Data Mining and Business Intelligence

Lecture 5: Text Mining Applications

Jing Peng
University of Connecticut

2/20/20

Recap

- Which approach is more time consuming, stemming or lemmatization?
- In the TF-IDF matrix, what do TF and IDF capture, respectively?

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

Interpretations of Matrices Produced by SVD

- **$A = U \Sigma V^T$ - example: Users to Movies**

SciFi-concept

Romance-concept

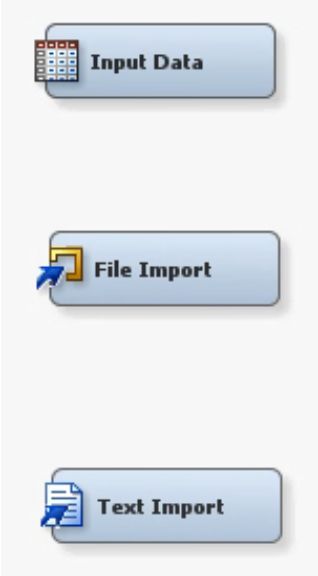
$$
\begin{array}{c}
\\ \text{SciFi} \\ \\ \\ \text{Romance} \\ \\ \\
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
0.13 & 0.02 & -0.01 \\
0.41 & 0.07 & -0.03 \\
0.55 & 0.09 & -0.04 \\
0.68 & 0.11 & -0.05 \\
0.15 & -0.59 & 0.65 \\
0.07 & -0.73 & -0.67 \\
0.07 & -0.29 & 0.32
\end{bmatrix}
\times
\begin{bmatrix}
12.4 & 0 & 0 \\
0 & 9.5 & 0 \\
0 & 0 & 1.3
\end{bmatrix}
\times
\begin{bmatrix}
0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\
0.40 & -0.80 & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

Jure Leskovec, Stanford CS246: Mining Massive Datasets

Text Mining in SAS

Loading Data

- **Input Data** node: SAS dataset (automatically created while dragging it from Data Sources to the diagram)
- **File Import** node: non-SAS format dataset (.xlsx format is recommended over .csv for text data)
- **Text Import** node: a collection of texts stored in separate files (each file represents one record in the data)



Text Mining Nodes

Step	Action	Description	Tools
1	File Preprocessing	Create a SAS data set from a document collection that is used as input for the **Text Parsing** node.	**Text Import** node, %TMFILTER macro, or SAS DATA step.
2	Text Parsing	Decompose textual data, and generate a quantitative representation that is suitable for data mining purposes. Parsing might include: • stemming • automatic recognition of multi-word terms • normalization of various entities such as dates, currency, percent, and year • part-of-speech tagging • extraction of entities such as organization names, product names, and addresses • support for synonyms • language-specific analyses	**Text Parsing** node
3	Text Filtering	Transform the quantitative representation into a compact and informative format; reduce dimensions.	**Text Filter** node
4	Document Analysis	Cluster, classify, predict, or link concepts.	**Text Topic** node, **Text Cluster** node, **Text Rule Builder** node, **Text Profile** node, and SAS Enterprise Miner predictive modeling nodes

SVD in SAS (Text Cluster Node)

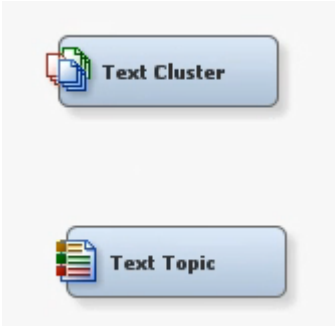
- **Max SVD Dimension**: Maximum allowed latent dimensions (actually used dimension is often smaller) for words

- **Number of Clusters**: Maximum number of clusters for documents
  - Exact or Maximum Number: Whether to create an exact or maximum number of clusters

- Is it possible to have more clusters than SVD dimensions?

