# Learning Equality - Curriculum Recommendations

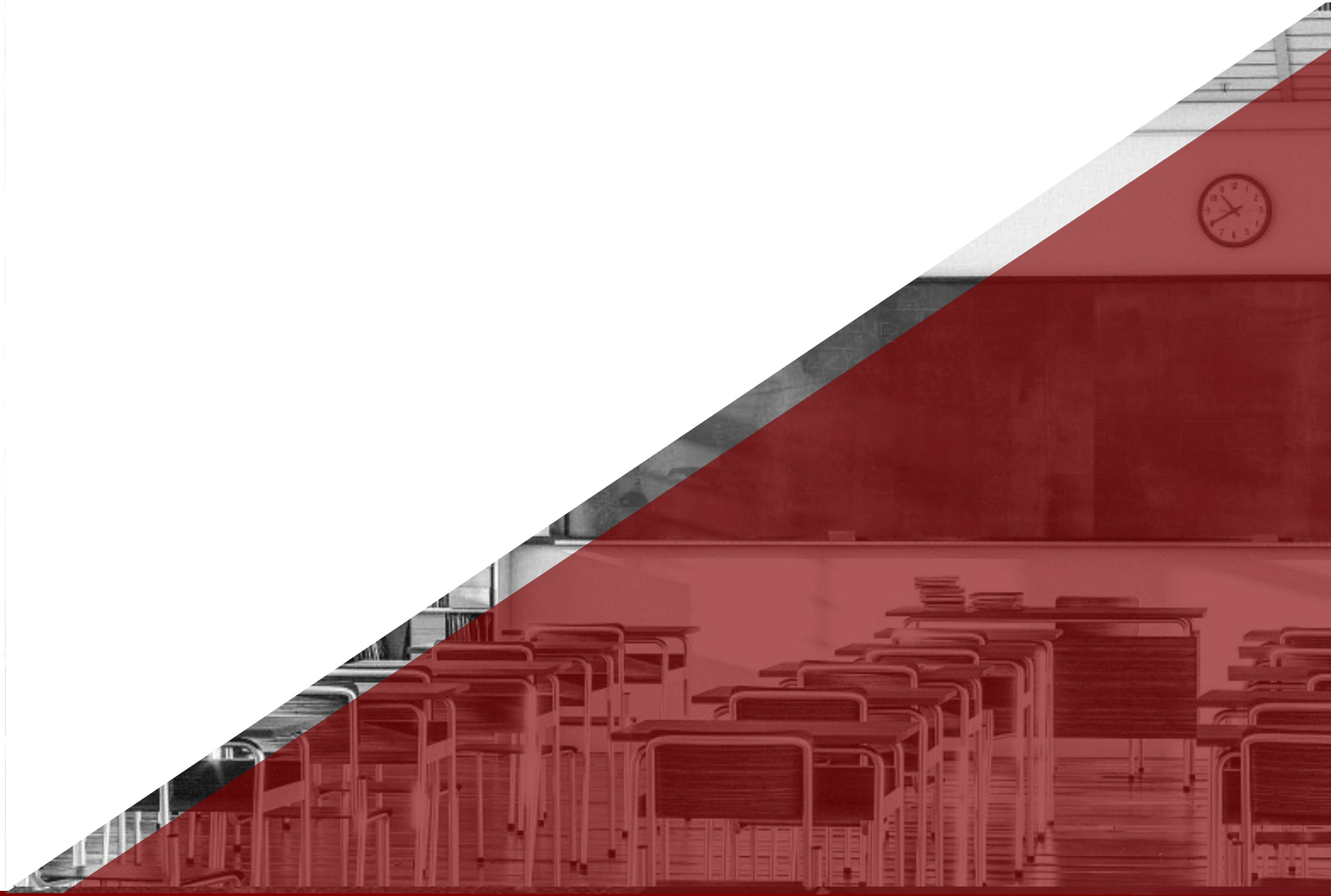
*Enhance learning by matching K-12 content to target topics*

Naman Mehta, Puneet Modi, Kishor Reddy, Kargil Thakur, and Ekansh Trivedi

THE UNIVERSITY OF CHICAGO

# Agenda

- Problem Statement
- Prototype Solution
- Data Description
- EDA
- Feature Engineering
- Model Selection
- Conclusion
- Future Scope



# Business Problem

Streamline the process of matching educational content to specific topics in a curriculum. We are developing an accurate and efficient model trained on a library of K-12 educational materials that have been organized into a variety of topic taxonomies. These materials are in diverse languages and cover a wide range of topics, particularly in STEM (Science, Technology, Engineering, and Mathematics).

