Mohammad Saad
Applied Machine Learning NLP Decision Tree Report

Details:

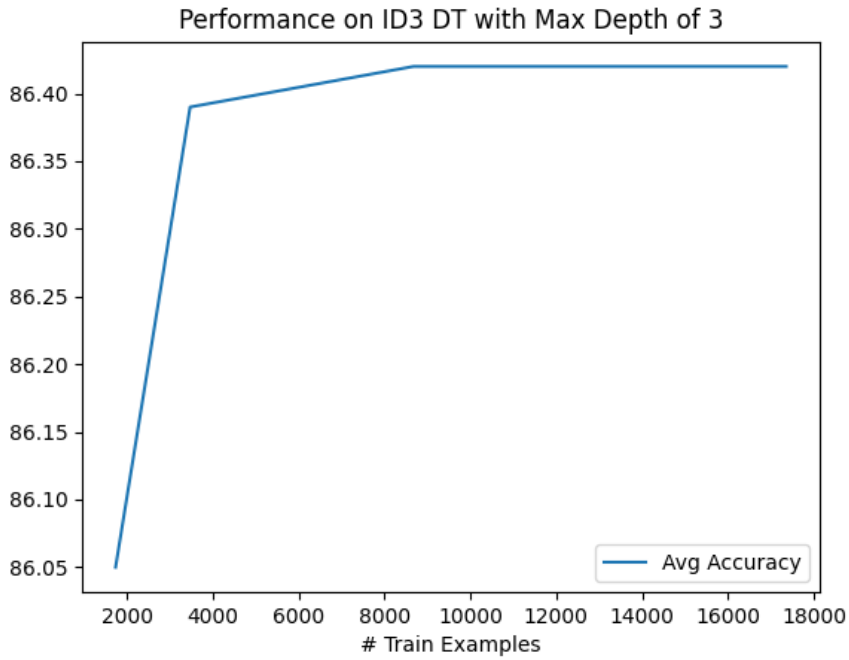
- Features/Attributes:
  - We will be using 2 bag-of-words features. 1 for the word that appears before, and another for the word that appears after.
  - For example: `The_Before` attribute will have a value of True or False depending on whether the word 'the' appears before the labeled word.
  - Because the size of the dictionary is very large, we will only have attributes for the top 50 words that appear before and top 50 that appears after the labeled word.
  - This means that in total we will have 100 attributes
- Preprocessing:
  - we need to create a dictionary of the words that appear before and another dictionary for the word that appear after the labeled word. Then we need to extract the top 50 most occurring words from the 2 dictionaries
  - we will parse the training data file and create a table. This will help organize our data and also allow us to use the pandas library to make operations on the table easier.
  - The table will have a column for each of the attributes and an extra column to store the label (weather/whether).
  - Each row in the table will correspond to a single training example
  - we will also create a table for the test examples in a similar fashion
- Decision Tree:
  - Each node in the tree will have a table. The root will have the full table (full training data set) and the children's nodes will have a sub-table (or subset) of their parent node's table.
  - We will build the tree using the ID3 decision tree algorithm
  - The stopping condition for building the tree is if all the labels in the table of the current node are homogenous (all the same label). Another stopping condition is if we pass the max depth allowed for the tree.
  - NOTE: the functions in the decision tree class are not recursive since python has a max recursive depth. To work around this, the functions will be implemented iteratively rather than recursively.
- Evaluation:
  - We will be evaluating 3 decision trees with max depths 3, 5, and 10
  - for each run, we will generate a new tree and test on random samples of the training data set. The sample sizes are 10%, 20%, 50%, 80%, and 100%
  - also whenever we test our tree on the sample we will average the accuracy over 3 trials.
  - We store number of training examples and the average accuracy in a table and print it out at the end of the program
  - We will also plot the relationship between the number of training examples and the average accuracy.

Mohammad Saad

Run 1: Max Tree Depth = 3:

RESULT:

	# Train Examples	Avg Accuracy
0	1735	86.05
1	3471	86.39
2	8678	86.42
3	13885	86.42
4	17357	86.42

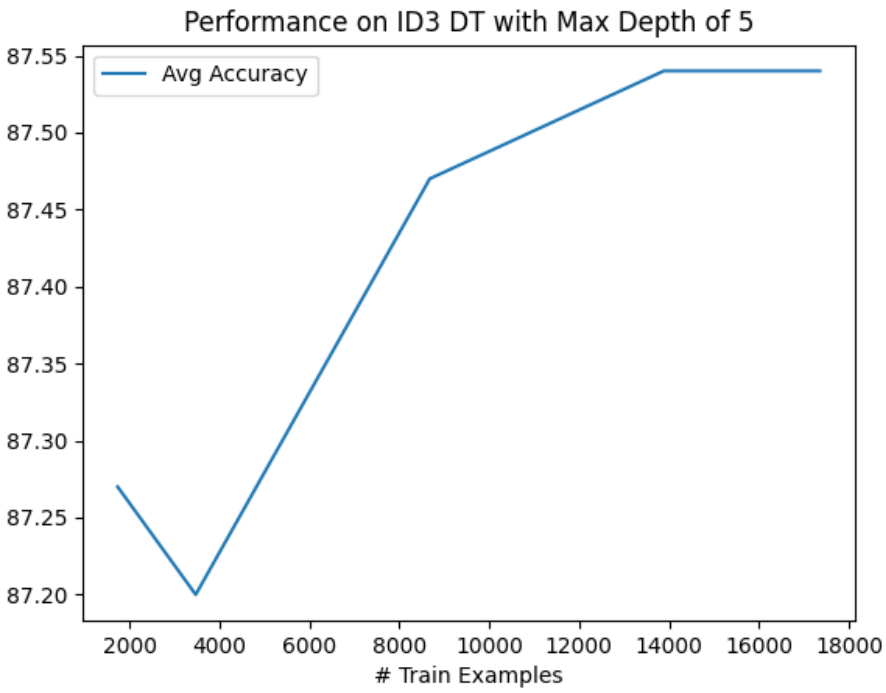


Mohammad Saad

Run 2: Max Tree Depth = 5:

RESULT:

	# Train Examples	Avg Accuracy
0	1735	87.27
1	3471	87.20
2	8678	87.47
3	13885	87.54
4	17357	87.54



Mohammad Saad

Run 3: Max Tree Depth = 10:

RESULT:

	# Train Examples	Avg Accuracy
0	1735	88.04
1	3471	88.75
2	8678	88.62
3	13885	88.58
4	17357	88.55