Live FAQ – Workshop

Fall 2022 CS 6220 – Big Data Systems & Analytics Project

Group #8 – Andrea Covre, Anshul Gupta, Prahlad Jasti, Akash Nainani, Rishabh Th

Problem

- Virtual Events are the new reality
- Live Events now have an associated public commentary
- Conferences, Online Classes, Live Sports Events: Examples of such events

- Attendees often come up with questions while participating
- Often common and easily resolved in an in-person event
- But clarification and responses is not up to mark in virtual events

High-Level Approach/Solution

Collect Question

Identify Semantical information

Cluster together similar questions

Gener
F

Sentence Level Embeddings

- Using Pretrained NLP SOTA models to obtain embedding
- Comparative analysis of results from each embedding
- Use embedding to obtain a semantic representation

Data

- Target Live Data but Training Questions Dataset
  - Question Classification Experiment Dataset – Topic and Subtopic Categorization
  - Quora Question pair dataset
  - Kaggle Question Dataset
- Clustering Evaluation: Question Classification with labels as ground truth
- Question Classification has ~5500 questions split across 6 topics and 50 subtopics

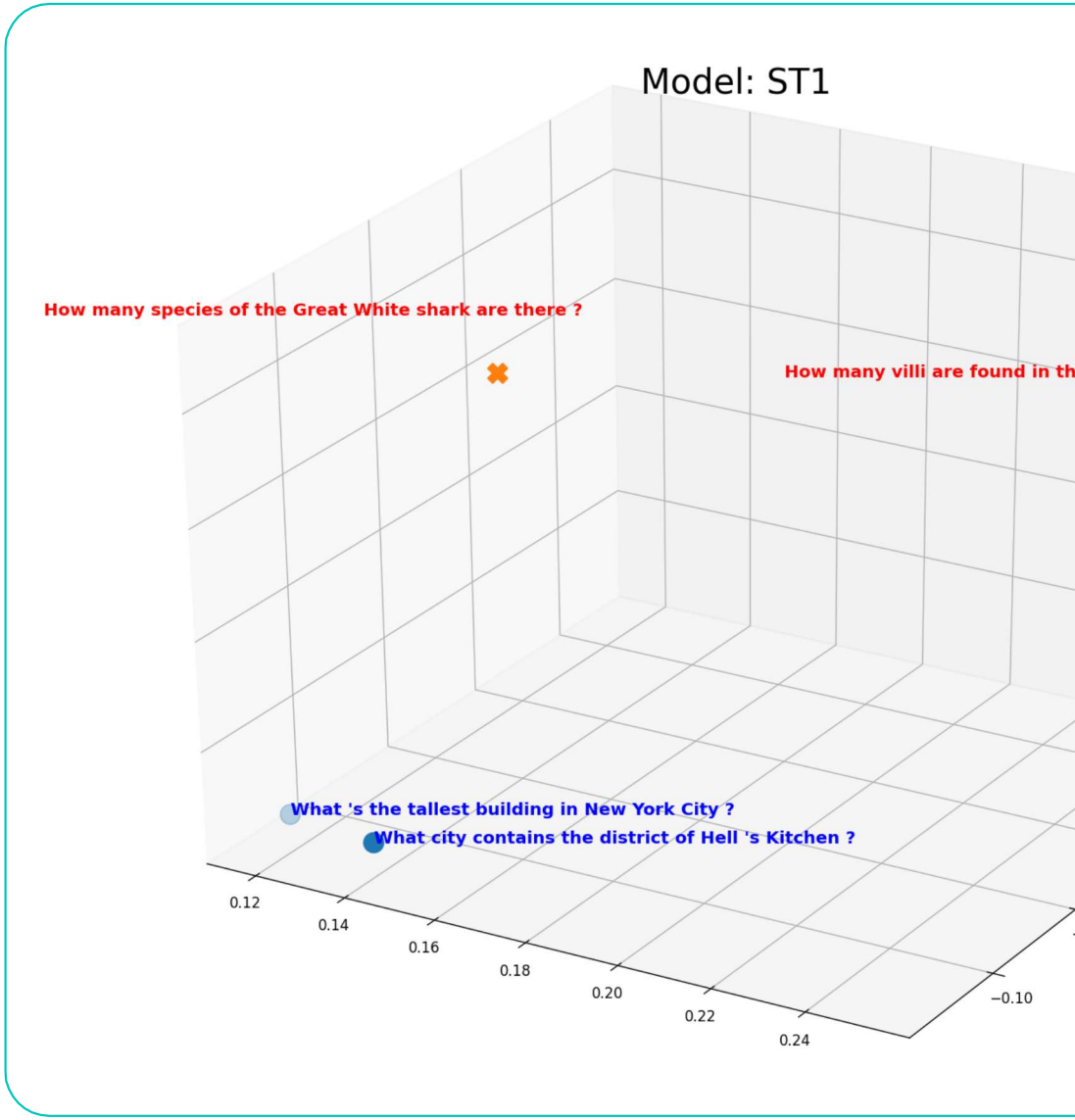
Clustering

- Clustering Methods
  - K-Means
    - Euclidean
    - Cosine
  - GMM
    - Full Covariance
    - Diagonal Covariance
  - Agglomerative/Hierarchical
- Clustering Metrics
  - Homogeneity Score
  - Adjusted Mutual Info S
  - Silhouette Coefficient

Model 1 – Sentence Transformer 1

- Model: all-mpnet-base-v2
- Maps sentences & paragraphs to a 768-dimensional dense vector space
- Pretrained microsoft/mpnet-base model and fine-tuned in on a 1.17 Billion sente dataset
- Training Corpus: Reddit Comments, Yahoo Answers, Wiki Answers, Stack Exchan