## Content Search

Select a topic

Dissolution of Partner

### Description:

. this is the topic tree cbse >> cbse content - final >> cbse >> elements of book keeping and accountancy >> dissolution of partnership firm >> dissolution of partner

Search

# Learning Content Recommender

Search Results

Made with Streamlit

# Data Description

### Topic

| Column | Description |
|---|---|
| ID | Topic ID |
| Title | Title of the topic |
| Description | Description text |
| Channel | Topic Tree ID |
| Category | Origin of the topic |
| Language | Language of the topic |
| Parent ID | Parent ID of the topic |
| Level | The depth within its topic tree |
| Has Content | 1 if any content item is mapped |

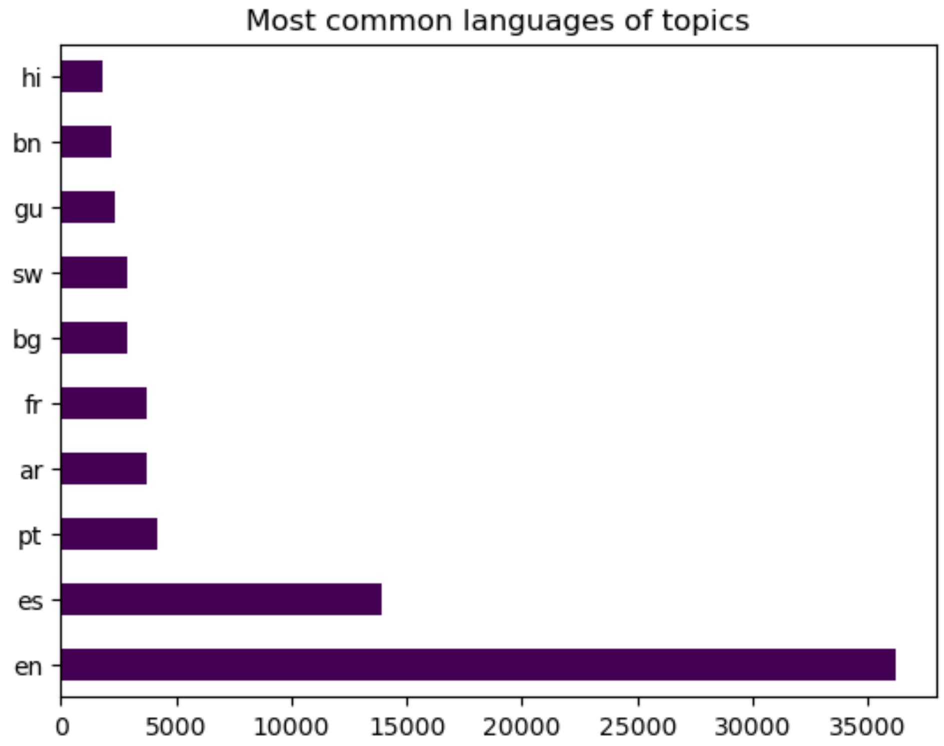
### Content

| Column | Description |
|---|---|
| ID | Content ID |
| Title | Title of the content |
| Description | Summary of the content |
| Text | Full details about the content |
| Language | Language of the content |
| Kind | HTML, Document, Video, etc. |

