Supplementary materials for “History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining” published by Solar Energy

D. Yang, J. Kleissl, C. A. Gueymard, H. T. C. Pedro and C. F. M. Coimbra

November 10, 2017

The article “History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining” comes with supplementary materials. This document contains the instructions for reproducing some of the results presented in the article. The code is written in R, and several packages (and, of course, their dependencies, see `http://stat.ethz.ch/R-manual/R-patched/library/tools/html/package.dependencies.html`) need to be installed before the scripts can be executed.

- Package `ggplot2` (`https://cran.r-project.org/web/packages/ggplot2/index.html`) is a system for ‘declaratively’ creating graphics, based on “The Grammar of Graphics”.
- Package `tm` (`https://cran.r-project.org/web/packages/tm/index.html`) provides a framework for text mining applications within R. It is one of the most popular text mining tools.
- Package `stringr` (`https://cran.r-project.org/web/packages/stringr/index.html`) is used to detect text strings based on regular expression.
- Package `tidytext` (`https://cran.r-project.org/web/packages/tidytext/index.html`) manipulates text data with tidy tools.
- Package `dplyr` (`https://cran.r-project.org/web/packages/dplyr/index.html`) is a grammar of data manipulation. It provides simple “verb” functions, such as `select()`, `filter()`, and `mutate()`, that correspond to the most common data manipulation tasks, to help you translate your thoughts into code.
- Other packages include `readtext`, `pluralize`, `extrafont`, `tidyr`, `igraph`, `ggraph`, `ldatuning`, and `topicmodels`.

Section 4 The raw HTML files – from the Google Scholar search “solar + irradiance + PV + power + forecasting” made from Singapore on 2017-07-23 – are provided in the “HTML” folder. These files were used to generate Figs. 1–4 in the paper. Since we plan to use the code in a subsequent paper, it is kept proprietary at the moment. Interested individual can request the code by contacting the corresponding author at `yangdazhi.nus@gmail.com`. Access will be granted on a case-by-case basis.

Section 5 This section performs abbreviation extraction. For demonstration purposes, a sample text file, `SampleText.txt`, and the code, `Abbreviations.R`, are provided in this supplementary material. The sample text file is taken from Ref. [1] with moderate preprocessing. After running the R code, 20 <short form, long form> pairs will be detected; the program output is depicted below. In addition to the abbreviation extraction algorithm, a complete version of Fig. 5 – with both the short and long forms – is also provided.

```
$abbrev
 [1] "PV"     "RPS"    "NREL"   "CAISO"  "DAM"    "RTM"    "VPP"    "NWP"    "MOS"    "GLS"    "MinT"   "OLS"    "WLS"   
[14] "ETS"    "SARIMA" "WWSIS"  "SPDIS"  "BU"     "nRMSE"  "nMBE"  

$full
 [1] "photovoltaic"                                      "renewables_portfolio_standard"                  
 [3] "national_renewable_energy_laboratory"              "california_independent_system_operator"         
 [5] "day_ahead_market"                                  "real_time_market"                               
 [7] "virtual_power_plant"                               "numerical_weather_prediction"                   
 [9] "model_output_statistic"                            "generalized_least_square"                       
[11] "minimum_trace"                                     "ordinary_least_square"                          
[13] "weighted_least_square"                             "exponential_smoothing"                          
[15] "seasonal_autoregressive_integrated_moving_average" "western_wind_and_solar_integration_study"       
[17] "solar_power_data_for_integration_study"            "bottom_up_approach"                             
[19] "normalized_root_mean_square_error"                 "normalized_mean_bias_error"                     

$undetect
character(0)
```

Section 6 Using the preprocessing sequence described in Section 6, the PDF files of six emerging technologies are preprocessed and saved as `EmergingTech.RData`. By running the script `PDF.R`, various plots shown in Section 6 of the paper can be reproduced.

References

[1] Dazhi Yang, Hao Quan, Vahid R. Disfani, and Licheng Liu. “Reconciling solar forecasts: Geographical hierarchy.” *Solar Energy* 146 (2017): 276-286.