Doc:63, Page:583
- WIND/OM_WinGD-X82-B.pdf
- Doc:64, Page:449
- WIND/OM_WinGD-RT-flex50DF.pdf
- Doc:65, Page:567
- WIND/WinGD-12X92DF-Development-of-the-Most-Powerful-Octo-Engine-Ever_CIMAC2019_paper_425_Dominik-Schneider.pdf
- WIND/WinGD-paper_engine_selection_for_very_large_container_vessels_201609.pdf
- Doc:67, Page:9
- WIND/WinGD-WiDE-Brochure.pdf
- WIND/WinGD_Engine-Booklet_2021.pdf
- Doc:68, Page:12
- Doc:69, Page:13
- Doc:69, Page:14
- Doc:69, Page:15
- Doc:69, Page:16
- Doc:69, Page:17
- Doc:69, Page:18
- Doc:69, Page:19
- Doc:69, Page:20
- Doc:69, Page:22
- Doc:69, Page:23
- Doc:69, Page:24
- Doc:69, Page:25
- Doc:69, Page:26
- Doc:69, Page:27
- WIND/WinGD_moving-inlet-ports-concept-for-optimization-of-2-stroke-uni-flow-engines_patrick-rebecchi.pdf
- WIND/X-DF-Engines-by-WinGD.pdf
- Doc:71, Page:4
- WIND/X-DF-FAQ.pdf

A 'Computation finished' dialog box is shown, displaying: 'nbr_true_pos: 690, nbr_fals_pos: 2, nbr_true_neg: 17327, nbr_false_neg: 0' and 'Precision: 0.997, Recall: 1.000'.

The right side of the screenshot shows a document titled 'MAN Energy Solutions' with the section '3 Technical data and engine performance'. The subsection '3.1 Performance data – Mechanical propulsion applications, IMO Tier II' is expanded, showing '3.1.1 MAN 20V175D-ML, 220 kW/cyl, 2,000 rpm, IMO Tier II'. The table below shows engine performance data:

Units	100 %	85 %	75 %	50 %	25 %	10 %	
Engine output	kW	4,400	3,740	3,300	2,200	1,100	440
Engine speed (FPP-curve)	rpm	2,000	1,885	1,817	1,587	1,280	928
Specific fuel oil consumption ¹⁾	g/kWh	199.0	196.0	196.0	190.0	201.0	300.0
Total fuel oil consumption ²⁾	l/h	1,047.0	876.0	789.0	500.0	285.0	161.0
Lube oil consumption ³⁾	g/kWh	0.11	-	-	-	-	-

Footnotes:

- ¹⁾ Tolerance +5 %.
- ²⁾ Based on ISO reference conditions (according to ISO 15550:2002; ISO 3046:2002) and a lower calorific value of 42,700 kJ/kg and engine equipped with attached lube oil (pumps), fuel oil (pumps), HT- and LT cooling water pumps).
- ³⁾ Relevant for engine's certification for compliance with the NO_x limits according to E3 Test cycle.
- ⁴⁾ Total fuel oil consumption [l/h] calculated based on above stated specific fuel oil consumption [g/kWh] and a density of 837 kg/m³.
- ⁵⁾ See accordingly section 'Lube oil consumption, Page 220'.

Table 25: Marine mechanical propulsion light duty, 220 kW/cyl, 2,000 rpm, IMO Tier II

Reference conditions

Units	ISO	Limit conditions ⁴⁾	
Air temperature	°C	25	45
Seawater inlet temperature	°C	18	32
Air pressure ⁵⁾	mbar	1,000	-
Exhaust back pressure ⁶⁾	mbar	50	-
Relative humidity	%	30	60