Embedding: Results



Clustering: Results

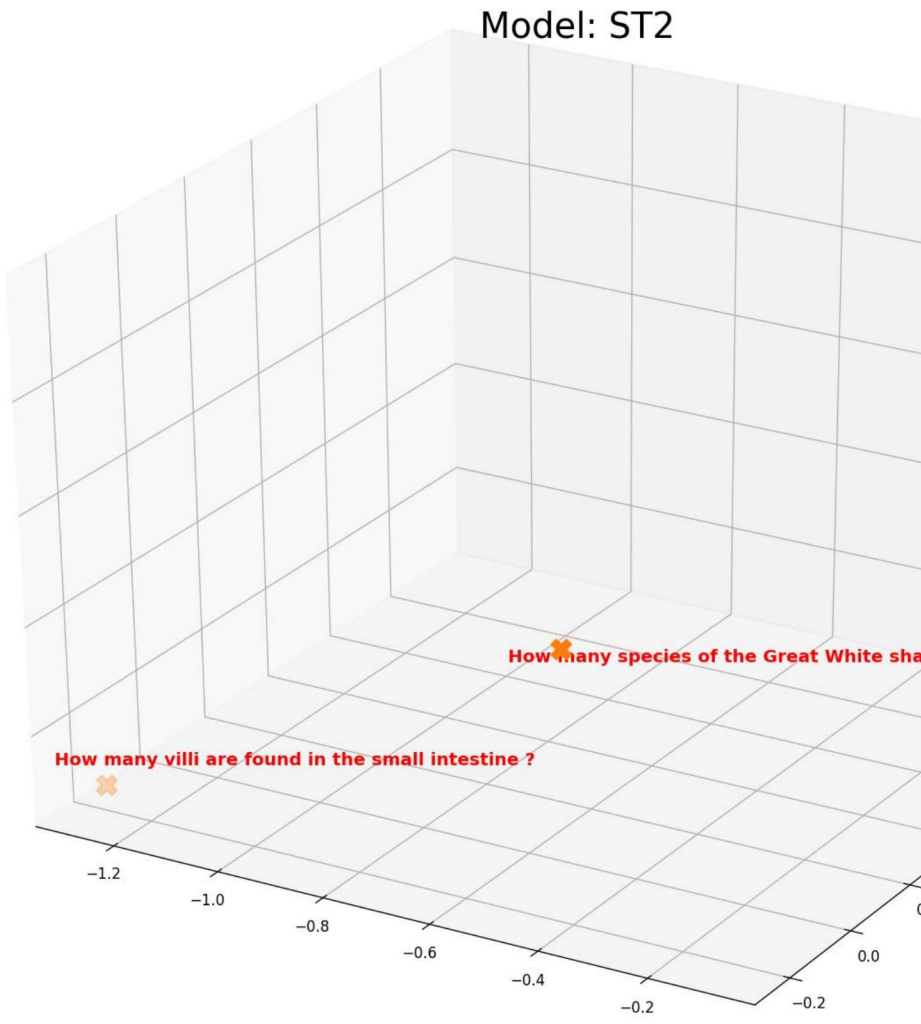
Clustering Metrics

Model	Algorithm	Homogenity Score	AMI Score	Silh
ST1	KMeans - Euclidean	0.420150696	0.34114179	
ST1	KMeans - Cosine	0.412456088	0.33271543	
ST1	Agglomerative/Hierarchical	0.332324997	0.26080778	
ST1	GMM - Full Covariance	0.415197024	0.33582861	
ST1	GMM - Diagonal Covariance	0.410623804	0.33230363	

Model 2 – Sentence Transformer 2

- Model: multi-qa-mpnet-base-dot-v1
- Maps sentences & paragraphs to a 768-dimensional vector designed for seman
- Concatenated multiple datasets to fine-tune with total about 215 Million (questi answer) pairs
- Training Corpus: Quora, Search QA, Yahoo Answers, Wiki Answers, Stack Exchan