### Mapping

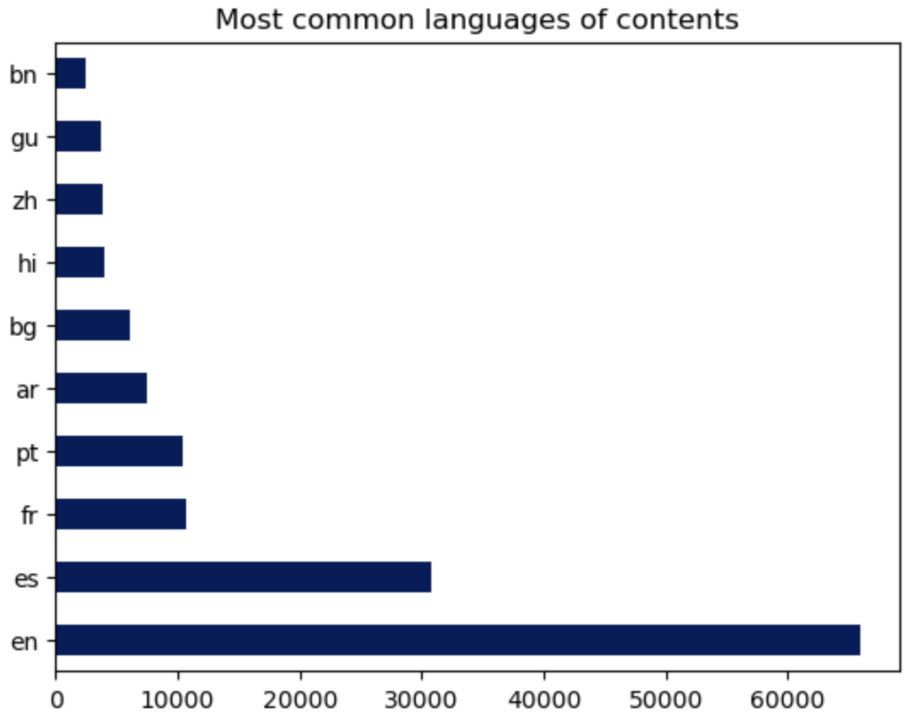
| Column | Description |
|---|---|
| Topic ID | Topic ID |
| Content IDs | Related content for the topic |

# Exploratory Data Analysis

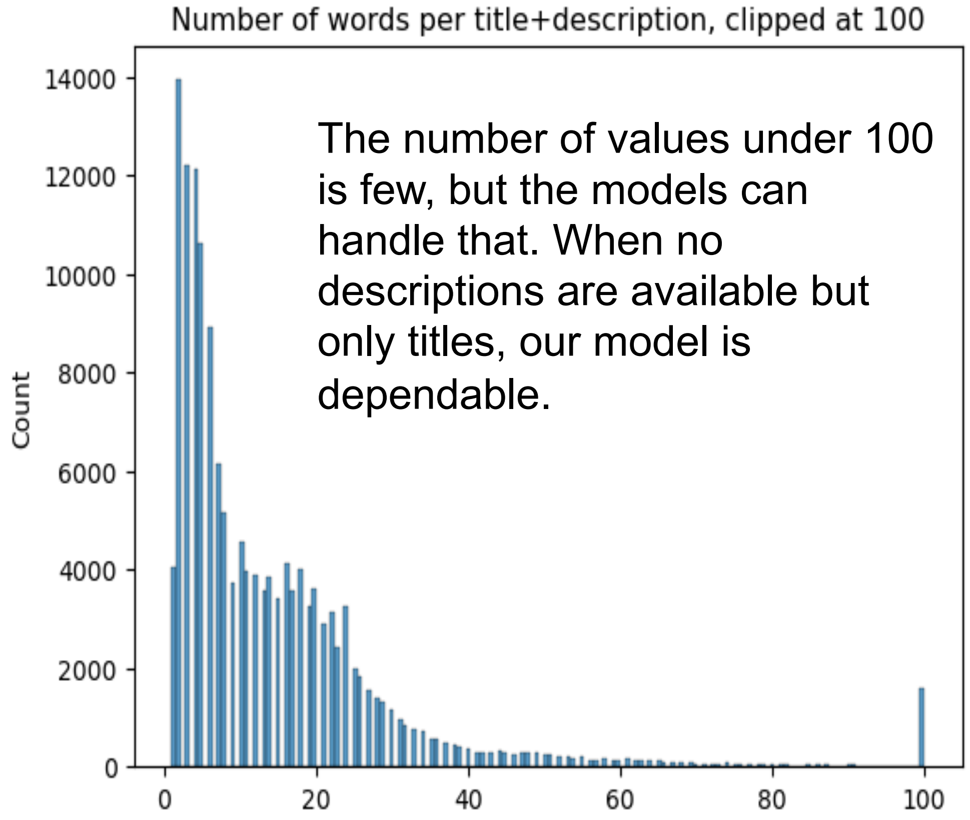



The most common language for topics and contents is English, followed by Spanish.