TextCluster vs. TextTopic

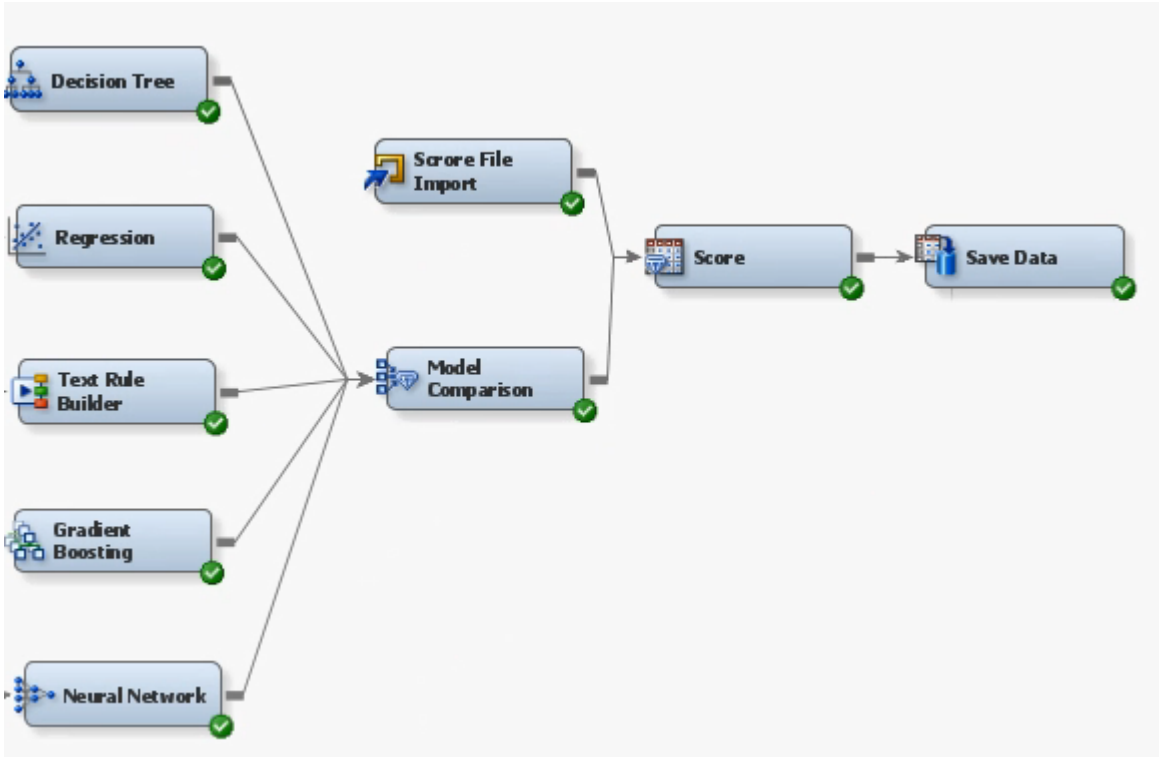
- **TextCluster node**
  - Performs SVD and Clustering
  - Soft group assignment



- **TextTopic node**
  - Performs clustering, but only retain terms and documents whose relevance exceeds certain thresholds
  - Hard group assignment

- Bug: SAS incorrectly labels TextCluster_prob as binary and TextTopic as interval.

Model Selection and Prediction



Save Data Node

Output Options → Variables: specify which variables you want to save

Output Data → Select Roles: specify which subset to save (train, valid, test, or score)

Output Format: which format to save as

Property	Value
General	
Node ID	EMSave
Imported Data	
Exported Data	
Notes	
Train	
Output Options	
Variables	
Filename Prefix	dt5
Replace Existing Files	Yes
All Observations	Yes
Number of Observations	1000
Output Format	
File Format	Excel Spreadsheet (.xlsx)
SAS Library Name	LECTURE1
Directory	P:\lecture10\Data
Output Data	
All Roles	No
Select Roles	
Status	
Create Time	11/3/16 3:16 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