Embedding: Results



Clustering: Results

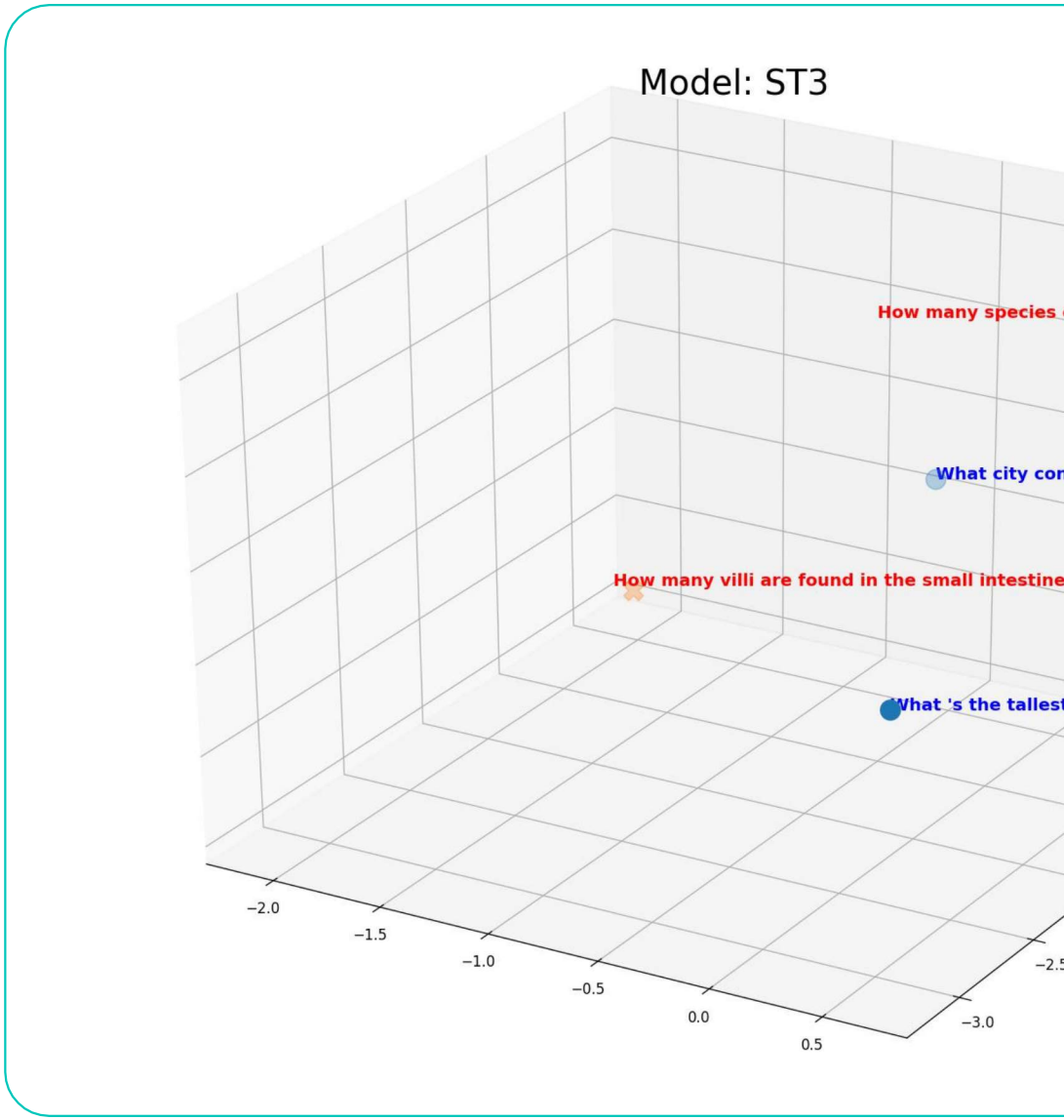
Clustering Metrics

Model	Algorithm	Homogenity Score	AMI Score	Silho
ST2	KMeans - Euclidean	0.386968498	0.309111641	
ST2	KMeans - Cosine	0.392326598	0.31383509	
ST2	Agglomerative/Hierarchical	0.328385803	0.257067261	
ST2	GMM - Full Covariance	0.376071663	0.300468646	
ST2	GMM - Diagonal Covariance	0.398971553	0.321492183	

Model 3 – Sentence Transformer 3

- Model: nq-distilbert-base-v1
- Optimized for question-answer retrieval
- Maps sentences & paragraphs to a 768-dimensional vector designed for seman
- Training Corpus: 100k real search queries from Google with the respective, relev passage from Wikipedia