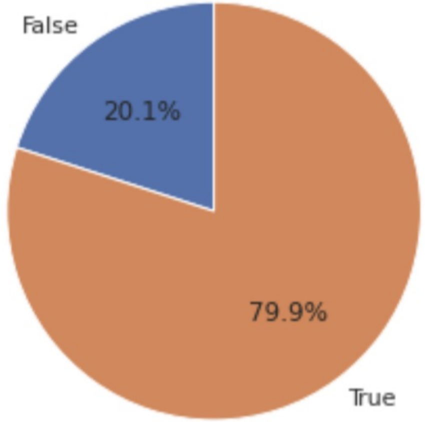
An important thing to note is that the topics and contents must be of the same language to recommend.



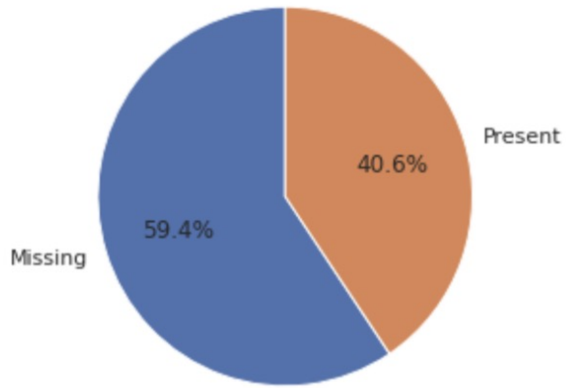
# Exploratory Data Analysis



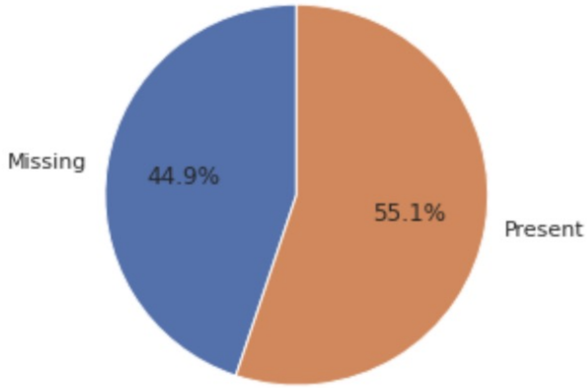

Name: has_content, dtype: int64



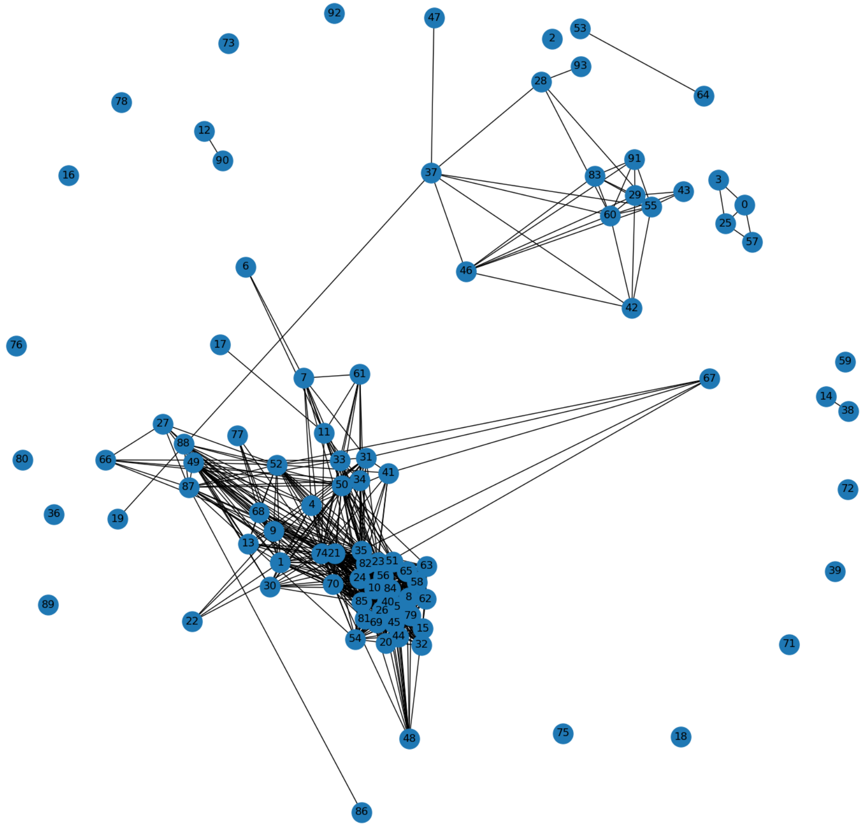

Num of content have description: 26762
Num of content don't have description: 39177




