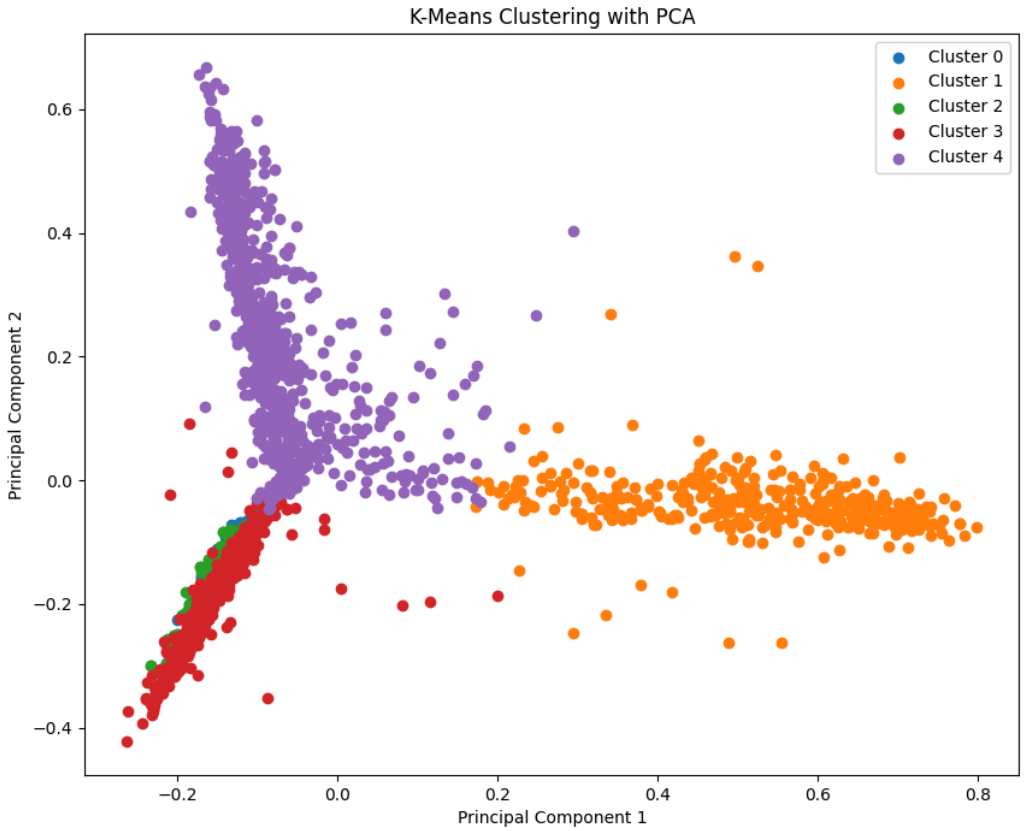
Web Scraping Analysis

Clustering (K-Means)

Note- First clustering was 5, later adjusted to 3 based on the results from elbow test. The results from the sentiment scores and below are based off the 3-cluster result.

Parameters

- **Number of clusters- 5**



Cluster Analysis

1. **Number of Articles in Each Cluster:**
    - **Cluster 0:** 103 articles
    - **Cluster 1:** 378 articles
    - **Cluster 2:** 174 articles
    - **Cluster 3:** 638 articles
    - **Cluster 4:** 827 articles