Embedding: Results



Clustering: Results

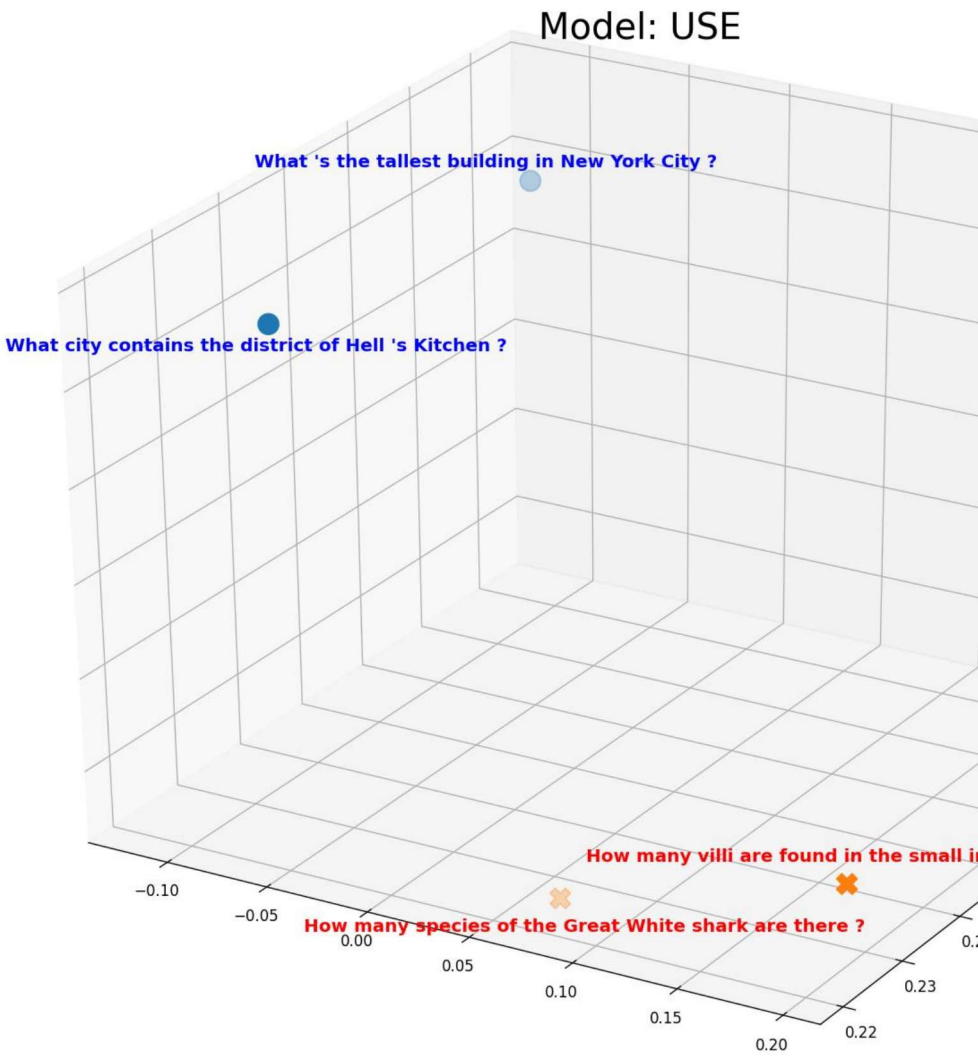
Clustering Metrics

Model	Algorithm	Homogenity Score	AMI Score	Silh
ST3	KMeans - Euclidean	0.419448891	0.339101789	
ST3	KMeans - Cosine	0.416712974	0.3360484	
ST3	Agglomerative/Hierarchical	0.331755941	0.261090344	
ST3	GMM - Full Covariance	0.423726098	0.345137056	
ST3	GMM - Diagonal Covariance	0.438488037	0.360037428	

Model 4 - USE

- Universal Sentence Encoder by Google
- Transformer and Transfer Learning based model
- Training Corpus: Wikipedia, web news, web question-answer pages and discuss
  Stanford Natural Language Inference (SNLI) corpus

