Challenge 4: Crop-Specific and Crop-Independent Questions

Word Frequency Analysis: Visualizing Bigrams, Trigrams and QuadGrams of Questions by Kenyan Farmers in the Producers Direct Dataset of WeFarm SMS

Producers Direct Dataset

Original CSV file of 20,304,843 rows & 24 columns included 5,865,819 unique questions asked by 1,026,367 farmers

Time to answer questions: mean=4.8 days, s.d. = 4.3 days

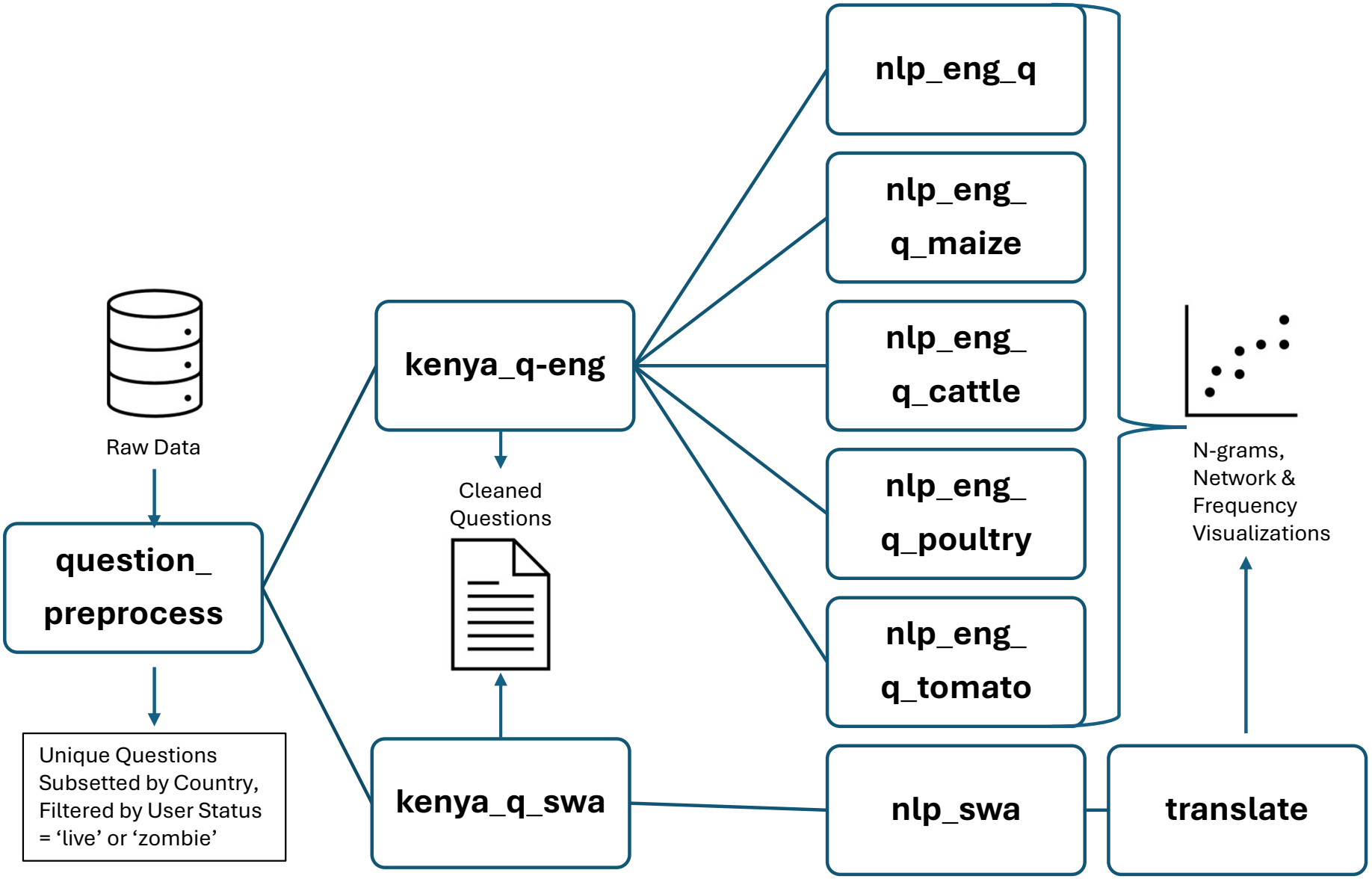
Unique Question Counts by:

Language		Country		Topic (148)		User Status	
eng	50.1%	Kenya	41.3%	null	28.5%	live	65.3%
swa	37.9%	Uganda	33.7%	poultry*	12.0%	zombie	19.1%
nyn	7.4%	Tanzania	25.0%	maize	9.1%	blocked	8.2%
lug	4.5%	Gambia	–-	cattle	5.5%	destroyed	7.4%
				tomato	5.6%		
				cranberry	–-		

* Includes chicken

Order of Jupyter Notebooks

Notebooks contain more detail about the steps, inputs, outputs, and dependencies



Challenges of Translating Swahili into English

Google Translate has a 5,000 character limit per call, so it's impractical to translate the full text of > 2 mm questions

A possible workaround is to extract and translate the most frequent combination of words in the Swahili questions to derive meaningful insights....