2. **Topics Covered in Each Cluster:**

- **Gardening**
    - **Garden Tools**
        1. **Cluster 0:** Covers topics about garden tools, their uses and best practices. Topics in this cluster include: *['wheelbarrow'], ['garden', 'scissors'], ['dibber'], ['expandable', 'garden', 'hose', 'adjustable', 'sprayer'], ['stool', 'knee', 'pad'].*
    - **Gardening**
        1. **Cluster 2:** Covers topics on gardening maintenance, preparation, garden tools storage and maintenance. The topics covered include: **Cluster 2:** Covers topics on gardening maintenance, preparation, garden tools storage and maintenance. The topics covered include:
            a. **Plant Watering and Lawn Care:** ['plant', 'watering', 'lawn', 'care', 'weed'], ['gardening', 'tip', 'maintaining', 'lawn', 'healthy']
            b. **Winter Preparation:** ['making', 'outside', 'water', 'tap', 'winterproof'], ['right', 'preparation'], ['cutting', 'back', 'bush', 'hedge', 'tree']\
            c. **Storing Garden Equipment:** ['put', 'battery', 'lawn', 'mower', 'robotic', 'winter'], ['instruction', 'storing', 'highpressure', 'cleaner'], ['storing', 'garden', 'tool', 'correctly']
            d. **Garden Tool Maintenance:** ['removing', 'rust'], ['regularity'], ['maintaining', 'bed']
            e. **General Gardening Tips:** ['tip', 'leaf', 'go', 'exemplary', 'disposal'], ['considering','need', 'plant', 'soil']
            f. **Plant Watering and Lawn Care:** ['plant', 'watering', 'lawn', 'care', 'weed'], ['gardening', 'tip', 'maintaining', 'lawn', 'healthy']
            g. **Winter Preparation:** ['making', 'outside', 'water', 'tap', 'winterproof'], ['right', 'preparation'], ['cutting', 'back', 'bush', 'hedge', 'tree']\
            h. **Storing Garden Equipment:** ['put', 'battery', 'lawn', 'mower', 'robotic', 'winter'], ['instruction', 'storing', 'highpressure', 'cleaner'], ['storing', 'garden', 'tool', 'correctly']
            i. **Garden Tool Maintenance:** ['removing', 'rust'], ['regularity'], ['maintaining', 'bed']
            j. **General Gardening Tips:** ['tip', 'leaf', 'go', 'exemplary', 'disposal'], ['considering','need', 'plant', 'soil']
        2. **Cluster 3:** this covers growing and caring for various plants. Topics covered are:
            a. **Vegetable and Herb Gardening:** ['repot', 'tomato', 'seedling', 'bigger', 'better'], ['conquer', 'blossom', 'end', 'rot', 'save', 'tomato'], ['grow', 'fava', 'bean', 'cover', 'crop', 'get', 'started']

b. **Planting Techniques:** ['grow', 'garlic', 'big', 'yield', 'easy', 'beginner'], ['plant', 'strawberry', 'patch', 'bare', 'root'], ['grow', 'tomato', 'potseven', 'without', 'garden']
c. **Harvesting Tips:** ['harvest', 'fava', 'bean', 'leaf', 'flower', 'vegetable'], ['harvest', 'fennel', 'pollen', 'rare', 'expensive', 'delicious'], ['harvest', 'garlic', 'scape', 'easy', 'way', 'gourmet']
d. **Pest Control and Plant Care:** ['make', 'recycled', 'newspaper', 'pot', 'seedling', 'starter'], ['brew', 'compost', 'tea', 'better', 'plant', 'growth'], ['deerresistant', 'plant', 'flower', 'fawn', 'avoid']
e. **Gardening Tips:** ['best', 'fertilize', 'tomato', 'ultimate', 'beginner'], ['much', 'plant', 'year', 'worth', 'food'], ['plant', 'windowsill', 'create', 'charming', 'indoor']

- **Pets**
  i. **Cluster 1:** Coves topics about cats, their health, care, behaviour and nutrition. The topics covered include- *['daschund', 'grooming', 'guide', 'groomers'], ['manx', 'cat'], ['regular', 'grooming', 'key'], ['monitor', 'skin', 'condition'], ['regular', 'veterinary', 'checkup'].*

  ii. **Cluster 4:** covers dog care, encompassing various aspects of dog health, grooming, nutrition, and training.
    1. **Dog Grooming and Care:** ['ear', 'care'], ['brushing'], ['recommended', 'shampoo', 'rottweiler']
    2. **Dog Nutrition:** ['papaya', 'pineapple', 'healthy', 'fruit', 'dog'], ['right', 'age', 'switch', 'dog', 'food'], ['key', 'food', 'help', 'support', 'dog', 'mental', 'health']
    3. **Dog Training:** ['effective', 'dog', 'training', 'technique', 'professor', 'akira'], ['brain', 'training', 'dog', 'founded', 'monica', 'elkhalifa']
    4. **Dog Health:** ['monitor', 'skin', 'condition'], ['regular', 'veterinary', 'checkup'], ['skin', 'problem']

Sentiment Score

The sentiment score is grouped into clusters. The elbow test shows 3 clusters is the optimal number to use sentiment test is as follows:

Cluster:

Most articles are in Cluster 0 (1246 articles), followed by Cluster 2 (455 articles) and Cluster 1 (419 articles).

