Algorithm for a Classification Tool

Objective

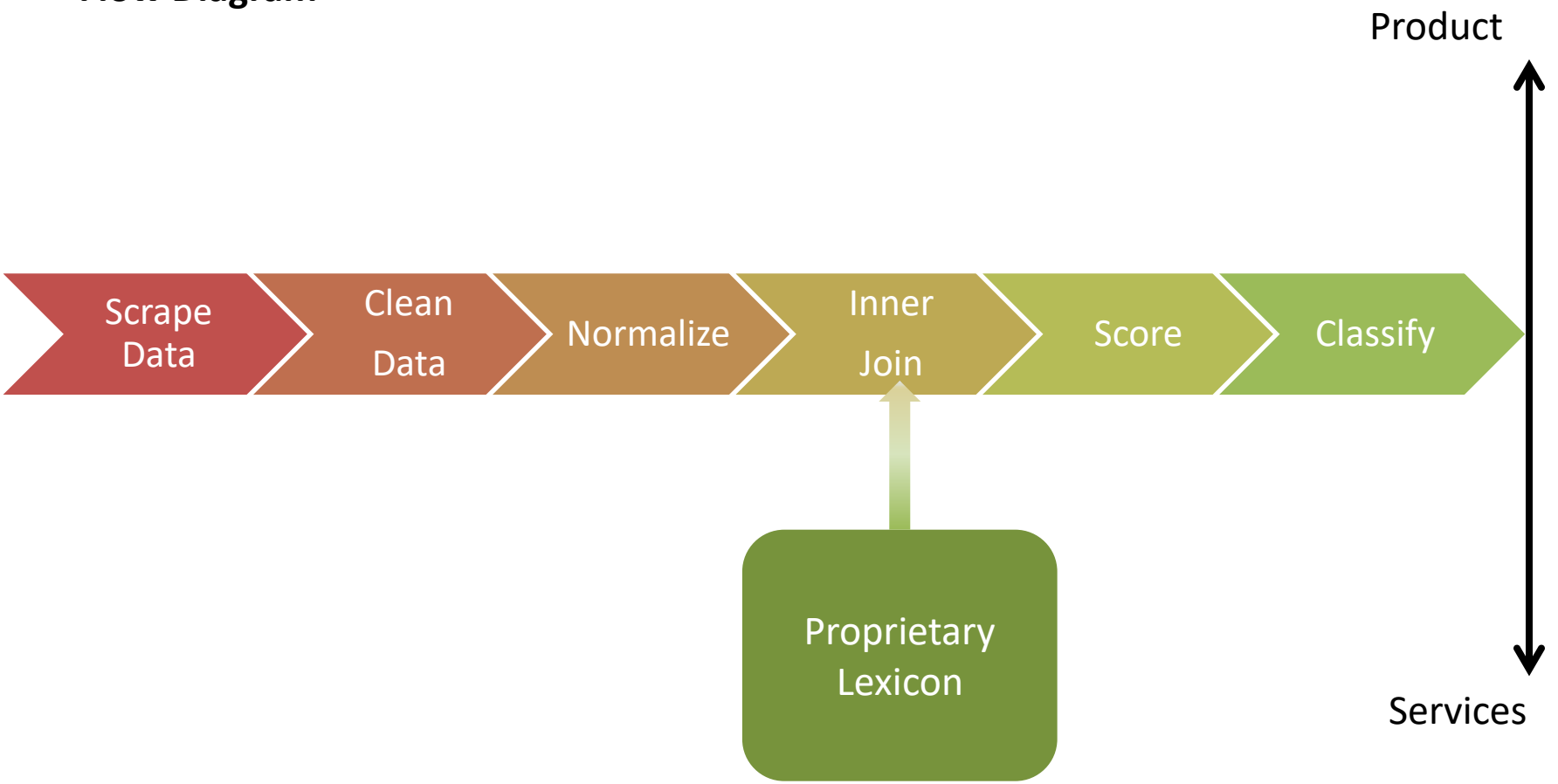
A method to classify a website as:

- Product
- Service

Assumptions:

- Product and Service set is Collectively Exhaustive
- Scale is Linear
- A Taxonomy development is a given skill of the set up

Flow Diagram



Algorithm for the Workflow

1. IDE – Set up an Environment (R/Python)
2. Scraping Package: For R use `rvest`; for Python use selenium
3. Taxonomy Package (Proprietary): Develop a separate
4. Classification Algorithms: -- Compare Algorithms RMSE
5. Scoring:
    - Assigning the depth and width of Hierarchical clusters
    - Determining cut-offs as per **Business case** in practical scenario
    - Plotting on the Product/Service scale
6. Fitting, Data Massaging
7. Compile Results
8. Manually revisit step 4-6 to determine improvements/change in Algorithms Tweaking the Taxonomy libraries (Naive Bayes Classifiers)

Taxonomy Development

Model: Bag of Words
Algorithms: Decision Trees, KNN, Bayes
Visualization: Dendrograms

Order of operations:

1. Create a corpus
2. Tidy the corpus
3. Create TDM
4. Create TF-IDF
5. Score the Terms (say a Likert of 0-7 for Product to Service)
6. Cluster the segments (use Bayes, KNN)
7. Build the Lexicon
8. Revisit as the corpus grows reliably