But Swahili is an under-resourced language in Natural Language Processing:

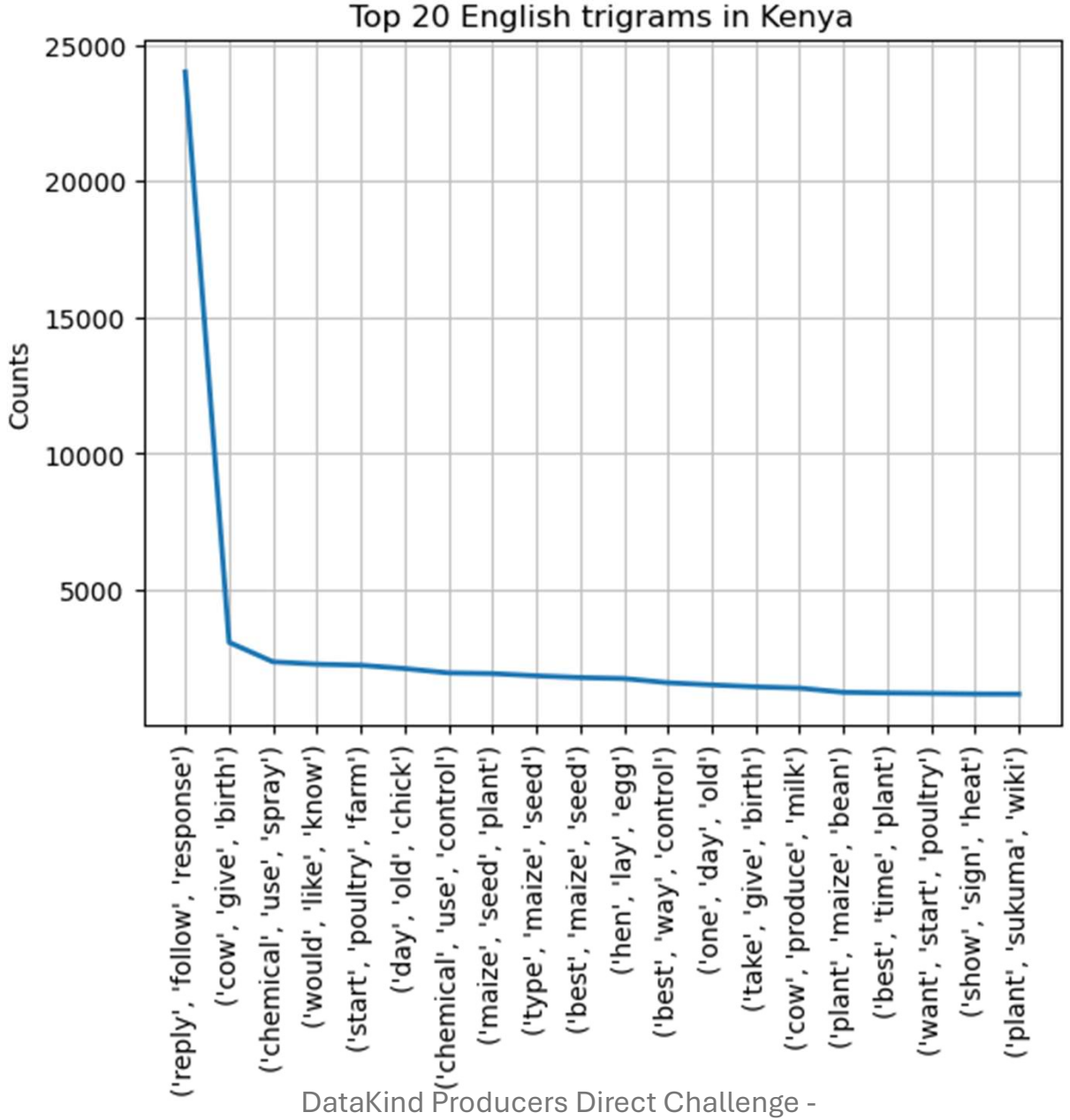
- Commonly used Python packages such as SpaCy, NLTK, or Gensim do not have inherent Swahili support
- It is an agglutinative language: prefixes, roots, and suffixes are combined into one word. It also has complex noun class structures, that affect verb agreement. These can lead to ineffective lemmatization.

An attempt was made to generate and translate n-grams using Swahili word lists sourced from the Mendeley Data repository, but the translated n-grams was not particularly informative. A more robust analysis requires an agricultural corpus on rural farming in Africa, and custom lists of words and lemma dictionary.