Embedding: Results



Clustering: Results

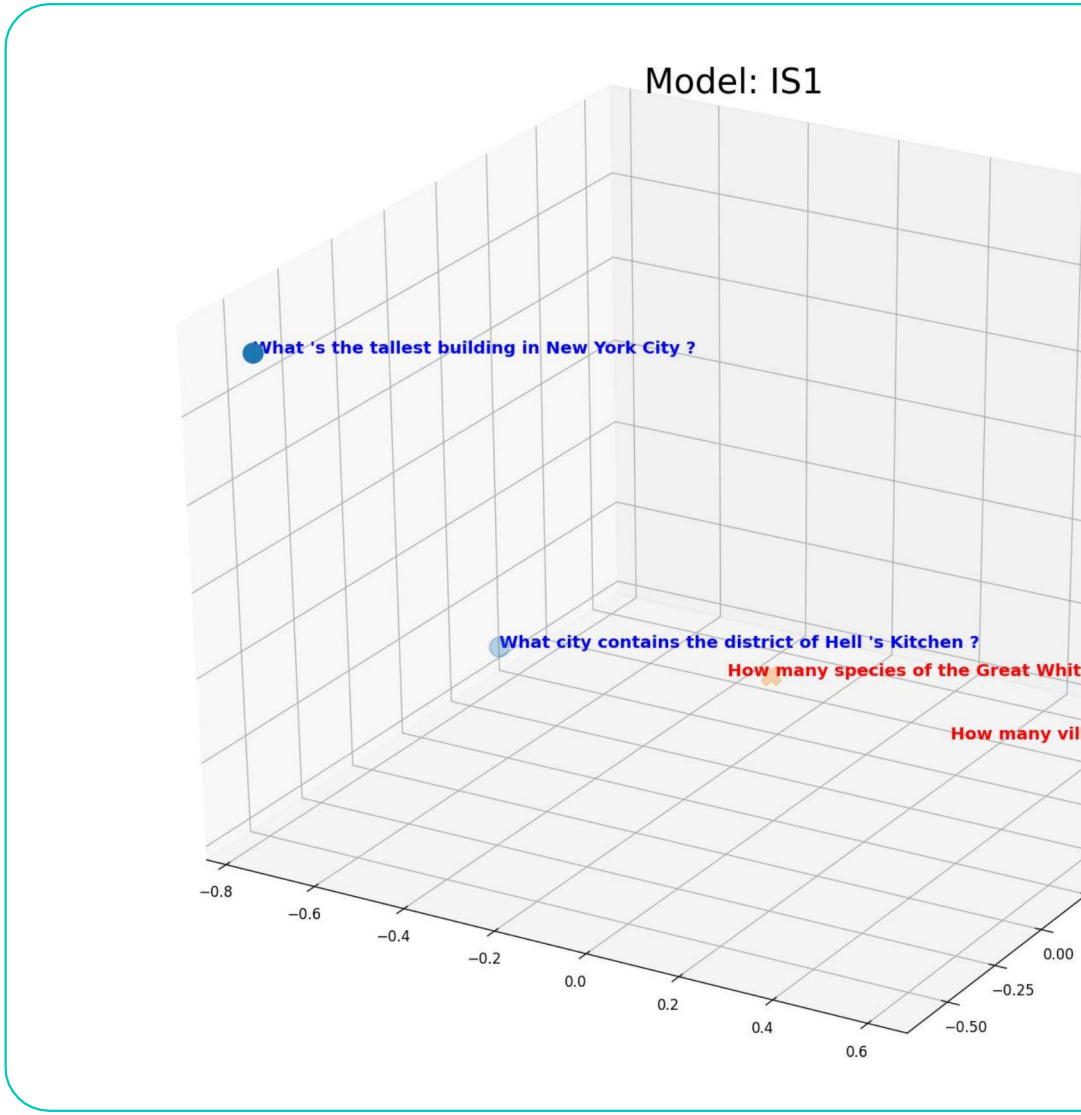
Clustering Metrics

Model	Algorithm	Homogenity Score	AMI Score	Silh
USE	KMeans - Euclidean	0.504007275	0.423351968	
USE	KMeans - Cosine	0.495579341	0.413522818	
USE	Agglomerative/Hierarchical	0.362255337	0.290531284	
USE	GMM - Full Covariance	0.500522032	0.418846994	
USE	GMM - Diagonal Covariance	0.523115464	0.442591815	

Model 5 - InferSent1

- Developed by Facebook Research
- Bi-directional LSTM architecture + GRU + Pooling
- Training Corpus: Stanford Natural Language Inference (SNLI) dataset (570k sente pairs)
- Designed for general purpose
- Embedding Dimensions: 512, 1024, 2048, 4096
- Word Embeddings are initialized by GloVe embeddings