Num of content have text: 36347
Num of content don't have text: 29592

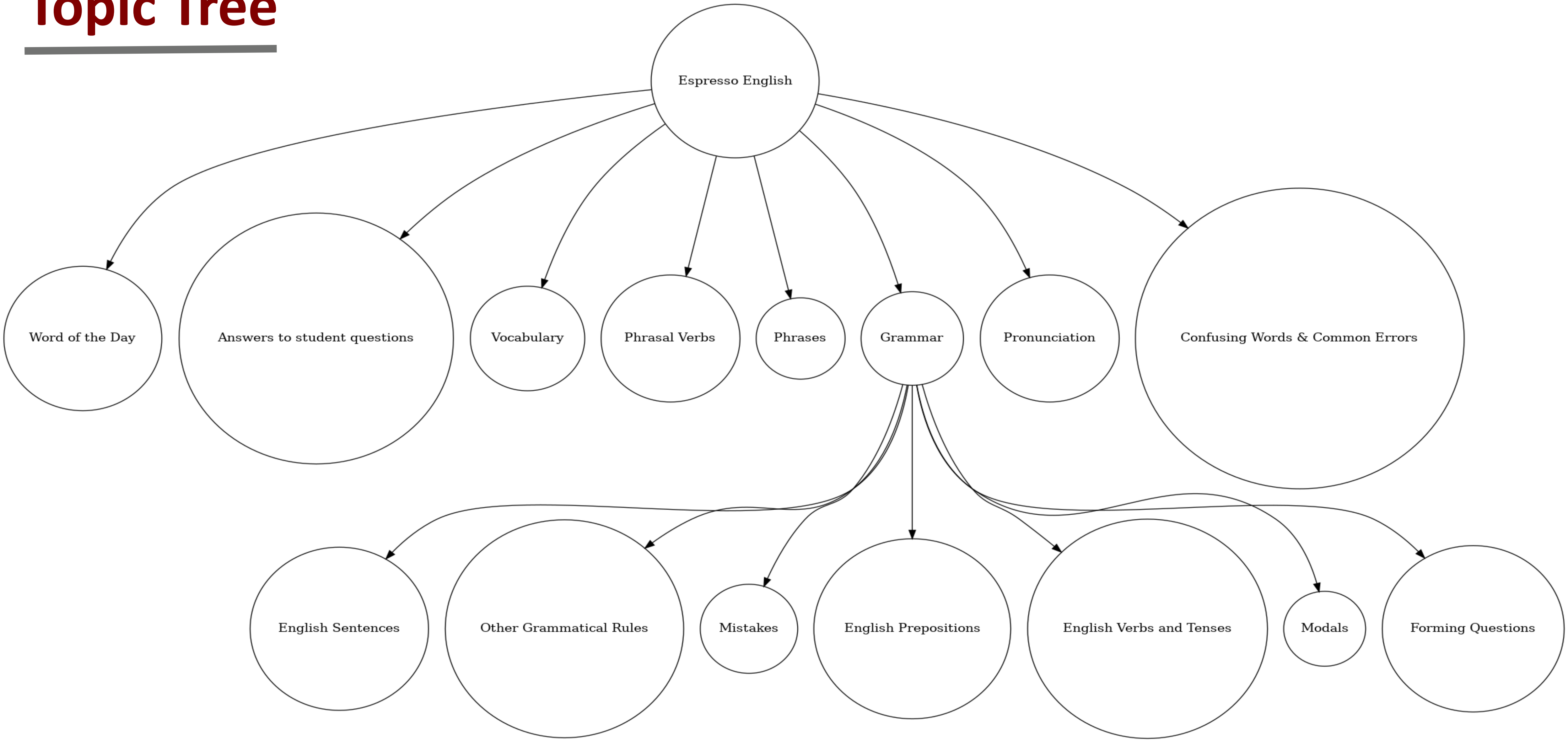


# Exploratory Data