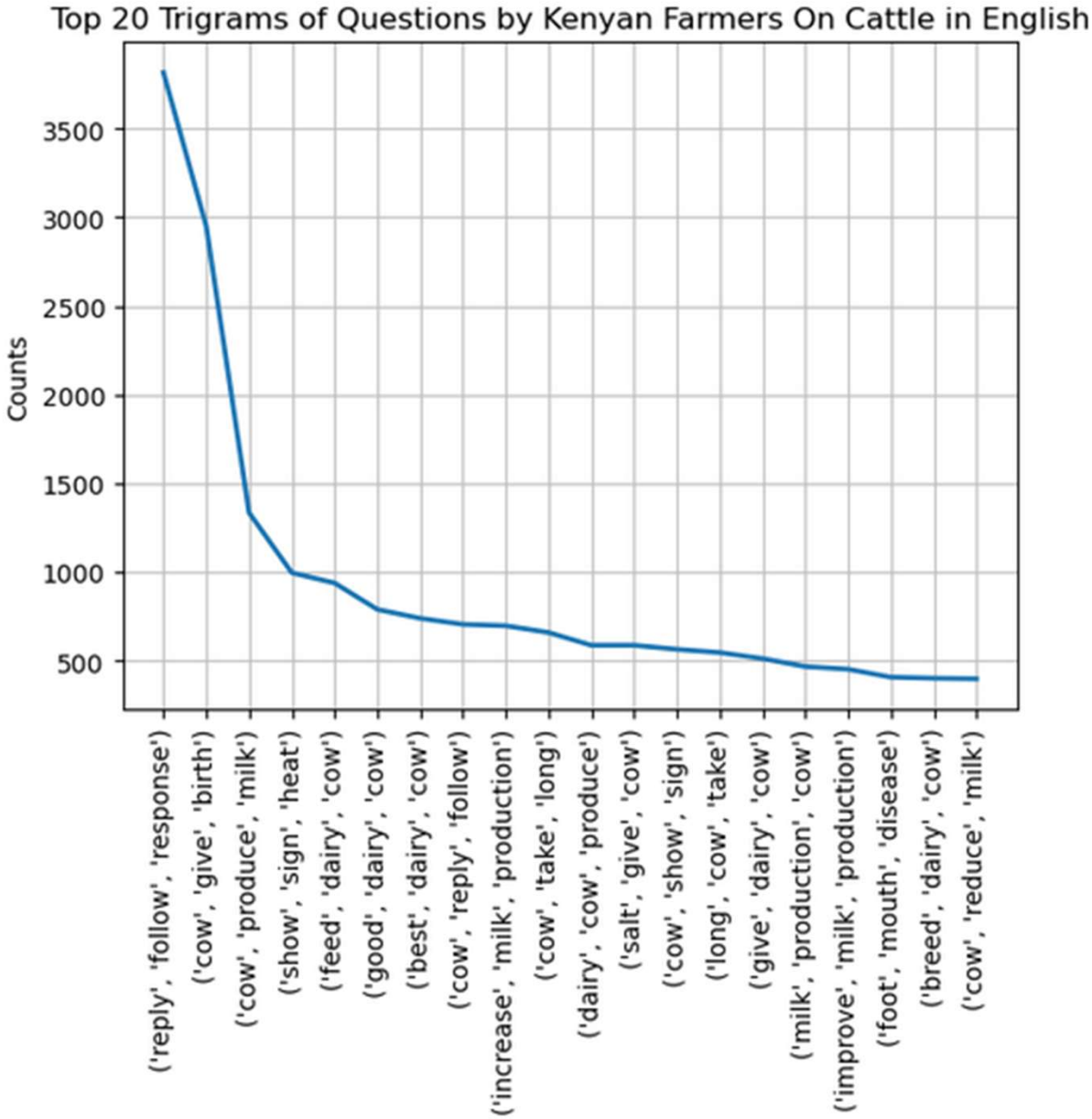
With These Challenges, Swahili Trigrams Translated in English – All Topics – Were More Informative Than Bi- or Quadgrams.

1. what_medicine_to_use
2. what medicine should i use
3. what_good_medicine
4. what_seed_india
5. can_get_me
6. what_good_medicine
7. what_drug_to_use
8. naeza_pata_mpi (I can get water)
9. get_the_seed
10. where is the problem?
11. to_lay_the_egg
12. how_long_take
13. medicine_can_I
14. I want to_fuga_ku (I want to raise / domesticate)
15. medicine_can_you

Out of 5 Million Trigrams, Farmers Most Frequently Asked About Birthing Calves, Starting Poultry Farms, Using Treatments / Chemicals, and Maize Seeds



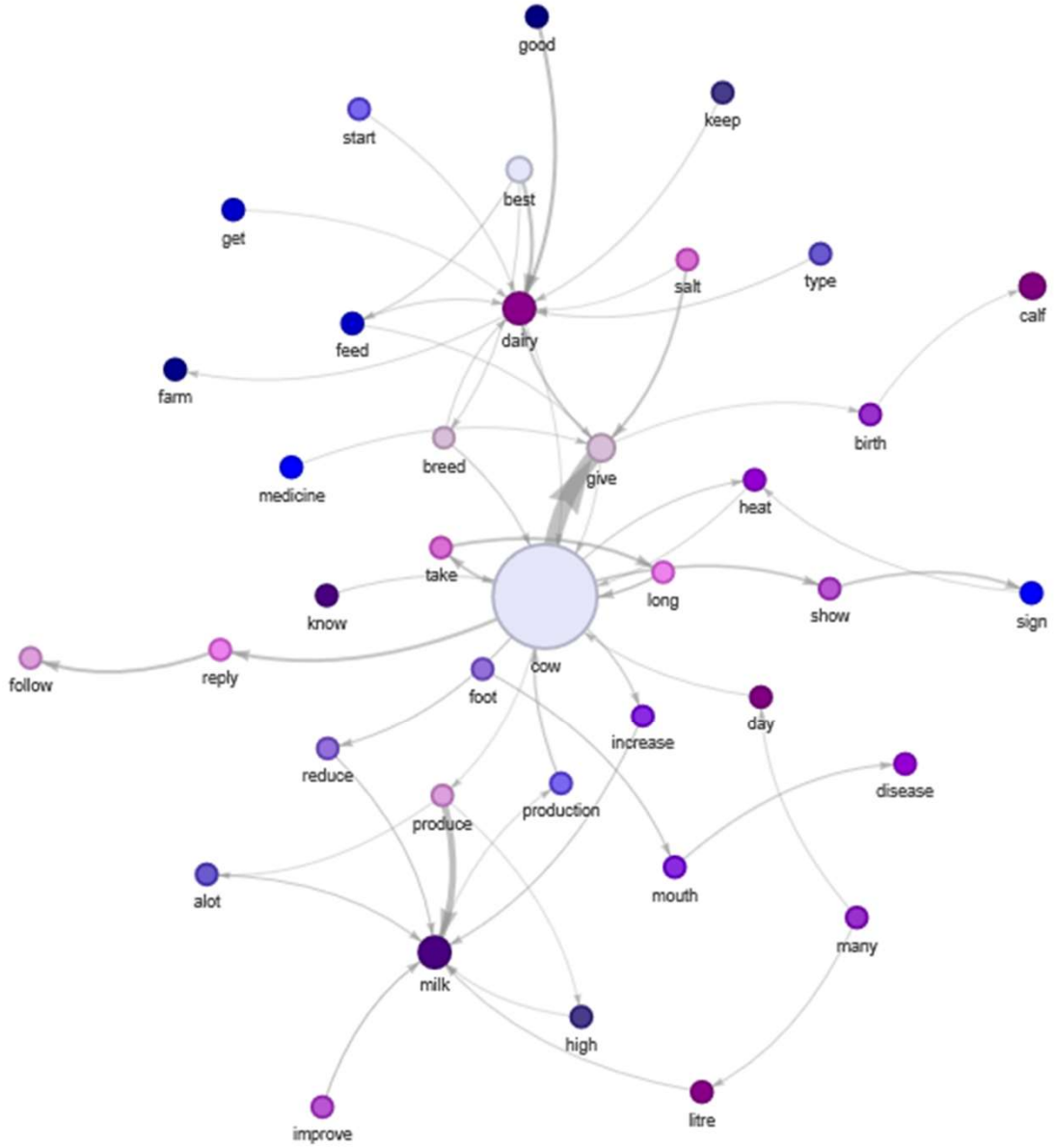
Cattle Kenyan Farmers Focused on Giving Birth, Milk Production, and Caring for Cows