Sentiment Scores per Cluster:

Cluster 0

- **Low Sentiment Score**: -0.60
- **High Sentiment Score**: 0.75
- **Average Sentiment Score**: 0.129

Cluster 1

- **Low Sentiment Score**: -0.60
- **High Sentiment Score**: 0.75
- **Average Sentiment Score**: 0.301

Cluster 2

- **Low Sentiment Score**: -0.60
- **High Sentiment Score**: 0.75
- **Average Sentiment Score**: -0.029

Summary of Topics Covered in Each Cluster

Cluster 0

- **Dominant Topics**:
    - `cat` (253 occurrences)
    - `dog` (154 occurrences)
    - `plant` (99 occurrences)
    - `garden` (76 occurrences)
    - `tip` (66 occurrences)
- **Summary**: This cluster primarily focuses on pets (cats and dogs) and gardening, including tips related to plant care and gardening practices.

Cluster 1

- **Dominant Topics**:
    - `cat` (36 occurrences)
    - `dog` (56 occurrences)
    - `plant` (37 occurrences)
    - `garden` (39 occurrences)
    - `best` (17 occurrences)
    - `retriever` (29 occurrences)
    - `Golden` (23 occurrences)
- **Summary**: Similar to Cluster 0, this cluster also emphasizes pets and gardening. Additionally, it highlights specific breeds like retrievers and aspects of quality or "best" practices/items.

Cluster 2

- **Dominant Topics**:
    - `cat` (64 occurrences)
    - `dog` (24 occurrences)
    - `care` (14 occurrences)
    - `cutting` (19 occurrences)
    - `food` (20 occurrences)
    - `garden` (19 occurrences)
    - `plant` (19 occurrences)

Chatgpt analysis

Top 10 Positive Correlations between Topics and Sentiment Score:

1. **retriever**: 0.135
2. **golden**: 0.125
3. **dog**: 0.092
4. **best**: 0.068

5. **garden**: 0.057
6. **tip**: 0.054
7. **plant**: 0.049
8. **watering**: 0.034
9. **care**: 0.018
10. **guide**: 0.006

Top 10 Negative Correlations between Topics and Sentiment Score:

1. **cutting**: -0.085
2. **lawn**: -0.050
3. **eat**: -0.042
4. **cat**: -0.030
5. **video**: -0.024
6. **food**: -0.017
7. **way**: -0.013
8. **pet**: -0.010
9. **grow**: -0.002
10. **guide**: 0.006 (appears twice due to its low impact)

Outliers

There were a total of 43 outliers and non of them were from cluster 0.

Cluster 1:

Outlier Sentiment Mean: 0.546

Sentiment Mean: 0.301

Cluster 2

Outlier Sentiment Mean: -0.321

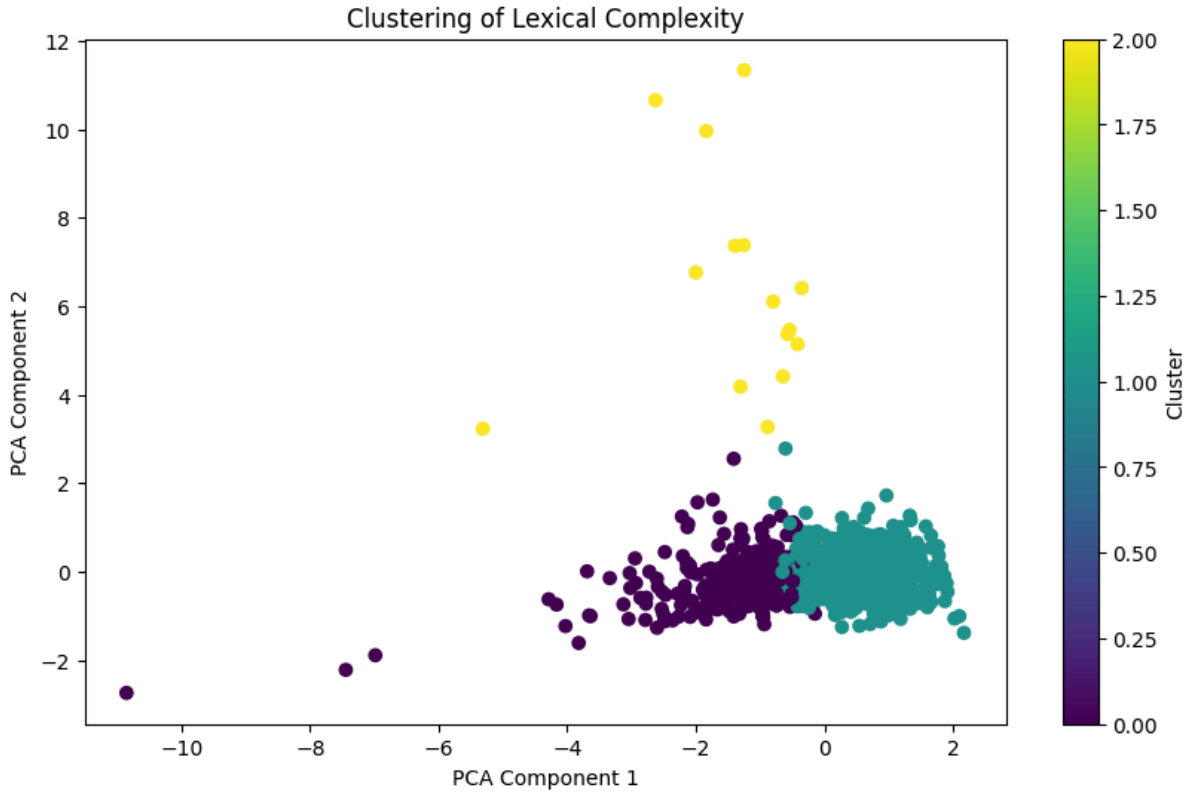
Sentiment Mean: -0.029

Cluster 1 outliers seem to have a more positive connotation compared to others inthe cluster. Topics in the outliers included words that lean to a positive sentiment score I.E. best garden equipments, safe dog spaces, While cluster 2 which leaned more to the negative end covered topics such as skin infections, managing diseases, allergies etc. This is not indicitive of the topics themselves as the topics in itself isnt negative.

Lexical Complexity

Parameters for lexical complexity:

- Average word length
- Average Sentence length
- Lexical Diversity (Using Type-Token Ratio:ratio of the number of unique words (types) to the total number of words (tokens) in a text.)



Cluster 0:

- **Number of Articles:** 1036
- **Average Word Length:** 6.02 (range: 4.00 to 9.94)
- **Average Sentence Length:** 102.80 (range: 1 to 459)
- **Lexical Diversity:** 0.79 (range: 0.575 to 1.000)

Cluster 1:

- **Number of Articles:** 1069
- **Average Word Length:** 5.93 (range: 4.66 to 8.61)
- **Average Sentence Length:** 610.53 (range: 6 to 3245)
- **Lexical Diversity:** 0.53 (range: 0.313 to 0.720)

Cluster 2:

- **Number of Articles:** 15

- **Average Word Length:** 14.75 (range: 10.65 to 20.92)
- **Average Sentence Length:** 89.07 (range: 24 to 411)
- **Lexical Diversity:** 0.67 (range: 0.353 to 0.805)

Cluster 0:

Word	Frequency
Garden	142
Also	99
Water	75
Use	69
Like	61
Time	58
Tool	56

Cluster 1:

Word	Frequency
Plant	442
Garden	138
Soil	122
Leaf	118
Like	69
Water	65
Grow	61

Cluster 2:

Word	Frequency

Dog	413
Pet	77
Help	46
Need	46
Time	46
May	44
Keep	40

Cluster 3:

Word	Frequency
Cat	337
May	117
Food	75
Also	68
Help	47
Need	39
Time	37

Cluster 4:

Word	Frequency
Pet	153
Cat	53
Food	48
Might	27
Also	24
Health	18

Make	14

Outliers

Total number of outliers in the lexical complexity clusters are 56 with almost all of cluster 2 in the outliers (13 out of 15). The outliers had a higher average word length- 12.71 compared to 5.54 for the non outliers. The also featured less lexical diversity- 0.73 compared to 0.85 of the non-outliers. Regarding cluster 2, the topics were almost all topics covering gardening community and a few articles on sustainable gardening.