Analysis



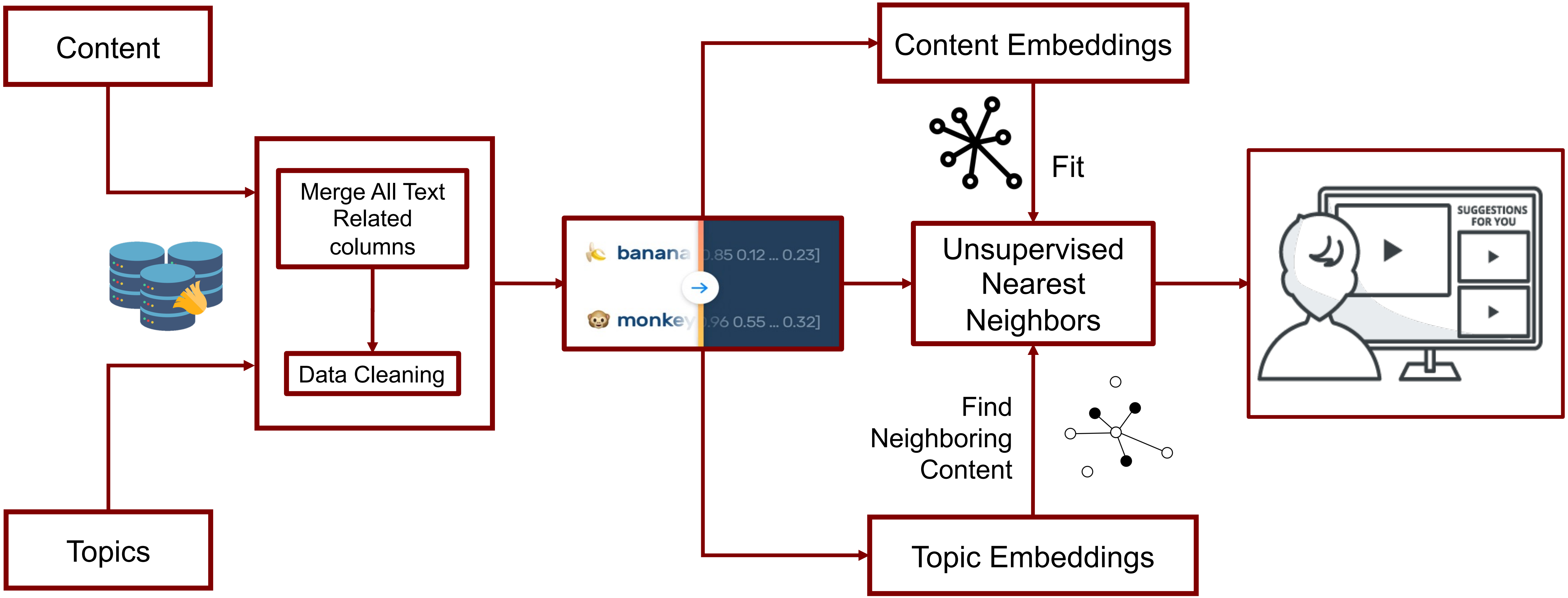
Network graph where the channels are the nodes, and the content ids are the edges

