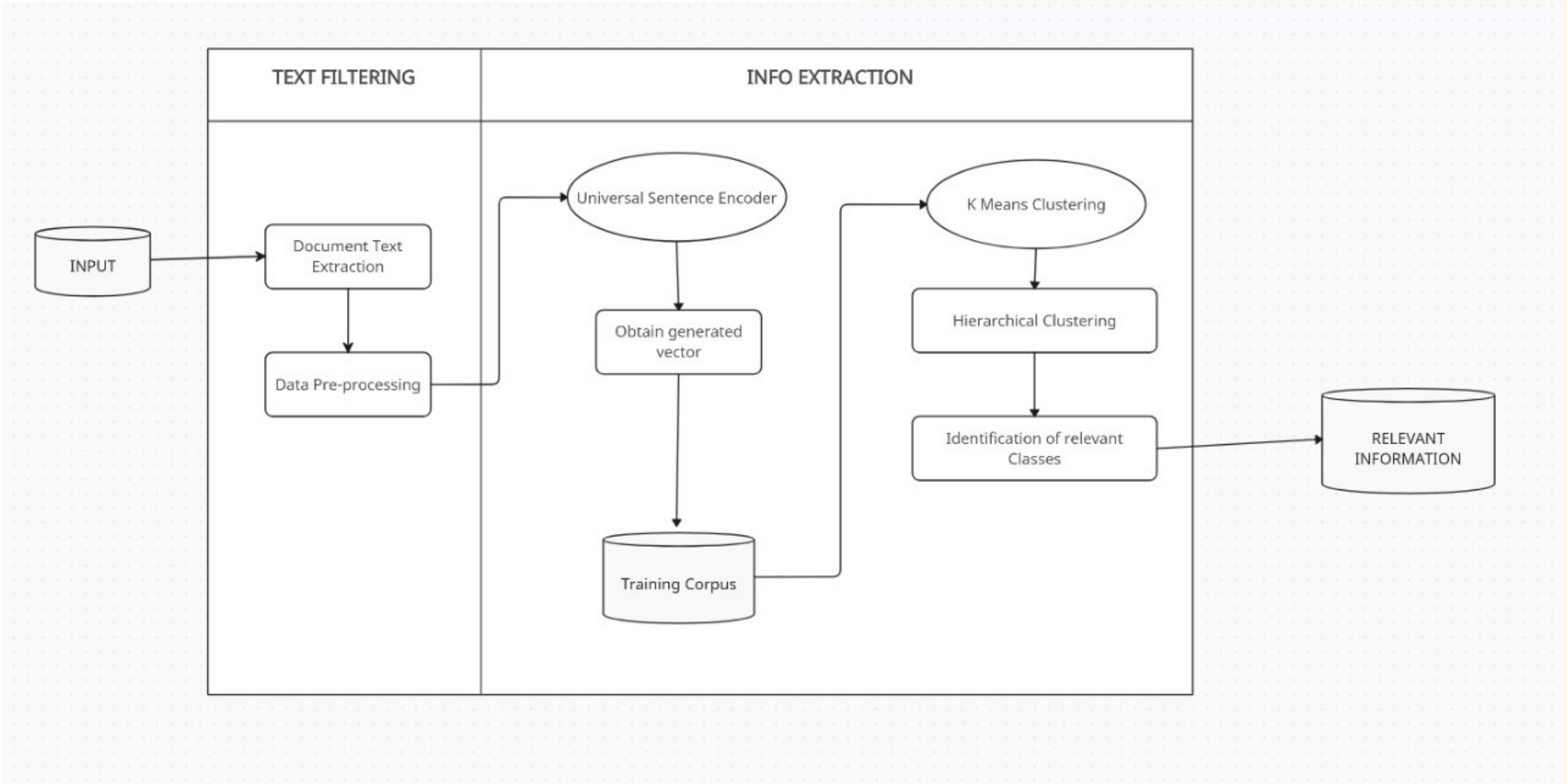
EXTRACTING SIGNIFICANT INFORMATION USING LARGE LANGUAGE MODELS(LLM)

Team Members:
Rakshith Arya - 4NI20CS075
Shashank K R - 4NI20CS094
Shashank Shandilya - 4NI20CS096
Vishal M V - 4NI20CS124