DataKind Producers Direct Challenge -
December 2025, Beatrice Liu

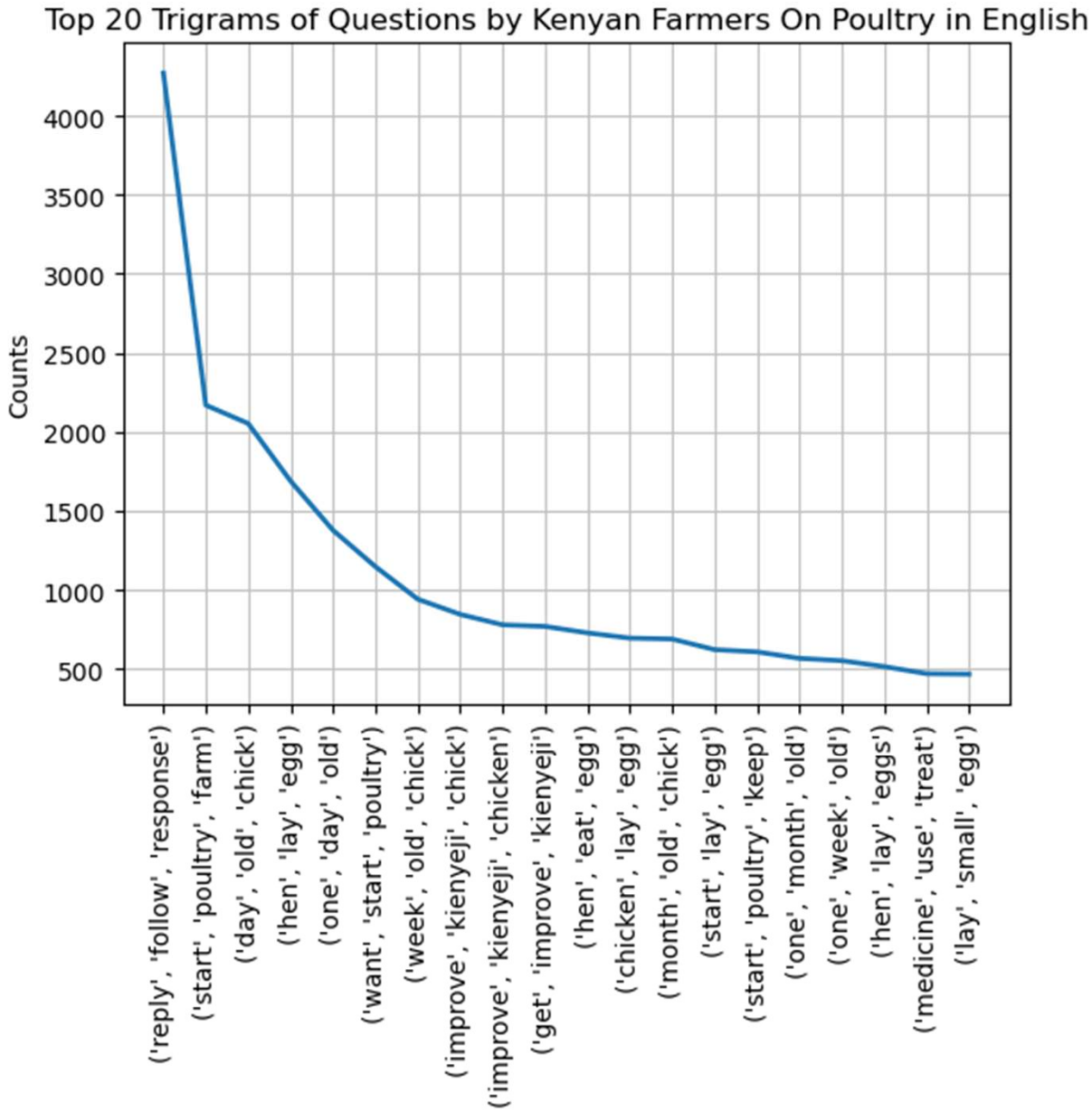
Network Graph: Top 40 Trigrams from Kenyan Farmers Questions on Cattle in English

Word circle size = word frequency, arrow width = trigram frequency, Source: WeFarm 2022 SMS Platform



