

**MIS 506 – Module 2 – Assignment 2**  
**Pre-Processing Text for Quantitative Text Analytics**  
**Orange Data mining Tool**

In this assignment you will complete a variety of tasks related to cleaning and preprocessing text using the Orange Data mining Tool. For this assignment, save your work under “Cleaning Text” (or similar) folder from last assignment.

**Submission Instruction:** Submit your OWS file and a copy of the Word Cloud, and the Bar Plot of word frequencies after preprocessing the text.

**Deliverables:**

**One:** Download and Install Orange Data Mining Software:

<https://orangedatamining.com/download/>

**Two:** Start the orange software → Close the welcome window

**Three:** Install Text mining Add-on: Open Orange Software → Go to Options tab → Add-ons → Text → Restart Orange for the Add-On to appear. Save your file under “Cleaning Text” (or similar) folder – keep saving your work throughout the assignment.

**Four:** Drag the **CSV File Import Widget** into the Canvas → double click it and Import the CSV file from Asheville Airbnb Reviews from your computer

- **Dataset:** Asheville Listing Reviews

**Five:** Drag the **Corpus Widget** into the Canvas → double click it and choose the text feature you want to use to analyze → You can read your Corpus into the **Corpus Viewer Widget** and explore the text data file.

**Six:** Drag the **Preprocess Text Widget** into the Canvas and connect it to the Corpus Widget → double click it and choose appropriate settings to tokenize and transform the words to their base forms

- You can follow the steps from the example in the module and prepare the text for analysis by tokenizing, removing the stop words, numbers, whitespaces, special characters, and any other necessary steps – you can create a text file for the undesirable words where you write each word on a separate line and import it as a stop word file.

**Seven:** Drag the **Word Cloud Widget** into the Canvas to see the most frequent words of this corpus → [Save the Word Cloud on your computer and upload it to Canvas](#)

**Eight:** Drag the **Data Table Widget** from the Data Category into the Canvas – Word Cloud has an output called Word Counts, which outputs a table with words and their frequencies in columns

→ connect Data Table Widget to the Word Cloud Widget → Double click the connection to connect Word Counts to Data

**Nine:** Drag the **Bar Plot Widget** from the Visualization Category into the Canvas → Connect Bar Plot Widget to the Word Cloud Widget → Double click the connection to connect Word Counts to Data → **Save the Bar Plot on your computer and upload it to Canvas**

**Ten:** Tell us about your experience with Orange Data Mining tool. Some possible things to post about: What was interesting or confusing about the tool? What are the advantages and disadvantages of tools like Orange vs. Programming with R for text analytics? Any other thoughts?