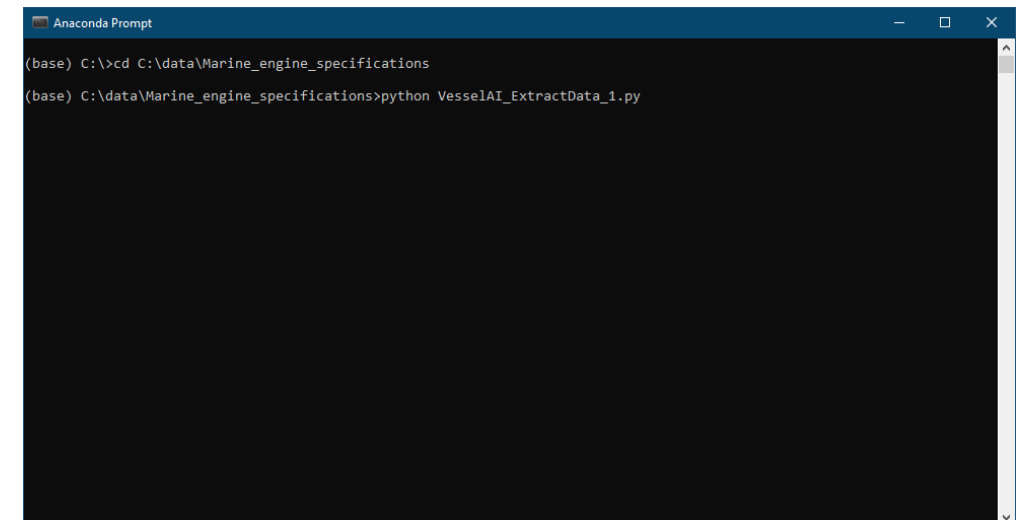
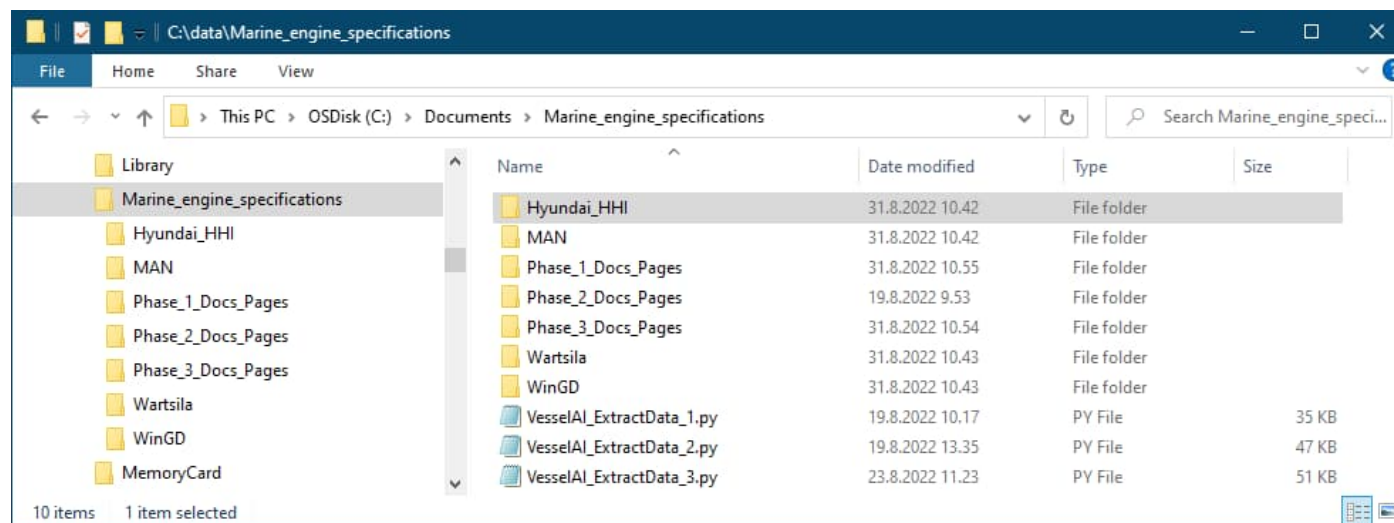
Footnotes:

- ¹⁾ Please contact MAN Energy Solutions if project specific limit conditions might be exceeded.
- ²⁾ Intake air depression up to 30 mbar allowed.
- ³⁾ Reference value for the difference pressure of exhaust gas line (plant) at MCR for IMO Tier II variant.

Table 26: Reference conditions – MAN 175D IMO Tier II

Installation

- The *.zip file contains a folder 'Marine_engine_specifications'
- This folder contains 4 Python files:
 - VesselAI_ExtractData_1.py
 - VesselAI_ExtractData_2.py
 - VesselAI_ExtractData_3.py
 - VesselAI_ExtractData_NN.py
- ...and also subfolders containing PDF files (Hyundai_HHI, MAN, Wartsila, WinGD) as well as empty subfolders where result data will be written (Phase_?_Docs_Pages)
- The PDF files must be downloaded from Internet, see slides 5-9