Project Guide:
Dr. Shabana Sultana

LIST OF CONTENTS

SYSTEM DESIGN



IMPLEMENTATION

DATA EXTRACTION

The code reads text files (specifically Microsoft Word .docx documents) and breaks them into individual paragraphs. This is essential because the later clustering techniques work best with smaller, self-contained units of meaning. Information extraction and analysis are much easier to perform when dealing with well-defined segments of text instead of one big block.

```
['ABSOLUTE SALE DEED\t\t\t                SITE NO : 08', 'This Deed of Sale of the Scheduled property is made on this 15th day of December, Two Thousand and Twenty One (15
['ABSOLUTE SALE DEED\t\t\t      \t\t   SITE NO : 24', 'This Deed of Sale of the Scheduled property is made on this 10th day of December, Two Thousand and Twenty One (10/12/
['ABSOLUTE SALE DEED\t\t\t              SITE NO : 53', 'This Deed of Sale of the Scheduled property is made on this 06th day of June, Two Thousand and Twenty Two (06-06-20
['ABSOLUTE SALE DEED \t\t\t\t      SITE NO: 03', 'This Deed of Sale of the Scheduled property is made on this 21st day of September, Two Thousand and Twenty One (21/09/2021
['ABSOLUTE SALE DEED\t\t\t              SITE NO : 45', 'This Deed of Sale of the Scheduled property is made on this 9th day of November, Two Thousand and Twenty One (09/11
['ABSOLUTE SALE DEED\t\t\t                   SITE NO:16', 'This Deed of Sale of the Scheduled property is made on this 4th day of February, Two Thousand and Twenty Three (
['ABSOLUTE SALE DEED\t\t\t              SITE NO : 42', 'This Deed of Sale of the Scheduled property is made on this 08th day of June, Two Thousand and Twenty Two (08-06-20
['ABSOLUTE SALE DEED\t\t\t                 SITE NO : 22', 'This Deed of Sale of the Scheduled property is made on this 11th day of January, Two Thousand and Twenty Three (
['ABSOLUTE SALE DEED\t\t\t      SITE NO:46', 'This Deed of Sale of the Scheduled property is made on this 13th day of October, Two Thousand and Twenty One (13/10/2021) by -
['ABSOLUTE SALE DEED\t\t\t      SITE NO:43', 'This Deed of Sale of the Scheduled property is made on this 8th day of October, Two Thousand and Twenty One (08/10/2021) by --
['ABSOLUTE SALE DEED\t\t\t                SITE NO : 06', 'This Deed of Sale of the Scheduled property is made on this 26th day of December, Two Thousand and Twenty Two (2
['ABSOLUTE SALE DEED\t\t\t      SITE NO:05', 'This Deed of Sale of the Scheduled property is made on this 18th day of October, Two Thousand and Twenty One (18/10/2021) by -
['ABSOLUTE SALE DEED\t\t\t              SITE NO : 40', 'This Deed of Sale of the Scheduled property is made on this 5th day of September, Two Thousand and Twenty Two (05-0
['ABSOLUTE SALE DEED \t\t\t\t      SITE NO: 09', 'This Deed of Sale of the Scheduled property is made on this 21st day of September, Two Thousand and Twenty One (21/09/2021
['ABSOLUTE SALE DEED\t\t\t                SITE NO : 20', 'This Deed of Sale of the Scheduled property is made on this 27th day of June, Two Thousand and Twenty Two (27-06-
['ABSOLUTE SALE DEED\t\t\t      \t\t   SITE NO:19', 'This Deed of Sale of the Scheduled property is made on this 27th day of October, Two Thousand and Twenty One (27/10/202
['ABSOLUTE SALE DEED\t\t\t                 SITE NO : 58', 'This Deed of Sale of the Scheduled property is made on this 18th day of January, Two Thousand and Twenty Three (
['ABSOLUTE SALE DEED\t\t\t              SITE NO:07', 'This Deed of Sale of the Scheduled property is made on this 18th day of October, Two Thousand and Twenty One (18/10/2
['SALE AGREEMENT', 'This Agreement of Sale has been made on Nineteenth Day of August Two Thousand Twenty Two (19-08-2022)', 'Sri.SHIVSHANKAR GANGADHAR DUDHALE ALIAS SHIVSHA
['ABSOLUTE SALE DEED\t\t\t      \t\t   SITE NO:15', 'This Deed of Sale of the Scheduled property is made on this 25th day of October, Two Thousand and Twenty One (25/10/202
['ABSOLUTE SALE DEED \t\t\t\t      SITE NO: 23', 'This Deed of Sale of the Scheduled property is made on this 20th day of September, Two Thousand and Twenty One (20/09/2021
['ABSOLUTE SALE DEED\t\t\t                  SITE NO:12', 'This Deed of Sale of the Scheduled property is made on this 13th day of October, Two Thousand and Twenty One (13/
['ABSOLUTE SALE DEED\t\t\t      \t\t   SITE NO:41', 'This Deed of Sale of the Scheduled property is made on this 30th day of October, Two Thousand and Twenty One (30/10/202
['SALE AGREEMENT', 'This Agreement of Sale has been made on Twenty Seventh Day of October Two Thousand Twenty Two (27-10-2022)', 'Sri. SHIVSHANKAR GANGADHAR DUDHALE ALIAS S
['ABSOLUTE SALE DEED\t\t\t              SITE NO : 29', 'This Deed of Sale of the Scheduled property is made on this 24th day of February, Two Thousand and Twenty Two (24/0
['ABSOLUTE SALE DEED\t\t\t              SITE NO : 57', 'This Deed of Sale of the Scheduled property is made on this 08th day of June, Two Thousand and Twenty Two (08-06-20
['SALE AGREEMENT', 'This Agreement of Sale has been made on First Day of October Two Thousand Twenty Two (01-10-2022)', 'Sri. SHIVSHANKAR GANGADHAR DUDHALE ALIAS SHIVSHANKA
['ABSOLUTE SALE DEED\t\t\t              SITE NO : 30', 'This Deed of Sale of the Scheduled property is made on this 24th day of February, Two Thousand and Twenty Two (24/0
['SALE AGREEMENT', 'This Agreement of Sale has been made on Twenty Fourth Day of January Two Thousand Twenty Three (24-01-2023)', 'Sri. SHIVSHANKAR GANGADHAR DUDHALE ALIAS
['ABSOLUTE SALE DEED\t\t\t\t\tSITE NO. 45', 'This Deed of Sale of the Scheduled property is made on this 24th  day of March, Two Thousand Twenty Two (24-03-2022) by ---', '
```