DataKind Producers Direct Challenge -
December 2025, Beatrice Liu

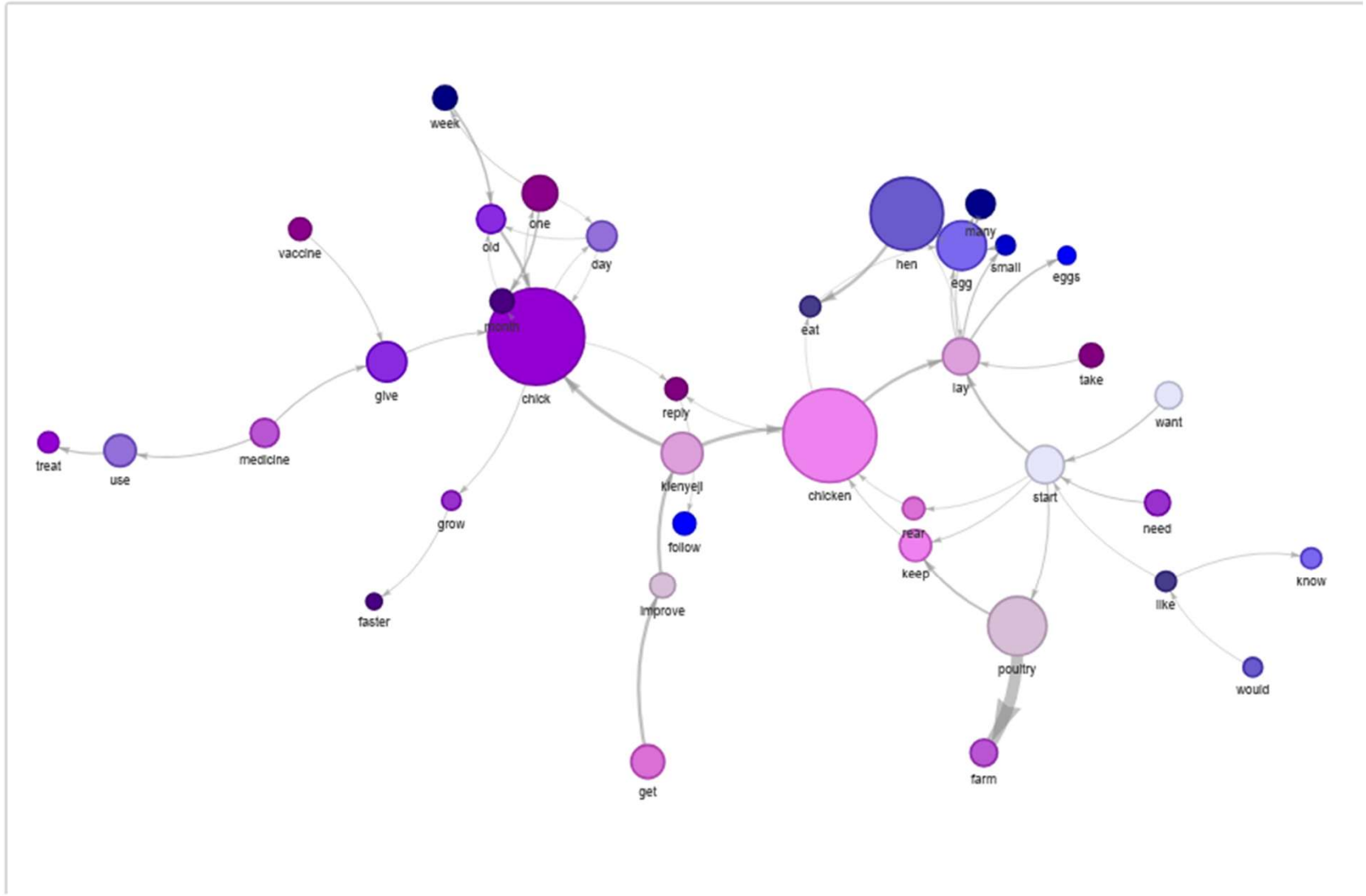
Kenyan Poultry Farmers Asked How to Start a Poultry Farms, Young Chicks, and Lay Eggs



