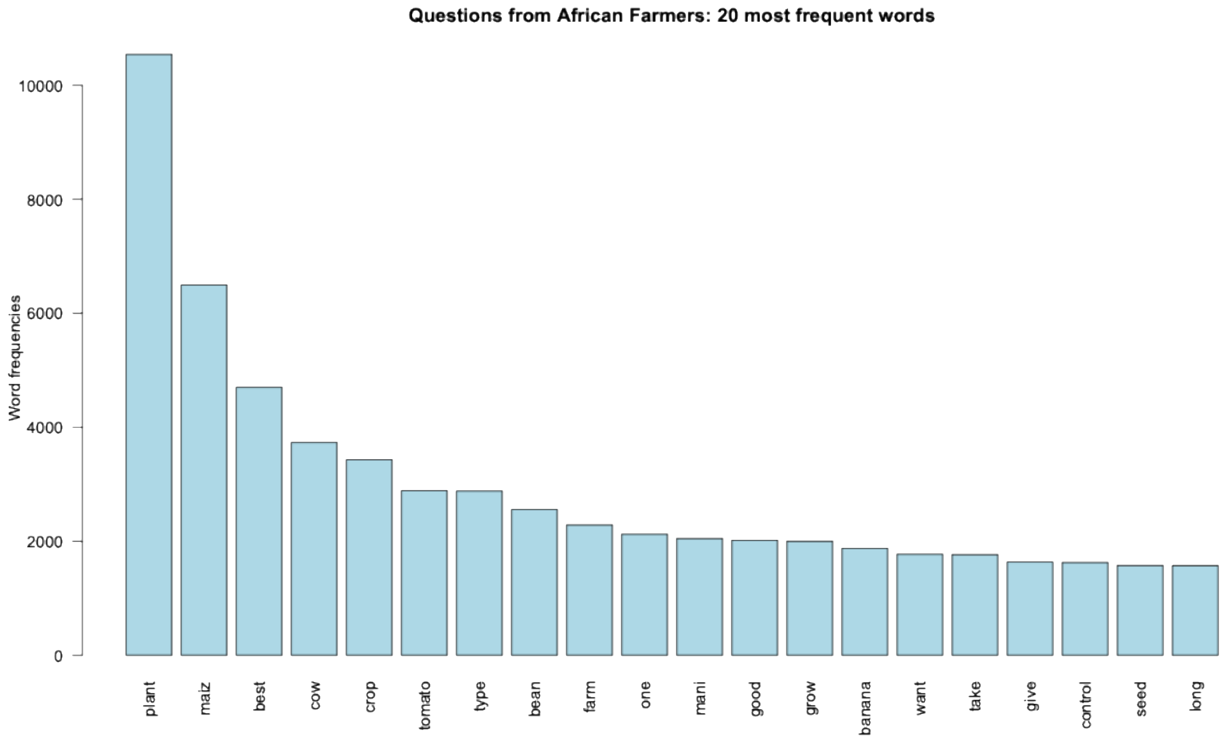
Small Holder Farmers Text Analysis for Questions

Holmes Finch

Most common words in farmers’ questions

- The 20 most common words in the farmers’ question posts appear in this graph.

- The 5 most common words were “plant”, “maize”, “best, “cow”, and “crop”.

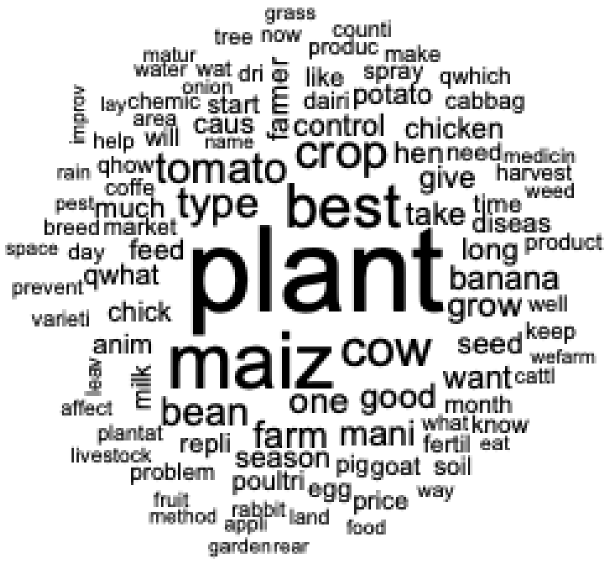


Word cloud for farmers' questions

- The word cloud presents word frequency in a different format.

- The larger the word in the cloud, the more frequently it occurs in the farmers' posts.

- The most frequent terms focused on planting, cattle, and various crops (e.g., maize, tomatoes, beans, and crops).



Most highly correlated words for most commonly used words

- The correlation coefficients for several of the most common post terms with other terms in the posts appear in this table.
- Positive correlations indicate that the two words are more likely to appear together in the posts.
  - E.g., Farmers who use the word “Cow” are more likely to also use the word “Milk”.
- Larger values of the correlation suggest a stronger relationship in the usage of the two words.
- For some target words (e.g., ”Plant”) there were fewer than 5 other words with which it was related.