DATA PRE-PROCESSING

Pre-processing usually involves removing extra character, misspellings or converting text to lowercase. In our case, we remove the punctuations and ensure that the paragraphs are split uniformly. We also remove misspellings and unrecognized text (Languages other than English) from the extracted paragraphs. Cleaning up text makes it more uniform. This helps the clustering to be more precise.

	A	B	C	E
1		id	filenar	para_text
2	0	7e5	/Users	ABSOLUTE SALE DEED SITE NO: 03
3	1	7d4	/Users	This Deed of Sale of the Scheduled property is made on this 21st day of September, Two Thousand and Twenty One (21/09/2021) by ---
4	2	d81	/Users	Sri. SHIVSHANKAR GANGADHAR DUDHALE ALIAS SHIVSHANKAR DUDHALE, S/o.Sri.Gangadhar Dudhale, (PAN No.AISPD6976G, AADHAAR No. 3494 9796 2818) aged about 54 years, residing at No. 44/A/404, ‚ÄúAmeesh CHS‚Äù, Near Tilak Nagar Police Station, Chemburu, Tilak Nagar, Mumbai-400089 Hereinafter called as the VENDOR.
5	3	839	/Users	AND
6	4	2ef	/Users	Smt. SWATI, D/o. Sri. Narasimha Bhat.C.H, (PAN No.DANPS0517N, AADHAAR No.6484 0543 7193) aged about 31 years, residing at No. ‚ÄúSwati House‚Äù, Near Government Junior College, J.C.Road, P.O. Sullia, Sullia Taluk, Dakshina Kannada District, Karnataka-574 239. Represented by her GPA Holder Sri. NARASIMHA BHAT.C.H, S/o. G. Narayana Bhat, (PAN No. AIOPC7995E, AADHAAR No. 9582 9138 0829) Hereinafter called the PURCHASER.

WORD EMBEDDING

The code uses a powerful model called the Universal Sentence Encoder. It transforms each paragraph into a set of numbers (a vector). Vectors that represent similar paragraphs will be positioned closer together in a "conceptual space." This is very helpful in this case as this research focuses on unstructured data that usually does not follow a pattern. LLM's are trained on many different types of documents and text making it a better choice compared to NLP for this research.