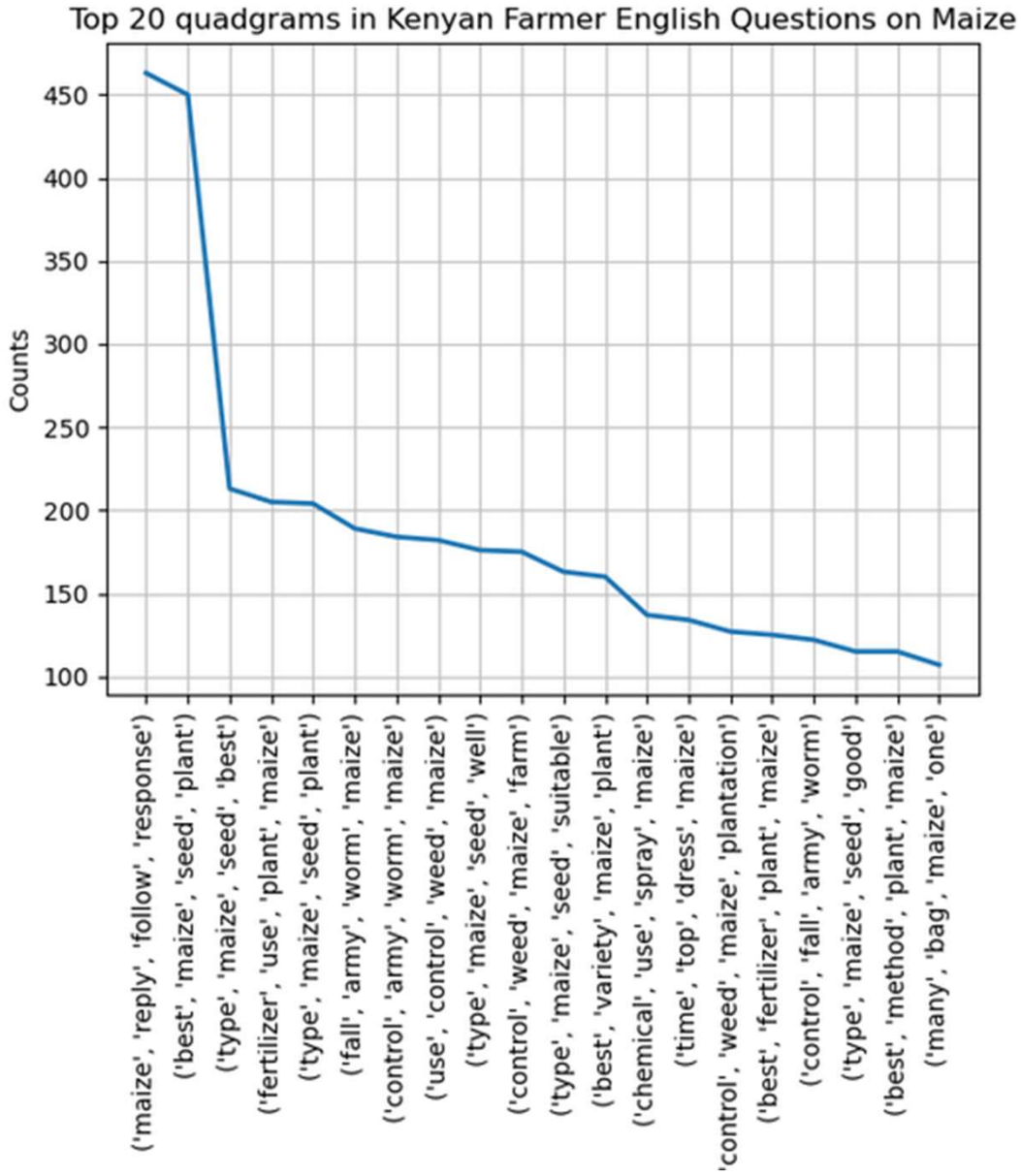
December 2025, Beatrice Liu

Network Graph: Top 40 Trigrams from Kenyan Farmers Questions on Poultry in English

Word circle size = word frequency, arrow width = trigram frequency, Source: WeFarm 2022 SMS Platform



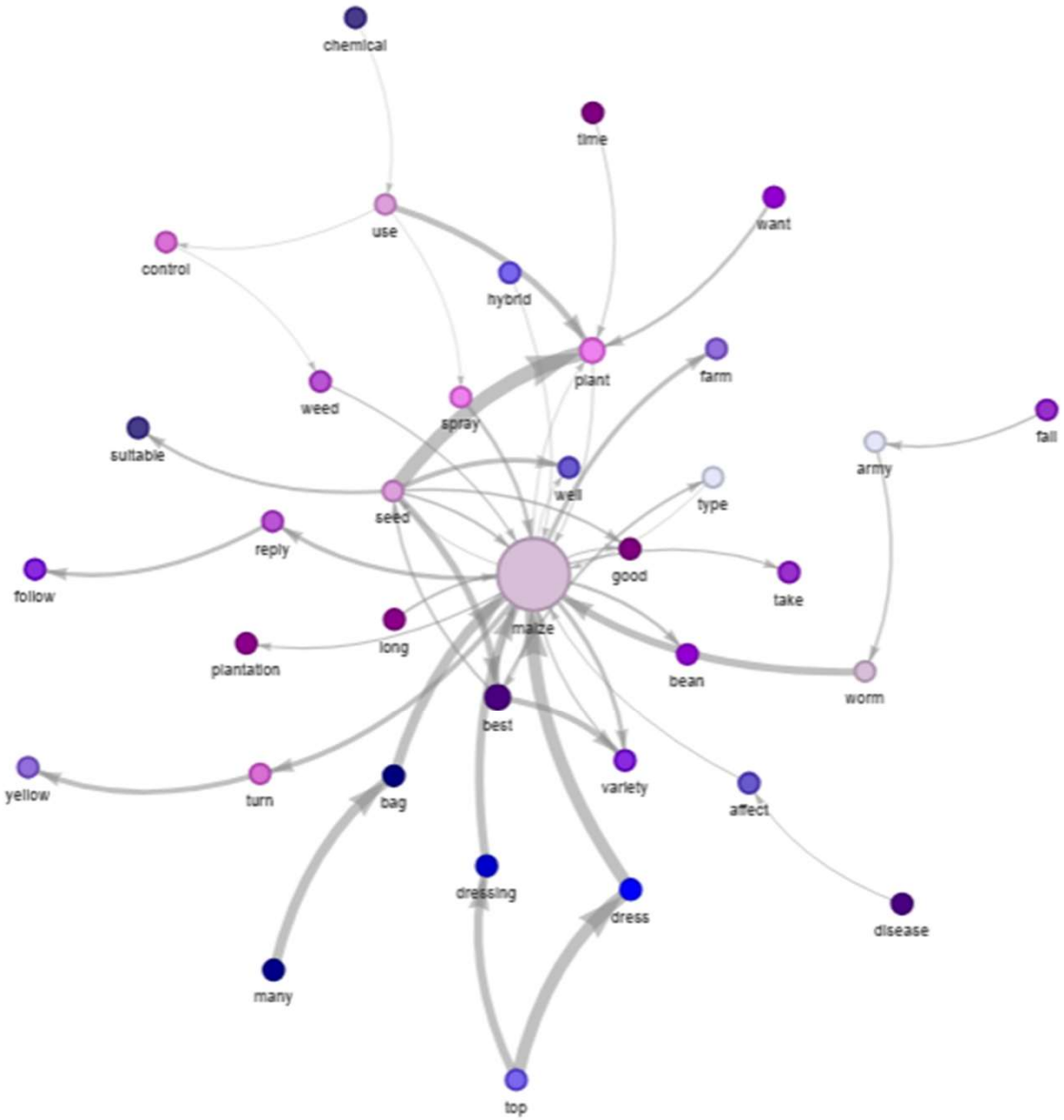
Maize Farmers in Kenya Asked About the Best Seeds, Best Fertilizer, Weed and Worm Control