Target Word	Word 1	Word 2	Word 3	Word 4	Word 5
Maize	Plant (0.16)	Seed (0.10	Worm (0.07)	Smut (0.07)	
Plant	Maize (0.16)	Season (0.15)			
Best	Variety (0.14)	Way ().12)	Which (0.12)		
Cow	Milk (0.29)	Dairy (0.18)	Heat (0.16)	Birth (0.13)	Pregnant (0.12)
Crop	Rotate (0.25)	Cash (0.19	Grown (0.12)		
Tomato	Blight (0.10)				
Bean	Soya (0.23)	French (0.17)			
Farm	Start (0.14)	Poultry (0.12)			
Poultry	Start (0.15)	Keep (0.12)	Farm (0.12)	House (0.10)	
Control	Pest (0.18)	Weed (0.14)	Disease (0.11)	Wilt (0.11)	
Soil	Erosion (0.30)	Clay (0.18)	Sandy (0.15)	Acid (0.14)	Loam (0.11)
Price	Current (0.21)	Kampala (0.11)	Per (0.11)		

Bigrams

- Bigrams reflect the frequency with which two words appear together in a post.
- The most frequent word pairs were “plant maize” and “maize plant”.
- Indeed, the 5 most common bigrams involved the term “plant”.
- The pair of ”dairy cow” was the most frequent bigram that did not involve planting and crops.

```
Bigrams
            ngrams freq
1      plant maiz  1050
2      maiz plant   602
3      want plant   457
4    plant tomato   430
5      crop plant   392
6       dairi cow   384
7      plant bean   369
8  banana plantat   368
9       maiz seed   347
10   sweet potato   319
11     good plant   311
12      maiz bean   301
13      type maiz   283
14        2 plant   280
15       best way   266
16   banana plant   262
17        lay egg   258
18   plant banana   258
19  passion fruit   250
20     many month   241
```

Trigrams

- The most frequent word trios (trigrams) were “plant maize bean” and “maize seed plant”.

- The most common non-crop trigrams were “start poultry farm” and “hen lay egg”.

- The majority of trigrams focused on techniques for planting crops, particularly maize.

```
 ngrams freq
1     plant maiz bean   131
2     maiz seed plant   107
3  start poultry farm    94
4     type maiz plant    79
5     best crop plant    70
6         hen lay egg    67
7      best maiz seed    66
8      type maiz seed    63
9   plant water melon    62
10    best time plant    62
11  best season plant    59
12 plant sweet potato    59
13    want plant maiz    57
14 plant irish potato    53
15     feed dairi cow    49
16    take give birth    47
17      day old chick    45
18     best way plant    45
19   best way control    45
20     top dress maiz    43
```

4-grams

- The most frequent word 4-grams were “best maize seed plant” and “one day old chick”.
- Other common 4-grams focused on the best seeds to plant, resistance to rain, and starting a poultry farm.

```
ngrams freq
1     best maiz seed plant    40
2        one day old chick    22
3     type maiz seed plant    21
4  hybrid bean resist rain    20
5  bean resist rain season    20
6      plant maiz bean one    18
7  want start poultri farm    16
8  need start poultri farm    15
9  best maiz varieti plant    15
10     feed best dairi cow    15
```

Topic modeling

- In addition to examining word frequency and co-occurrence, topic modeling was used in order to understand the underlying structure of the farmers' posts.

- Topic modeling groups the posts together based on word content.

- Posts that use similar word combinations belong to the same topic.

- Topics can then be characterized by the words that they have in common.

Topic modeling

- Several statistical methods were used to determine the number of topics likely to be present in the data.
  - Exclusivity – Common words in one topic are unlikely to appear in another; larger values indicate a better solution.
  - Coherence – Maximized when the most probable words in a topic frequently co-occur together; larger values indicate a better solution.
  - Lift - Divides word frequency in the target topic by their frequency in other topics; larger values indicate a better solution.
  - Score - Divides the log word frequency in the target topic by the log word frequency in other topics ; larger values indicate a better solution.
  - Residual dispersion - Check for large variance of the residuals for a topic solution; correctly specified models have a variance near 1.

Topic modeling

- Based on the statistical indices and word content, a 7-topic solution was found to be optimal.

- Topic 1: Business questions
- Topic 2: Tomatoes and non-cattle livestock
- Topic 3: Maize
- Topic 4: Pests and pest control
- Topic 5: Dairy cattle
- Topic 6: Coffee
- Topic 7: Poultry