Embedding: Results



Clustering: Results

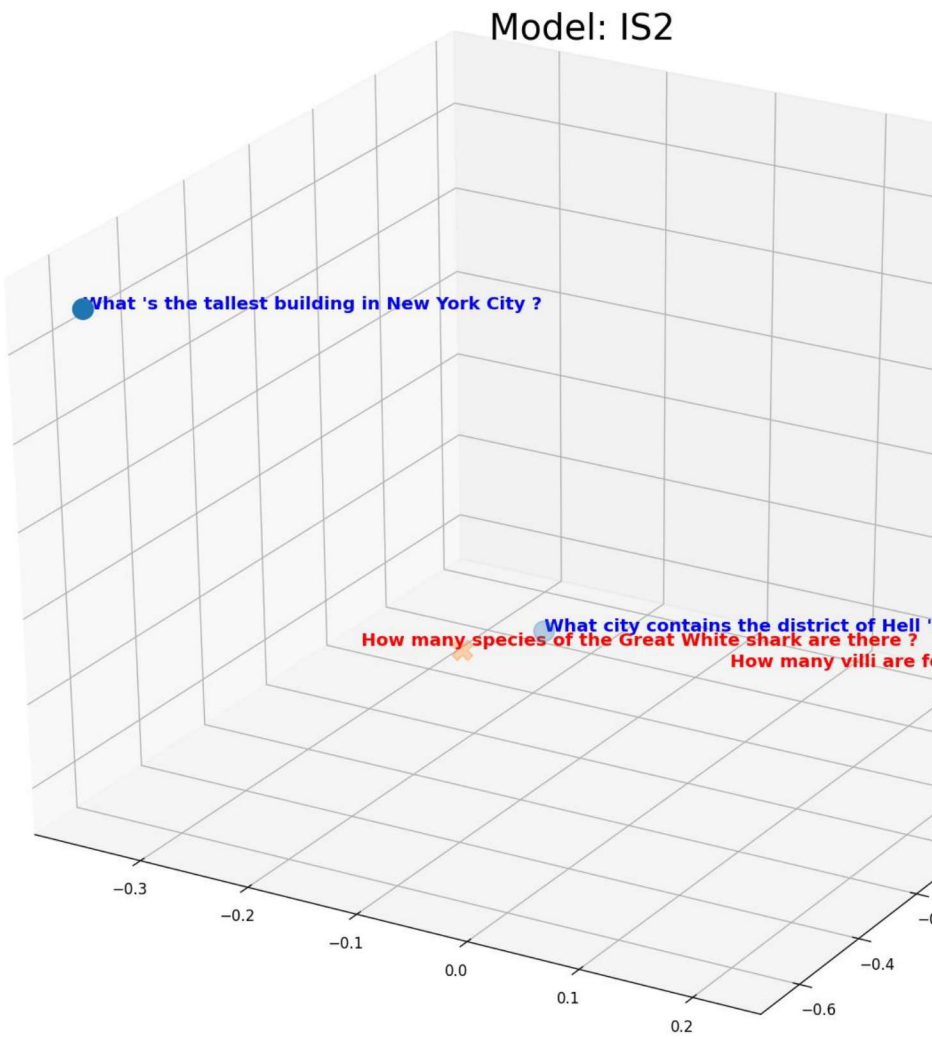
Clustering Metrics

Model	Algorithm	Homogenity Score	AMI Score	Silho
IS1	KMeans - Euclidean	0.316975274	0.241148594	
IS1	KMeans - Cosine	0.338050454	0.261439298	
IS1	Agglomerative/Hierarchical	0.316645453	0.243098864	
IS1	GMM - Full Covariance	0.330287743	0.254203281	
IS1	GMM - Diagonal Covariance	0.336738098	0.2601475	

Model 6 - InferSent2

- Developed by Facebook Research
- Same as InferSent1 except that Word Embeddings are initialized by FastText em

Embedding: Results



Clustering: Results

Clustering Metrics

Model	Algorithm	Homogenity Score	AMI Score	Silh
IS2	KMeans - Euclidean	0.345309114	0.267908473	
IS2	KMeans - Cosine	0.343147882	0.267467154	
IS2	Agglomerative/Hierarchical	0.327020738	0.256079877	
IS2	GMM - Full Covariance	0.333848458	0.258926877	
IS2	GMM - Diagonal Covariance	0.370199056	0.294539959	

