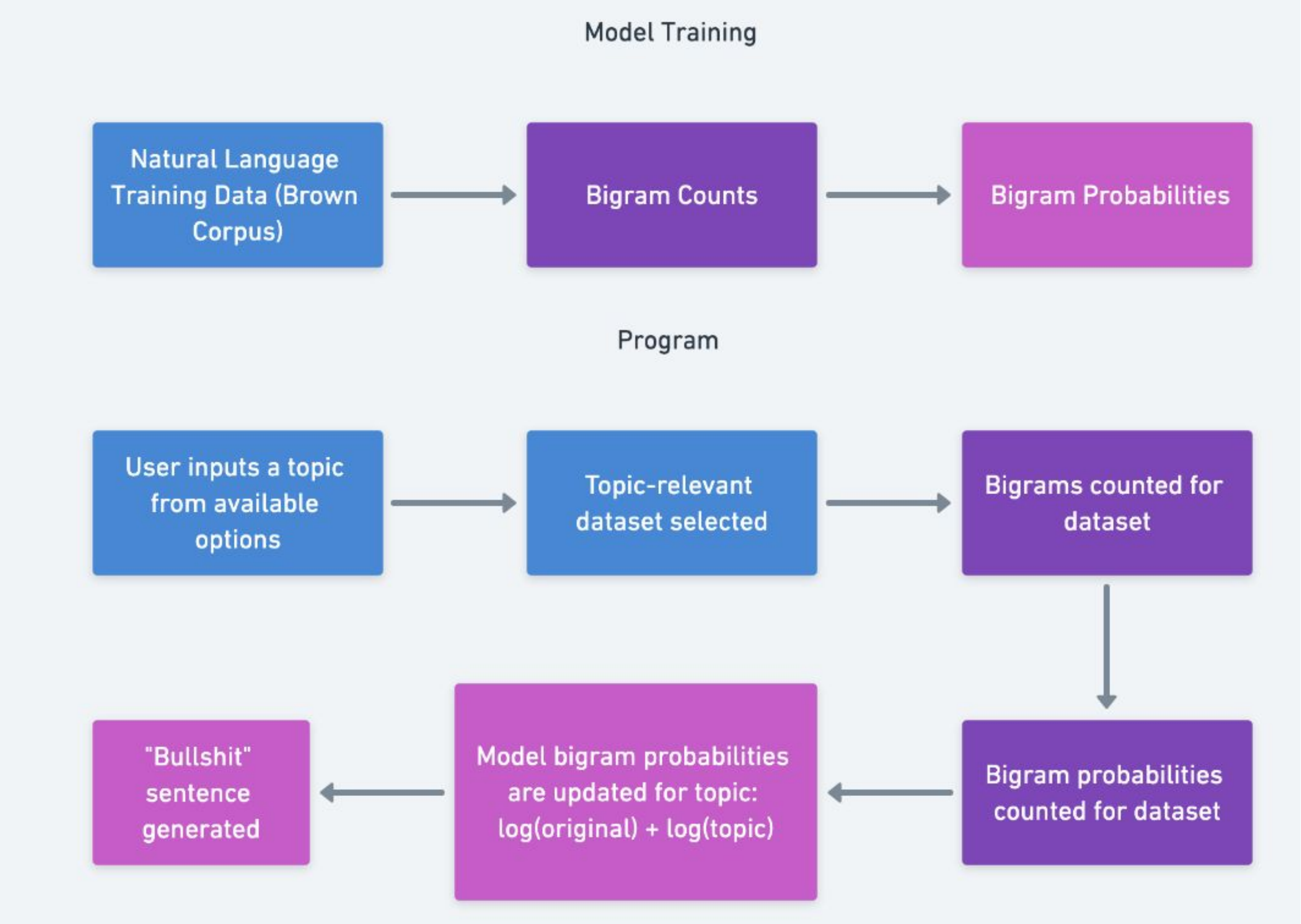
Generating “Bull”

Murage Kibicho and Uma Dwivedi

The Structure of Our Program



Model Training
Natural Language Training Data (Brown Corpus)
Bigram Counts
Bigram Probabilities
Program
User inputs a topic from available options
Topic-relevant dataset selected
Bigrams counted for dataset
Bigram probabilities counted for dataset
Model bigram probabilities are updated for topic: log(original) + log(topic)
"Bullshit" sentence generated

The Basic Language Model

- Bigram model trained on Brown's 1.1 million word corpus
- The corpus includes news articles, fiction, religious texts, nonfiction books, government documents, and many other sources.
- We selected it for the variety of sources it draws from: it's difficult to get natural language data in multiple registers, which was something we specifically wanted for our "bullshit" generator
- We used smoothing to take bigram counts and calculate log probabilities

Updating the Language Model with “Flavor”-Specific Data

Our program provides five “flavors” of bullshit for a user to select from:

- Jokes
- Sci-Fi
- Regency England
- Finance
- Medicine

“Flavor”-Specific Datasets

- Jokes: text from wocka.com
- Sci-Fi: *Star Trek* scripts
- Regency England: Jane Austen novels
- Finance: Headlines for news articles about stocks
- Medicine: Medical transcription data from mtsamples.com

Updating the Basic Language Model

- We counted the bigrams in the specified “flavor” dataset and calculated (smoothed) log probabilities
- We then added these to the log probabilities from our basic model to create a new set of probabilities
- The updated probabilities were used to generate a phrase according to the specified flavor

Outputs: The Good, the Bad, and the Fun

The Good

- “The sacrifice, he feared, would speak.” (Regency England)
- “The dekendi’s favourite is defenceless.” (Sci-Fi)
- “The bandages were touching the mole that turned away cephalically and preparations were consistent.” (Medicine)
- “The alvera trees, and utter oblivion” (Sci-Fi)
- “The dalai lama implied that without making sandwiches” (Jokes)
- “The country gives, and preach such universal improvement, though young persons was restlessly miserably forever!” (Regency England)

The Bad

- “The daughter, formed such true generosity and leisure hours, mrs.” (Regency England)
- “The camel, where you’re halfway across, and coolant” (Jokes)
- “The midline, non painful” (Medicine)
- “The LOS sales seeing major project stakes demand declining” (Finance)
- “The kendi system toenlist the mechanism to ziyal” (Sci-Fi)
- “The meeting you home work, get outside door” (Jokes)

The Fun

- “The grail again?” (Jokes)
- “The young-uns and stud back tail but don’t trust google!” (Jokes)
- “The inflation industrial product names no grace your pocket calculator” (Finance)
- “The suturing spaces four carbohydrate serving” (Medicine)
- “The world nor the interval would increase his nephew” (Regency England)

Moving Forward

- Incorporating part-of-speech data to increase subject/verb agreement
- Use probabilistically-weighted random choice to select the first word rather than starting every sentence with “the”
- Combining multiple datasets for each flavor
- Try preserving capitalization
- Using Wikipedia data for the base language model