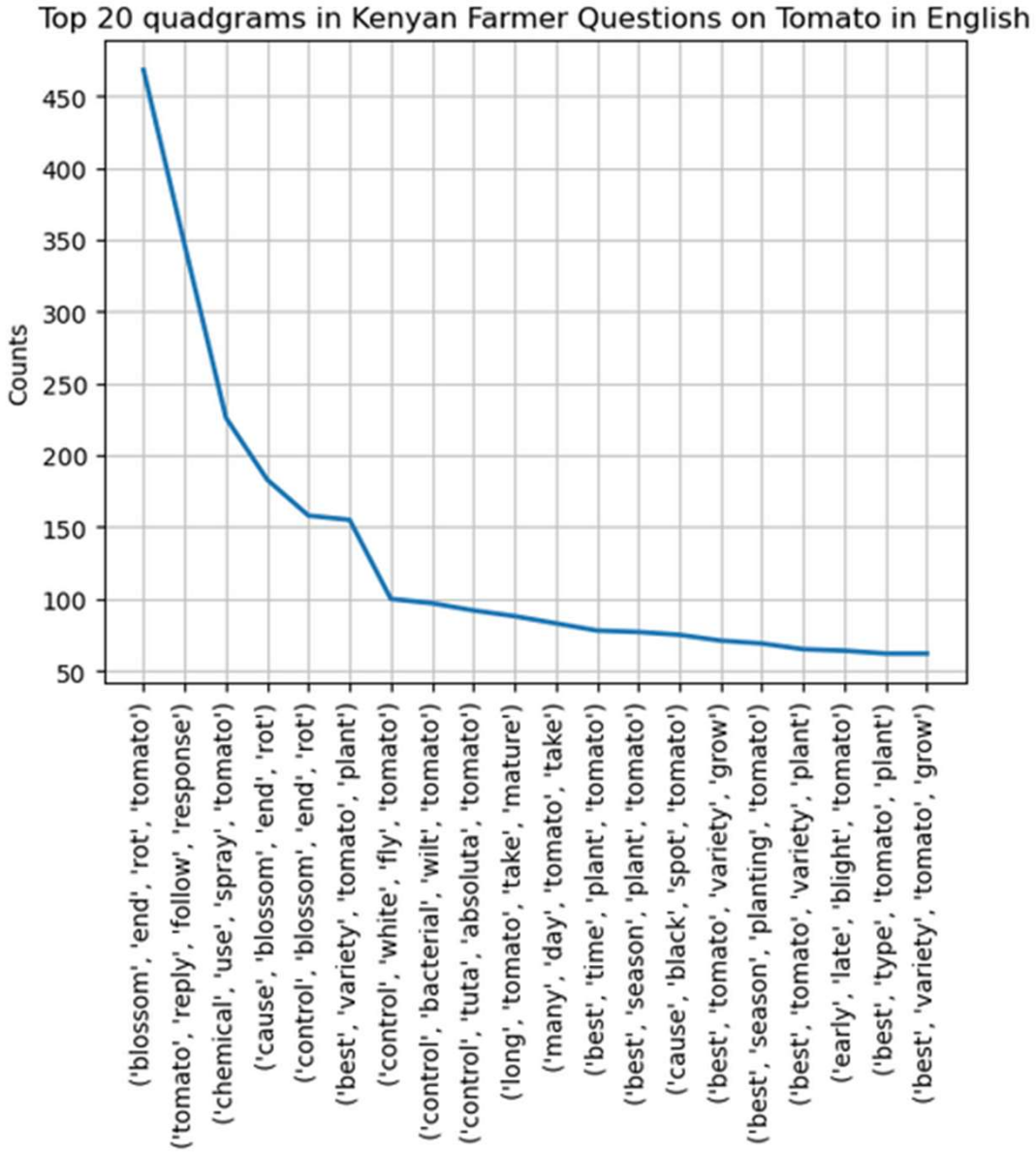
DataKind Producers Direct Challenge -
December 2025, Beatrice Liu

Network Graph: Top 40 Trigrams from Kenyan Farmer English Questions on Maize

Word circle size = word frequency, arrow width = trigram frequency, Source: WeFarm 2022 SMS Platform