Clustering Metrics

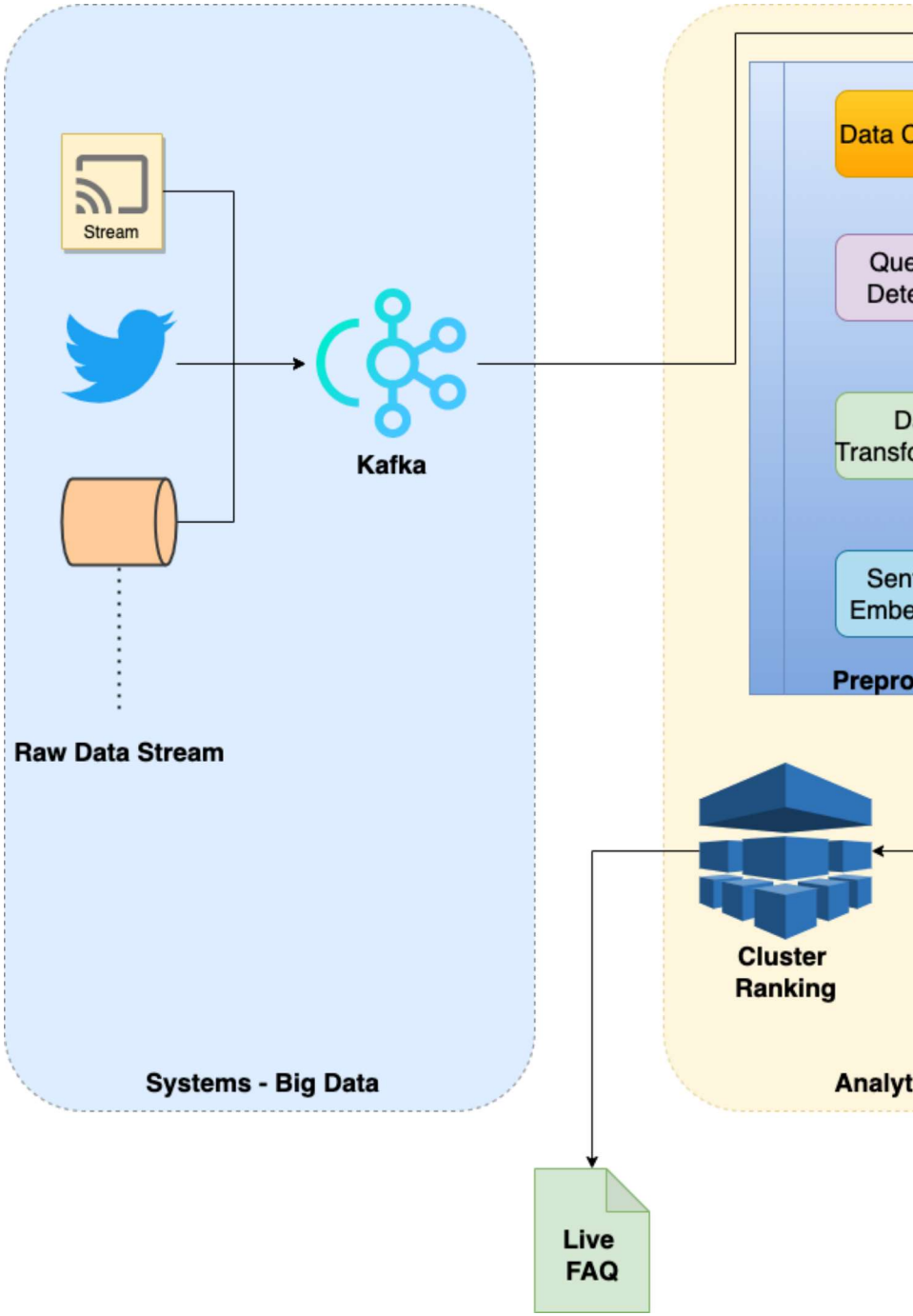
Clustering: Overall Results

Model	Algorithm	Homogenity Score	AMI Score
ST1	KMeans - Euclidean	0.420150696	0.3411417
ST1	KMeans - Cosine	0.412456088	0.3327154
ST1	Agglomerative/Hierarchical	0.332324997	0.2608077
ST1	GMM - Full Covariance	0.415197024	0.3358286
ST1	GMM - Diagonal Covariance	0.410623804	0.3323036
ST2	KMeans - Euclidean	0.386968498	0.3091116
ST2	KMeans - Cosine	0.392326598	0.313835
ST2	Agglomerative/Hierarchical	0.328385803	0.2570672
ST2	GMM - Full Covariance	0.376071663	0.3004686
ST2	GMM - Diagonal Covariance	0.398971553	0.3214921
ST3	KMeans - Euclidean	0.419448891	0.3391017
ST3	KMeans - Cosine	0.416712974	0.33604
ST3	Agglomerative/Hierarchical	0.331755941	0.2610903
ST3	GMM - Full Covariance	0.423726098	0.3451370
ST3	GMM - Diagonal Covariance	0.438488037	0.3600374
USE	KMeans - Euclidean	0.504007275	0.4233519
USE	KMeans - Cosine	0.495579341	0.4135228
USE	Agglomerative/Hierarchical	0.362255337	0.2905312
USE	GMM - Full Covariance	0.500522032	0.4188469
USE	GMM - Diagonal Covariance	0.523115464	0.4425918
IS1	KMeans - Euclidean	0.316975274	0.2411485
IS1	KMeans - Cosine	0.338050454	0.2614392
IS1	Agglomerative/Hierarchical	0.316645453	0.2430988
IS1	GMM - Full Covariance	0.330287743	0.2542032
IS1	GMM - Diagonal Covariance	0.336738098	0.26014
IS2	KMeans - Euclidean	0.345309114	0.2679084
IS2	KMeans - Cosine	0.343147882	0.2674671
IS2	Agglomerative/Hierarchical	0.327020738	0.2560798
IS2	GMM - Full Covariance	0.333848458	0.2589268
IS2	GMM - Diagonal Covariance	0.370199056	0.2945399

System Implementation

- Kafka Server – Scalability for varying applications
- Python Producer – Compatibility with existing software tools
- Python Consumer - Integration with clustering pipeline

High Level Architecture Diagram



Future Work & Challenges

- Clustering Metrics – Merged Data
- Hyperparameter Tuning for HDBScan
- Exploring Doc2Vec
- Live Streaming – Kafka Ingestion & BIRCH Clustering integration

QnA