Installation, continued

- Extract the contents of the *.zip file into a local folder, and then go to the folder 'VesselAI_InfoExtract' (inside the new folder) with your Python Prompt
- In Anaconda Python environment, enter command:
 - `conda env create -f env.yml`
- This will create Anaconda environment 'vesselai_ix', containing the necessary Python program libraries; then activate this new environment:
 - `conda activate vesselai_ix`
- Go to subfolder 'Marine_engine_specifications'
 - `cd Marine_engine_specifications`
- Remember that subfolders 'MAN', 'Wartsila', 'Hyundai_HHI', and 'WinGD' must contain PDF files you have downloaded, as instructed earlier
- Start the Phase 1 program, *_1.py
 - `python VesselAI_ExtractData_1.py`

```

Anaconda Prompt
Collecting oauthlib>=3.0.0
  Using cached oauthlib-3.2.2-py3-none-any.whl (151 kB)
Installing collected packages: tensorboard-plugin-wit, pyasn1, libclang, flatbuffers, wrapt, typing-extensions, termcolor, tensorflow-io-gcs-filesystem, tensorflow-estimator, tensorboard-data-server, rsa, pymupdf, pyasn1-modules, protobuf, opt-einsum, oauthlib, MarkupSafe, markdown, keras, h5py, grpcio, google-pasta, gast, cachetools, astunparse, absl-py, werkzeug, requests-oauthlib, google-auth, google-auth-oauthlib, tensorboard, tensorflow-intel, tensorflow
Successfully installed MarkupSafe-2.1.1 absl-py-1.3.0 astunparse-1.6.3 cachetools-5.2.0 flatbuffers-22.12.6 gast-0.4.0 google-auth-2.15.0 google-auth-oauthlib-0.4.6 google-pasta-0.2.0 grpcio-1.51.1 h5py-3.7.0 keras-2.11.0 libclang-14.0.6 markdown-3.4.1 oauthlib-3.2.2 opt-einsum-3.3.0 protobuf-3.19.6 pyasn1-0.4.8 pyasn1-modules-0.2.8 pymupdf-1.21.1 requests-oauthlib-1.3.1 rsa-4.9 tensorboard-2.11.0 tensorboard-data-server-0.6.1 tensorboard-plugin-wit-1.8.1 tensorflow-2.11.0 tensorflow-estimator-2.11.0 tensorflow-intel-2.11.0 tensorflow-io-gcs-filesystem-0.28.0 termcolor-2.1.1 typing-extensions-4.4.0 werkzeug-2.2.2 wrapt-1.14.1

done
#
# To activate this environment, use
#
#   $ conda activate vesselai_ix
#
# To deactivate an active environment, use
#
#   $ conda deactivate
#

Retrieving notices: ...working... done

(base) C:\VesselAI_Test\VesselAI_InfoExtract>conda activate vesselai_ix

(vesselai_ix) C:\VesselAI_Test\VesselAI_InfoExtract>cd Marine_engine_specifications

(vesselai_ix) C:\VesselAI_Test\VesselAI_InfoExtract\Marine_engine_specifications>python VesselAI_ExtractData_1.py
  
```

Installation, continued

- If you do not have the Anaconda Python, this will probably fail at first, because you have not installed all the necessary libraries to your Python environment
- But when starting the python program fails, the name of the missing library is shown. Then install this library (pip install <name_of_lib>). Continue until *_1.py starts
- Close this *_1.py immediately. Then try *_2.py and *_3.py. Continue installing missing libraries until all 3 programs start
- Then try to go through the whole workflow with default values: start *_1.py:
 - >python VesselAI_ExtractData_1.py
 - *_Phase_1: Maximize, move vertical slider right, press 'Refresh', (wait...), press 'Next >>'
 - *_Phase_2: Maximize, move vertical slider right, press 'Precomp', (wait...), select algorithm 'DBSCAN' from the drop-down, press 'execute', press 'Next >>'
 - *_Phase_3: Maximize, move vertical slider right. Next you should enter relevant line definitions for each cluster, then press 'Refresh', (wait...), press 'Results'.
 - => However, because this is tedious (DBSCAN produces 41 clusters!), delete all the files in folder 'Phase_3_Docs_Pages', and then start 'VesselAI_ExtractData_3.py' directly. And now:
 - *_Phase_3: Maximize, move vertical slider right, press 'Refresh', (wait...), press 'Results'
 - And now Notepad shows file 'Phase3_key_results.txt', containing relevant data lines only!