# Topic Tree




THE UNIVERSITY OF CHICAGO

# Solution Design



# Model Selection

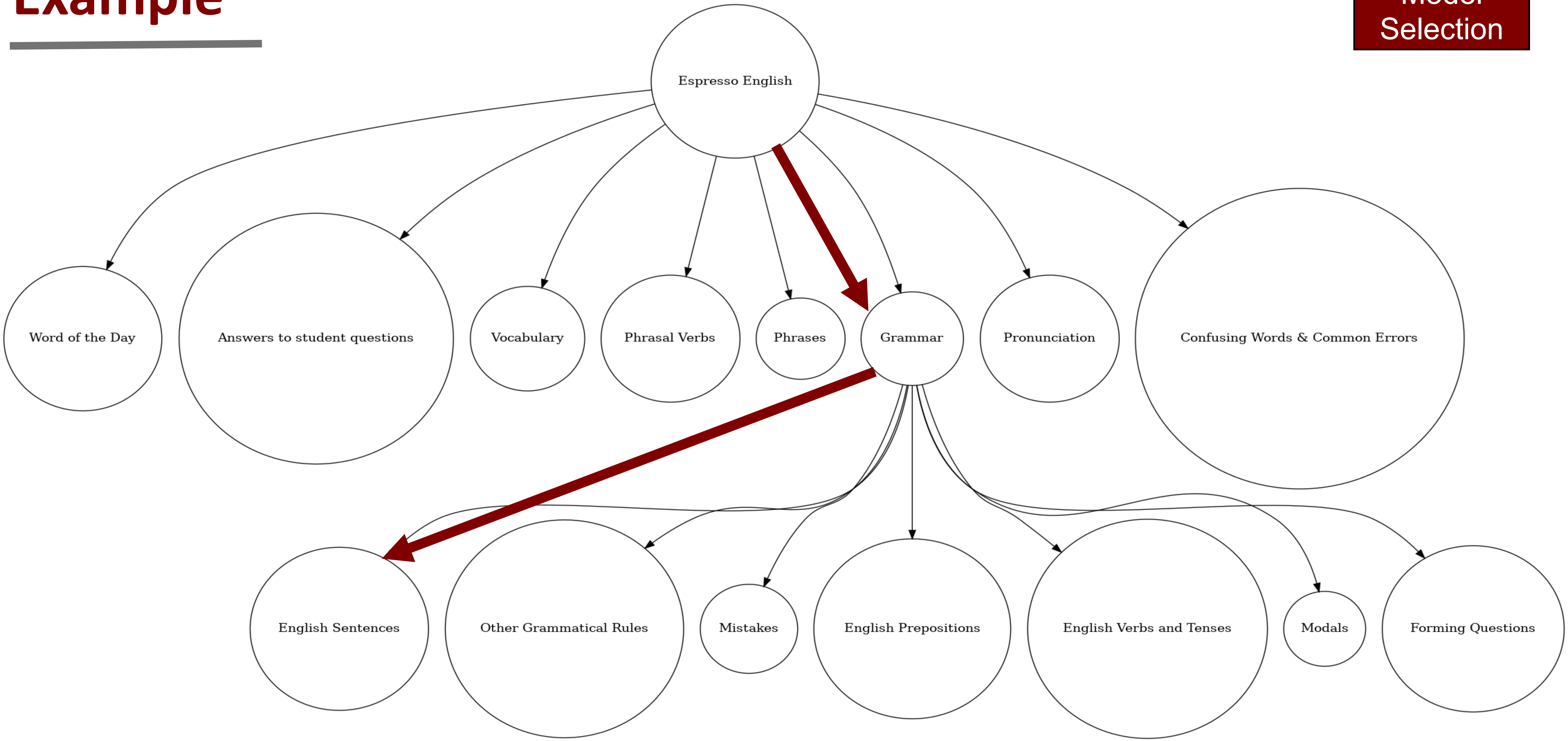
| Approach | Average F2 | Mean AP@ k |
|---|---|---|
| Count Vectorizer + Truncated SVD + Nearest Neighbors (Euclidean) | 0% | 0% |
| TF-IDF + Truncated SVD + Nearest Neighbors (Cosine) | 1% | 1.3% |
| Pre-Trained Model + Nearest Neighbors (Cosine) | 20.5% | 23% |

| Model Name | Performance Sentence Embeddings (14 Datasets) | Performance Semantic Search (6 Datasets) | Avg. Performance | Speed | Model Size |
|---|---|---|---|---|---|
| paraphrase-MiniLM-L3-v2 | 62.29 | 39.19 | 50.74 | 19000 | 61 MB |
| all-MiniLM-L6-v2 | 68.06 | 49.54 | 58.80 | 14200 | 80 MB |
| multi-qa-MiniLM-L6-cos-v1 | 64.33 | 51.83 | 58.08 | 14200 | 80 MB |
| paraphrase-multilingual-MiniLM-L12-v2 | 64.25 | 39.19 | 51.72 | 7500 | 420 MB |
| all-MiniLM-L12-v2 | 68.70 | 50.82 | 59.76 | 7500 | 120 MB |

Source: SBERT

# Example

Model Selection

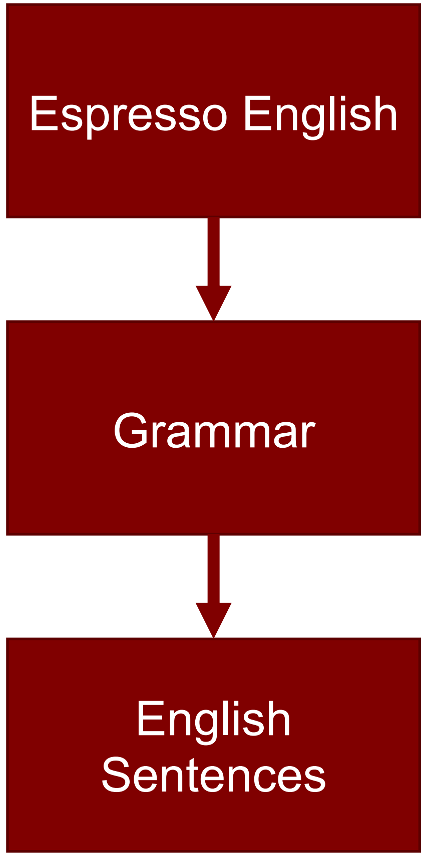


# Recommendation

| Current Content | Type of Content |
|---|---|
| **Sentence - Types and Structures** | Video |
| How to form English sentences with NOT ONLY + BUT ALSO | Video |
| **English Sentence Structure: 4 Types of English Sentences** | Video |

