Text Analytics using SAS EM

Dr. Goutam Chakraborty

Outline

- This assumes you have been exposed to basics of text analytics
  - Refresh your memory with notes from prior semester
- Goals for this session are
  - Use a Text Import node in SAS EM to read text files from a directory.
  - Use a Text Import node in SAS EM to read web pages.

Working with Text Mining Data Sources

- When documents are stored in separate files, the ***Text Import node*** can be used to create an appropriate SAS data set for text mining.
- When documents are stored together (for example, in a Microsoft Excel spreadsheet), then the Import Data Wizard or File Import node can be used to create a text mining data set.

Working with Text Mining Data Sources

- Two supported types of text mining data:
  - The data set contains at least one variable with the role ***Text***, and documents can be stored completely as a SAS character variable
    (limited to 32K).
  - The data set contains at least one variable with the role ***Text Location***, and some documents cannot be stored completely as a SAS character variable.
    - The location must be the full pathname of the document with respect to the Text Miner server.
    - An additional variable with the role ***Web Address*** can include the path to an unfiltered version of the document to be displayed in an interactive viewer such as the Interactive Filter Viewer.

Text Import Node in Text Miner

- The Text Import node can be used to import source data into Text Miner as a SAS data set.
- The source data can exist in a directory in any proprietary file format type. More than 100 file formats are supported by the Text Import node. Here are some of them:
    - Microsoft Word (.doc, .docx)
    - Microsoft Excel (.xls, .xlsx)
    - Microsoft PowerPoint (.ppt, .pptx)
    - Rich Text (.rtf)
    - Adobe Acrobat (.pdf)
    - ASCII Text (.txt)

Text Import Examples

- We will demonstrate the following:
  - How to extract text from multiple files and create a SAS data set
  - How to retrieve data from the web and create a SAS data set

Demo Using SAS EM

- Follow handout titled "How to Get Textual Data in SAS EM"

What if you have data in other formats?

- XML format:
  - You will need XML mapper (free download from SAS) and then you can import into SAS EM (https://support.sas.com/downloads/package.htm?pid=486)
  - Or, use SAS codes (https://support.sas.com/resources/papers/proceedings17/1318-2017.pdf)
- JSON format:
  - Use Proc JSON (https://support.sas.com/rnd/base/Tipsheet_PROC_JSON.pdf)
- You should also learn to use Python Scripts to scrape web sites.