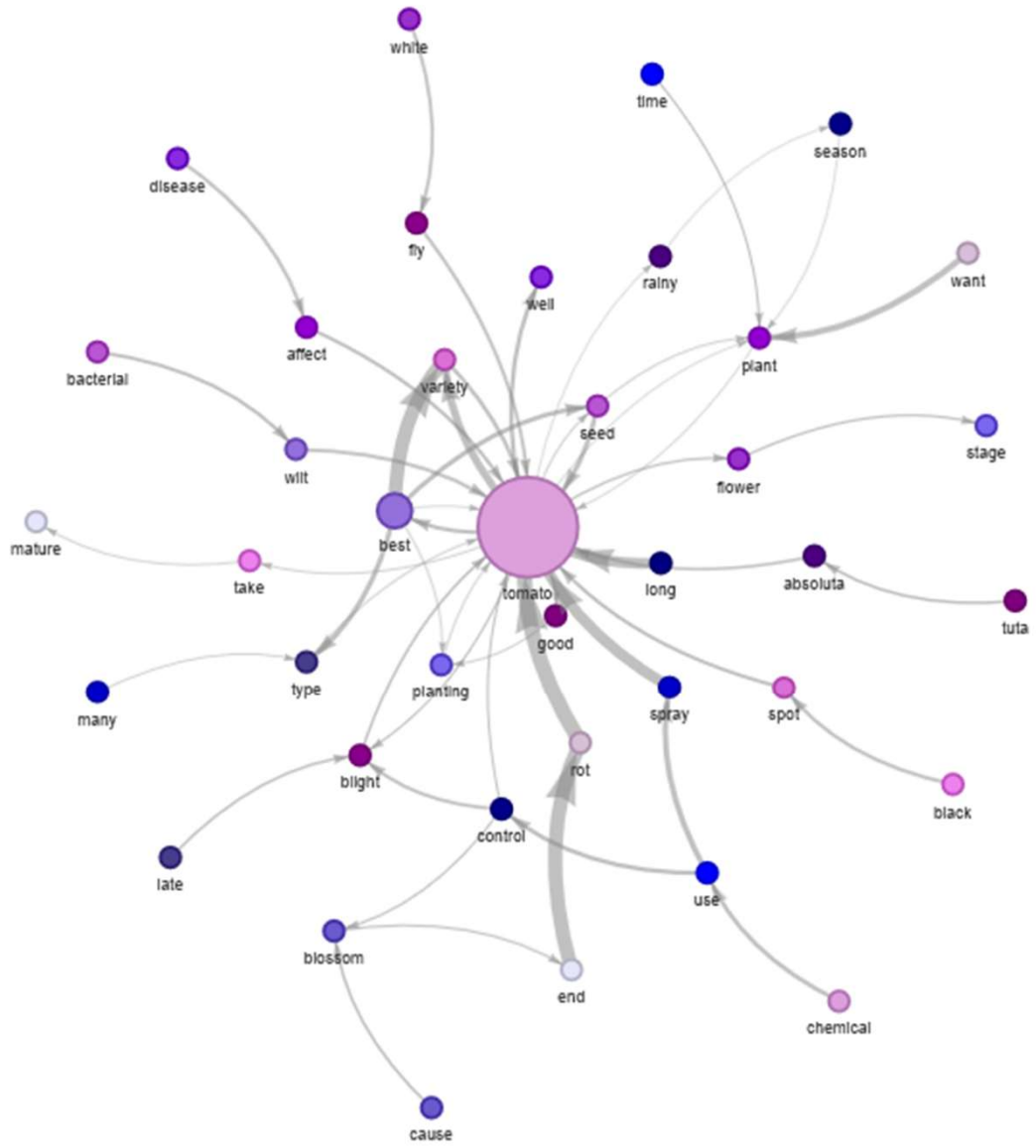
Tomato Kenyan Farmers Asked About Best Time to Plant, Blossom-End Rot, and Best Tomato Varieties



DataKind Producers Direct Challenge -
December 2025, Beatrice Liu

Network Graph: Top 40 Trigrams from Kenyan Farmers Questions on Tomato in English

Word circle size = word frequency, arrow width = trigram frequency, Source: WeFarm 2022 SMS Platform



DataKind Producers Direct Challenge -
December 2025, Beatrice Liu

GitHub Files: https://github.com/bl1412/datakit-smallholder-farmers-fall-2025.git

```
Bliu_analysis/
├── README.md (this file)
├── notebooks/
│   ├── question_preprocess.ipynb
│   └── kenya_q_swa.ipynb
│       └──nlp_swa.ipynb
│       └──translate.ipynb
│   ├── kenya_q_eng.ipynb
│       └──nlp_eng.ipynb
│       └──nlp_eng_cattle.ipynb
│       └──nlp_eng_poultry.ipynb
│       └──nlp_eng_maize.ipynb
│       └──nlp_eng_tomato.ipynb
├── 10 directed interactive network visualizations/
│   ├── {xx}bigram_eng_ken_{topic}_network.html
├── static visualizations/
│   └── top40trigrams_ken_eng_network.png
│   └── top40bigrams_ken_eng_network.png
│   └── top20quadgrams_ken_tomato_eng.png
│   └── top20trigrams_cattle_ken_eng.png
│   └── top20trigrams_poultry_ken_eng.png
│   └── top20quadgrams_ken_maize_eng.png
├── results - displayed in notebooks /
│   ├── farmers.bliu.pdf - sorry, no markdown file
└── translated n-grams from swahili to english data/
    └── ken_240quadgrams_swa2eng.txt
    └── ken_500bigrams_swa2eng.txt
    └── ken_500trigrams_swa2eng.txt
```

Additional Resources

Data files and visualizations created by these notebooks in Google Drive: https://drive.google.com/drive/folders/1tpwqTqoFfZCWvDvncJjaSbzzua0Y6Q_i?usp=sharing

Swahili Word Datasets:

- Common Swahili Stop-Words; https://data.mendeley.com/datasets/mmf4hnsm2n/1
- Swahili Agriculture Corpus: KILIMO: https://data.mendeley.com/datasets/d4yhn5b9n6/2/files/cfd0108d-863d-460d-b52c-a51ce4101f79
- Swahili Verb Conjugation Dataset for lemmatization: https://data.mendeley.com/datasets/rvt89578g5/1

References:

- Bernard Masua, Noel Masasi, "Enhancing text pre-processing for Swahili language: Datasets for common Swahili stop-words, slangs and typos with equivalent proper words", Data in Brief, Volume 33, 2020, 106517, ISSN 2352-3409, https://doi.org/10.1016/j.dib.2020.106517
- Mathayo, Irene; Kondoro, Alfred Malengo (2025), “Swahili Verb Conjugation Dataset: A Comprehensive Analysis of Agglutination and Verb Structure Across Tenses and Persons”, Mendeley Data, V3, doi: 10.17632/rvt89578g5.3
- Bernard Masua, Noel Masasi, "In the heart of Swahili: An exploration of data collection methods and corpus curation for natural language processing", Data in Brief, Volume 55,2024,110751,ISSN 2352-3409,https://doi.org/10.1016/j.dib.2024.110751