Most frequent word for each topic

Topic 1: feed, product, join, loan, care

Topic 2: tomato, sheep, dog, goat, donkey

Topic 3: Maize, plant, spray, line, start

Topic 4: Worm, bacteria, control, pest, type

Topic 5: Cow, milk, calf, birth, month

Topic 6: Coffee, season, rain, fertilizer, need

Topic 7: egg, chicken, hen, shell, soft

Correlations among topics

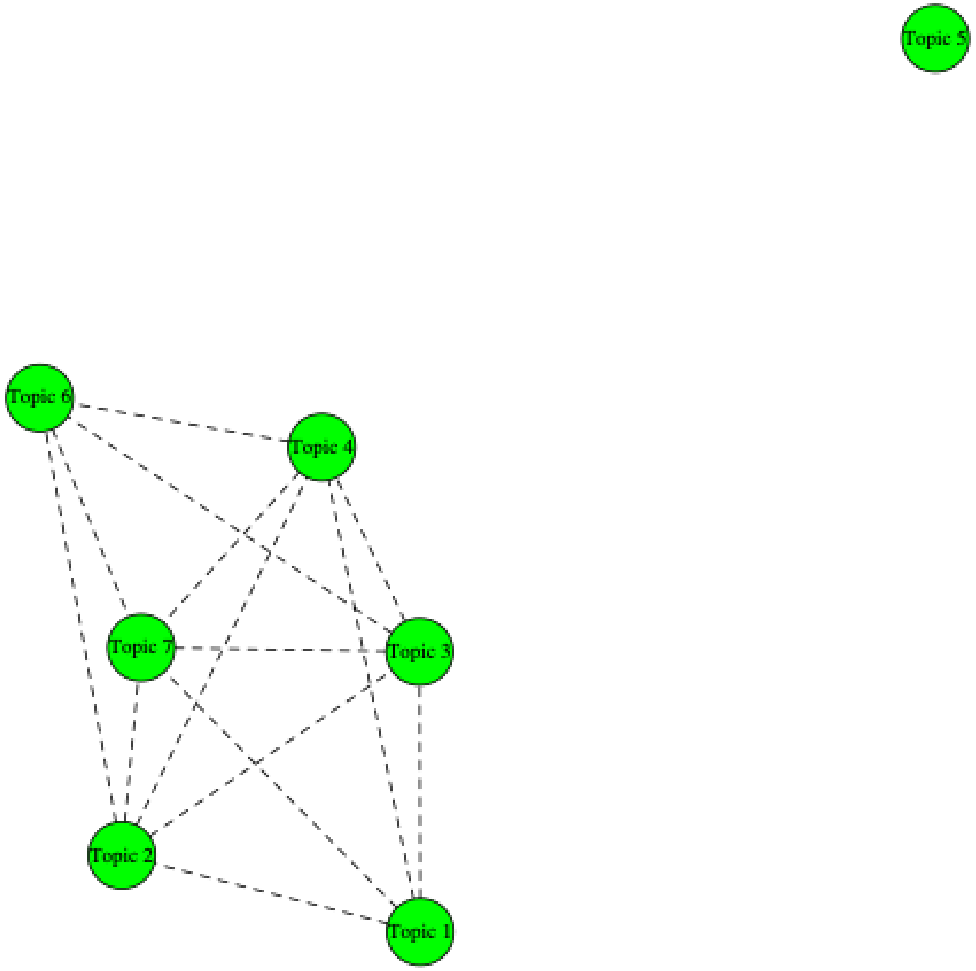
- It is useful to examine the correlations among topics.

- These correlations reflect the degree to which two topics are discussed in the same posts.
    - A large positive correlation indicates that two topics are likely to be mentioned in the same post.
    - A large negative correlation indicates that two topics are likely to be mentioned in different posts.

- The correlations are calculated using the estimated probabilities of a post belonging to a specific topic.

Correlations among topics

```
      [,1] [,2]  [,3]  [,4]  [,5]  [,6]   [,7]
[1,]  1.00  0.23  0.28  0.29 -0.42 -0.15  0.00
[2,]  0.23  1.00  0.17  0.28 -0.47  0.14  0.00
[3,]  0.28  0.17  1.00  0.38 -0.75  0.22  0.18
[4,]  0.29  0.28  0.38  1.00 -0.73  0.00  0.27
[5,] -0.42 -0.47 -0.75 -0.74  1.00 -0.39 -0.49
[6,] -0.15  0.14  0.22  0.00 -0.39  1.00  0.00
[7,]  0.00  0.00  0.18  0.27 -0.49  0.00  1.00
```



Correlations among topics

- Topics 1 (Business questions), 2 (Tomatoes and non-cattle livestock), 3 (Maize), and 4 (Pests and pest control) are relatively likely to be discussed together in posts.

- Topic 6 (Coffee) is relatively likely to be discussed with topics 2 (Tomatoes and non-cattle livestock) and 3 (Maize).

- Topic 7 (Poultry) is likely to be discussed with topics 3 (Maize) and 4 (Pests and pest control).

- Topic 5 (Dairy cattle) is unlikely to be discussed in the same post with any other topic.

Summary

- Farmers were most likely to post about maize, cattle, and best practices for planting in general.

- When posting about maize, they were most likely to ask about the best seeds to use and/or pest control.

- Posts about cattle were most likely to ask about issues around dairy.

- Questions about soil focused on problems with erosion, acidity, clay, and sand.

- Farmers were also likely to ask about starting a poultry farm.

Summary

- Statistical modeling identified 7 distinct topics among the farmers' posts.
- Several topics were relatively likely to be mentioned together in the same post.
- This was particularly for the maize topic, which was related to most of the others.
- Farmers who posted about dairy farming rarely asked about anything else.