| Recommended Content (Ranked by Similarity) | Type of Content |
|---|---|
| **Sentence - Types and Structures** | Video |
| GRAMMAR Modals (Part One) | Video |
| Subject Verb Object Sentences | Video |
| **English Sentence Structure: 4 Types of English Sentences** | Video |
| GRAMMAR Modals (Part Two) | Video |
| GRAMMAR Modals (Part Three) | Video |
| Subject and Object Questions in English | Video |

## Topic Branch



# Evaluation

## Average Precision @ K

Relevant Content

Recommended Content

Content Ranking based on Similarity

Precision = 1 | 1/2 | 1/3 | 2/4 | 2/5 | 2/6 | 2/7

Average Precision @ K $= \frac{1 + 2/4}{2} = 75\%$

Mean Average Precision @ K $= \frac{\sum AP\ @\ K}{\#\ of\ Topics}$

## F2 Score

$$F2\ Score = \frac{TP}{TP + (0.2 * FP + 0.8 * FN)}$$

$$True\ Positives = 2$$

$$False\ Positives = 5$$

$$False\ Negatives = 1$$

$$F2\ Score = \frac{2}{2 + (0.2 * 5 + 0.8 * 1)} = 52.6\%$$

$$\text{Mean } F2\ Score = \frac{\sum F2\ Scores}{\text{\# of Topics}}$$

# Conclusion

- ✓ Current efforts to align digital materials to national curriculum are manual and require time, resources, and curricular expertise, and the process needs to be made more efficient to be scalable and sustainable.
- ✓ As new materials become available, they require additional efforts to be realigned, resulting in a never-ending process.
- ✓ There are no current algorithms or other AI interventions that address the resource constraints associated with improving the process of curriculum alignment.

# Future Scope

- User feedback for recommendations and model improvement (collaborative filtering).
- Generating descriptions for the topics from their topic tree and then giving recommendations.
- Make our model multilingual.
- Retrain the model on pre-trained embeddings.
- Summarizing the topic description and content description.
- Add search engine feature on the app instead of drop down.

# THANK YOU!
# QUESTIONS?

# Appendix

| Model | Average F2 Score | Average Precision @k |
|---|---|---|
| all-MiniLM-L6-v2 | 19.73 | 21.04 |
| multi-qa-mpnet-base-dot-v1 | 21.58 | 25.14 |
| paraphrase-MiniLM-L3-v2 | 12.20 | 11.93 |
| multi-qa-MiniLM-L6-cos-v1 | 20.45 | 21.97 |

With Description + Topic Tree

| Model | Average F2 Score | Average Precision @k |
|---|---|---|
| all-MiniLM-L6-v2 | 14.22 | 15.30 |
| paraphrase-MiniLM-L3-v2 | 8.15 | 8.46 |
| multi-qa-MiniLM-L6-cos-v1 | 13.23 | 14.28 |

Only Topic Tree

## all-MiniLM-L6-v2

This is a sentence-transformers model: It maps sentences & paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search.

## multi-qa-mpnet-base-dot-v1

This is a sentence-transformers model: It maps sentences & paragraphs to a 768 dimensional dense vector space and was designed for **semantic search**. It has been trained on 215M (question, answer) pairs from diverse sources. For an introduction to semantic search, have a look at: SBERT.net - Semantic Search

## sentence-transformers/paraphrase-MiniLM-L3-v2

This is a sentence-transformers model: It maps sentences & paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search.

## multi-qa-MiniLM-L6-cos-v1

This is a sentence-transformers model: It maps sentences & paragraphs to a 384 dimensional dense vector space and was designed for **semantic search**. It has been trained on 215M (question, answer) pairs from diverse sources. For an introduction to semantic search, have a look at: SBERT.net - Semantic Search