Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

AY:2024-25

Class:	BE	Semester:	VII
Course Code:	CSDOL7011	Course Name:	Natural Language Processing

Name of Student:	**Parth Raut**
Roll No.:	**40**
Experiment No.:	**6**
Title of the Experiment:	**Chunking: Importance of Feature Selection & Training Size**
Date of Performance:	
Date of Submission:	

Evaluation

Performance Indicator	Max. Marks	Marks Obtained
Performance	5	
Understanding	5	
Journal work and timely submission	10	
Total	20	

Performance Indicator	Exceed Expectations (EE)	Meet Expectations (ME)	Below Expectations (BE)
Performance	4-5	2-3	1
Understanding	4-5	2-3	1
Journal work and timely submission	8-10	5-8	1-4

Checked by

Name of Faculty :

Signature :

Date :

Experiment 6

Aim: Perform chunking by analyzing the importance of selecting proper features for training a model and size of training.

Objective: To study POS Tagging and tag the part of speech for given input in english and an Indian Language.

Theory:

Chunking in machine learning refers to breaking down the dataset or tasks into smaller, manageable pieces for processing. When analyzing the importance of selecting proper features for training a model and the size of training, both play crucial roles in model performance and computational efficiency.

Importance of Selecting Proper Features

1. **Improves Model Accuracy:**
   - **Relevance:** Properly selected features that are highly relevant to the target variable enhance the model's ability to learn patterns, leading to better predictions.
   - **Noise Reduction:** Irrelevant or redundant features can introduce noise, making it harder for the model to find the underlying patterns.
2. **Reduces Overfitting:**
   - **Simplicity:** A model with fewer but more meaningful features is less likely to overfit the training data. Overfitting happens when the model learns the noise in the training data rather than the actual pattern.
   - **Generalization:** Proper feature selection helps in better generalization to unseen data, improving the model's performance on the test set.
3. **Enhances Computational Efficiency:**
   - **Reduced Complexity:** Fewer features mean less computational resources are needed for training the model, which can be critical when dealing with large datasets.
   - **Faster Training:** With fewer features, the model trains faster, allowing for quicker iterations and faster tuning of hyperparameters.
4. **Improves Interpretability:**

- **Simpler Models:** Models with a smaller number of relevant features are easier to interpret, which is crucial for understanding the decision-making process of the model, especially in regulated industries.

Importance of Training Size

1. **Model Performance:**
    - **Bias-Variance Tradeoff:** Larger training datasets help in reducing variance and improving the model's ability to generalize. A small training set might lead to high variance, while an extremely large set might reduce bias.
    - **Diverse Patterns:** A larger training dataset is likely to capture more diverse patterns, providing the model with a broader range of scenarios to learn from.
2. **Robustness:**
    - **Outlier Impact:** Larger datasets can mitigate the impact of outliers. With more data, outliers are less likely to distort the overall learning process.
    - **Error Reduction:** Larger datasets help in reducing errors in predictions by providing more information for the model to learn from.
3. **Training Time and Resources:**
    - **Balance Needed:** While larger datasets improve model performance, they also require more computational resources and longer training times. There is a tradeoff between the size of the training set and the available computational resources.
    - **Incremental Learning:** In some cases, chunking the training set into smaller batches and using techniques like mini-batch gradient descent can optimize training time and resource usage.
4. **Data Quality:**
    - **Quality over Quantity:** It's not just the size, but the quality of the training data that matters. A smaller, high-quality dataset can be more beneficial than a large, noisy dataset.
    - **Feature Engineering:** Proper feature engineering on a well-curated dataset can often outperform models trained on larger, less relevant data.

Code:

In [1]:
```
import nltk
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\admin\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\admin\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

Out[1]:
```
True
```

In [2]:
```
from nltk.chunk import RegexpParser
from nltk.tokenize import word_tokenize
```

In [3]:
```
sentence = "Educative Answers is a free web encyclopedia written by devs for devs."
```

Tokenization

In [4]:
```
tokens = word_tokenize(sentence)
```

In [5]:
```
tokens
```

Out[5]:
```
['Educative',
 'Answers',
 'is',
 'a',
 'free',
 'web',
 'encyclopedia',
 'written',
 'by',
 'devs',
 'for',
 'devs',
 '.']
```

POS tagging

In [6]:
```
pos_tags = nltk.pos_tag(tokens)
```

In [7]:
```
pos_tags
```

Out[7]:
```
[('Educative', 'JJ'),
 ('Answers', 'NNPS'),
 ('is', 'VBZ'),
 ('a', 'DT'),
 ('free', 'JJ'),
 ('web', 'NN'),
 ('encyclopedia', 'NN'),
 ('written', 'VBN'),
 ('by', 'IN'),
 ('devs', 'NN'),
 ('for', 'IN'),
 ('devs', 'NN'),
 ('.', '.')]
```

Chunking patterns

In [8]:
```
chunk_patterns = r"""
    NP: {<DT>?<JJ>*<NN>}  # Chunk noun phrases
    VP: {<VB.*><NP|PP>}  # Chunk verb phrases
"""
```

In [9]:
```
chunk_patterns
```

Out[9]:
```
'\n    NP: {<DT>?<JJ>*<NN>}  # Chunk noun phrases\n    VP: {<VB.*><NP|PP>}  # Chunk verb phrases\n'
```

Create a chunk parser

In [10]:
```
chunk_parser = RegexpParser(chunk_patterns)
```

In [11]:
```
chunk_parser
```
Out[11]:
```
<chunk.RegexpParser with 2 stages>
```

Perform chunking

In [12]:
```
result = chunk_parser.parse(pos_tags)
```

In [13]:
```
print(result)
```
```
(S
  Educative/JJ
  Answers/NNPS
  (VP is/VBZ (NP a/DT free/JJ web/NN))
  (NP encyclopedia/NN)
  written/VBN
  by/IN
  (NP devs/NN)
  for/IN
  (NP devs/NN)
  ./.)
```

Conclusion

In summary, both proper feature selection and the size of the training dataset are vital for developing effective machine learning models. Proper feature selection enhances accuracy, reduces overfitting, improves computational efficiency, and aids interpretability. A larger training size helps in capturing diverse patterns, reducing errors, and improving model robustness. However, a balance must be struck between dataset size, computational resources, and data quality to achieve optimal model performance.