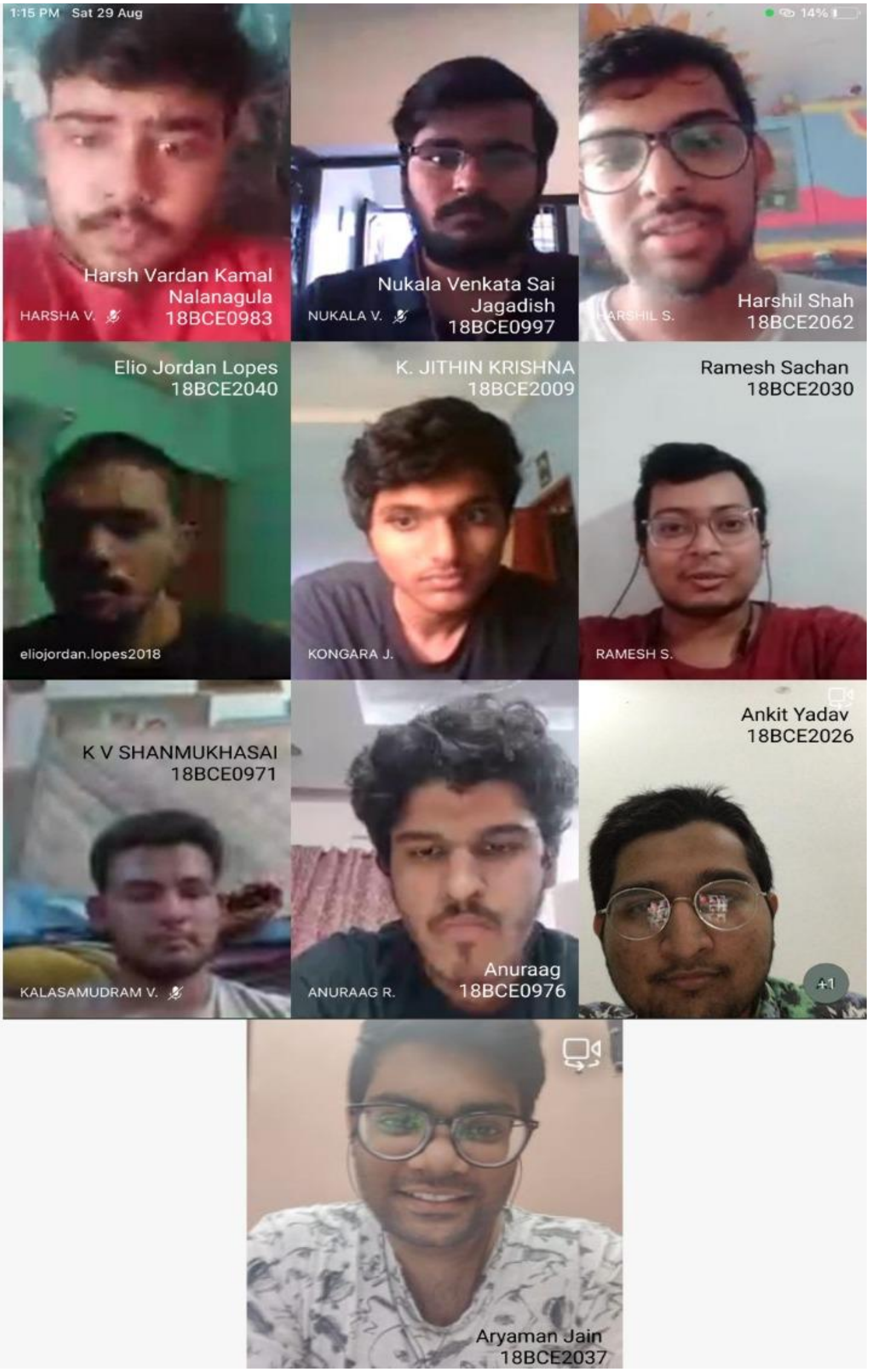
Nat Geo Group Activity - Hiring Ms. Lakshmi Sankaran

Team 5



Pipeline

1. Tokenization
2. Stemming (Optional see if it helps)
3. Remove stopwords
4. Create a dictionary with tf idf eMBEDDINGS
5. convert the paragraph to embeddings
6. Use PCA to reduces the vector size(optional)
7. Import k means and create cluster
8. use cosine similarity to find the similarity between two sentences
9. import Matplotlib Linkage dendograms
10. Create merging and plot dendogram

In this project, when we talk about basic structure, we used NLTK and Genism. NLTK is the famous library for NLP and easy to understand and it includes libraries for Tokenization, parsing, classificatiom, stemming, tagging and semantic reasoning. Genism is leading package for package for processing texts, working with vector models like Word2Vev.

Process:
After tokenizing the sentences, words. A dictionary is created using Genism which creates a dictionary object that maps each word to a unique id. Gensim reads the corpus and update the dictionary, without loading the entire textfile into system memory.

Bag Of words:
To work with genism we need to create corpus(A bag of words). This is basically a object which contains id ad its frequency in each document. Create a bag of words and pass the tokenized list of words to Dictionary.doc2bow()

TF-IDF:
It is calculated using TF and IDF. Term frequency is how many times words appear in document and IDF scales how rare the word. After this a similarity object is created, which builds an index for a given set of documents. Similarity class splits the index into several smaller sub-indexes, which are disk-based. After this, we will calculate how similar the corpus to each document and document similarities. After this query document similarity, we calculate average similarity by taking sum of similarities divided by number of similarities.

Document Similarity:
Calculate the similarity b/w all the documents

Average Similarity:
We calculated the average similarity using numpy

We used Colab notebook for executing the code. To view our code click here!