Tweets Popularity Prediction

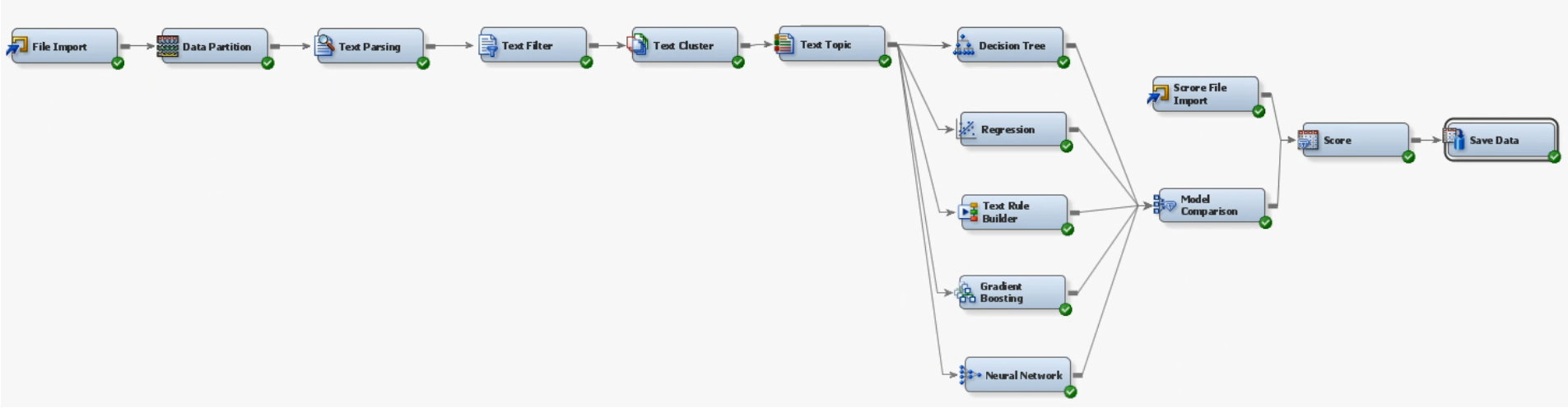
Data Collection

- 12 funniest brands on Twitter

- We focus on the top 3 of them
  - Moosejaw (screen name: MoosejawMadness)
  - Netflix (screen name: netflix)
  - KFC (screen name: kfc)

- Run brand_tweets.R

Enterprise Miner Diagram



Popularity.xml

Steps: Loading and Parsing Data

- Add **File Import** node (set is_popuar as a binary target variable)
- Append **Data Partition** node (70% train and 30% validation)
- Append **Text Parsing** node (default settings)
- Append **Text Filter** node (Filter Viewer)
  - Concept linking in Filter Viewer (double click to expand)
  - Add Term to Search Expression (remember to clear)
  - Treat as Synonyms (select multiple → right click)

Steps: Clustering and Modeling

- Append **Text Clustering** node (SVD Resolution: High, max SVD Dimensions: 20; Number of Clusters: 3)

- Add **Text Topic** node (Number of Multi-term Topics: 3) → Results (Descriptive Terms) → Topic Viewer

- Append **Decision Tree, Regression, Text Rule Builder, Gradient Boosting,** and **Neural Network** models

Steps: Evaluation and Prediction

- Add **Model Comparison** node for all models (selection statistic: misclassification rate; selection table: validation)

- Add **File Import** node to import score data (Role: Score). Append **Score** node to this new File Import node and **Model Comparison** node

- Append **SaveData** node to **Score** node (Filename Prefix: predictions; File Format: xlsx; Choose Directory)

- Press Ctrl + S in the diagram window to save the diagram as xml file

Readings

- **Highly Recommended**: Getting Started with SAS Text Miner
- SAS Enterprise Miner Documentation:  (press F1)
- Text Mining with R: ebook