	A	B	C	D	E
1		id	filenam	para_embedding	para_text
2	0	7e5	/Users/	0.01919835 -0.00022775 -0.01170207 -0.08491252 -0.05445177 0.01754978 -0.03626426 0.02691346 0.02986935 0.02120393 -0.04411981 -0.07442166 0.03225924 0.03864637 0.03802371 -0.05136845 0.04663963 0.05660635 0.03522041 -0.08006301 0.06875233 0.06793799 0.06745549 0.0455267 0.0033949 -0.06177441 0.00796594 -0.06026107 0.01271578 -0.05028609 -0.04109176 -0.03807456 -0.06231509 -0.04926367 -0.04240713 -0.05570034 0.05562764 0.05908606 -0.03294408 -0.01760802 0.01835042 0.02954552 -0.08458474 -0.01223406 0.05303831 -0.00773792 -0.05052388 -0.02522261 0.01962849 0.01503444 0.08459685 -0.0344249 0.03816418 -0.05546874 0.00776353 0.00672047 0.00477478 0.01369538 0.06360558 -0.0430297 0.01206642 0.0142039 0.02925353 0.02786294 0.02192249 0.00857271 0.04568698 0.06043746 -0.01983981 0.05821707 -0.02521453 0.06007453 -0.05577978 -0.05642657 0.05195663 -0.02640818 0.03002103 0.07379806 -0.06450544 -0.01367169 0.02434548 0.05942231 -0.02173761 0.03602776 -0.00462255 0.06937031 -0.03036185 -0.08573708 -0.07206458 -0.04423633 0.04479038 0.08640809 -0.05194912 0.05798733 -0.00965466 0.01064469 -0.03871823 -0.04317361 -0.038232 -0.04357472 0.03170309 0.01232866 -0.08247638 -0.07505359 -0.05186242 -0.06841182 0.00687262 -0.02720723 0.05431673 0.06820037 0.02520368 0.00222665 0.0590063 0.07175436 0.07861171 0.05419076 -0.05416563 -0.0063569 0.01833371 -0.08134428 -0.0503979 0.04761592 0.03752502 -0.03286331 0.04392654 -0.04278677 0.03843147 0.03151418 0.06571569 -0.05353945 0.02661062 -0.05478328 -0.02588535 -0.06062259 -0.00351928 0.03581516 0.00418107 0.02781007 -0.07986212 -0.04259096 -0.04747473 0.00160637 -0.05334257 -0.04858624 0.06254724 0.04831353 -0.02517878 -0.07564035 0.04334414 0.06929892 -0.00994972 -0.05478634 0.02482563 0.00398958 -0.01385602 0.00485301	ABSOLUTE SALE DEED SITE NO: 03

KMEANS CLUSTERING

K-means is a classic algorithm for finding groups when performing unsupervised learning. Clustering helps discover patterns in the paragraphs. Paragraphs in the same cluster probably address similar topics or themes. This algorithm is the perfect fit for this research as we are not clear about the patterns or the topics in the documents. The algorithm takes the paragraph vectors as input andtries to find "centers" where similar paragraphs are clustered. Each paragraph gets assigned to its closest center. The number of clusters is decided beforehand.

	id	filename	para_embedding	para_text	k_means_labels
0	7e506bd6-f9	/Users/shash	[-0.0460757 -0.03784995 0.00855281 -0.00814807 0.05272095 -	ABSOLUTE SALE DEED SITE NO: 03	0
1	7d409b82-e	/Users/shash	[-6.53896481e-02 -5.14134988e-02 2.72322465e-02 6.64128587e-02	This Deed of Sale of the Scheduled property is made on this 21st day c	1
2	d810c249-c	/Users/shash	[-0.04715949 0.06256256 -0.04148636 -0.05242401 -0.01417808	Sri. SHIVSHANKAR GANGADHAR DUDHALE ALIAS SHIVSHANKAR DUD	4
3	83963589-d	/Users/shash	[-1.41516859e-02 6.94209593e-04 5.53115122e-02 1.87204182e-02	AND	2
4	2ef1c954-27	/Users/shash	[-0.02329692 -0.03527451 -0.01527591 -0.00126762 0.02784951	Smt. SWATI, D/o. Sri. Narasimha Bhat.C.H, (PAN No.DANPS0517N, AA	4
5	dffafcc3-b16	/Users/shash	[0.00057216 -0.03169177 0.00496114 0.01897375 0.06189654	WHEREAS, M/s.Janani Developers and Builders represented by its Part	3
6	28353314-d	/Users/shash	[0.01308791 -0.01476652 -0.0343005 -0.06010677 0.05414017	WHEREAS, the Vendor along with M/s. Janani Developers and Builders	3
7	94496f0a-1a	/Users/shash	[4.49535809e-02 -5.15193231e-02 -8.32074974e-03 7.27138249e-03	No. 95 dated 24-11-2018. That being the owner in possession of the a	3
8	de9ebad0-f2	/Users/shash	[0.04283376 -0.01681864 -0.0036458 -0.04658139 0.02843381	WHEREAS, Sri.Shivshankar Gangadhar Dudhale alias Shivshankar Dud	3
9	75f1448d-e6	/Users/shash	[3.11471485e-02 -8.09351653e-02 4.24961969e-02 -5.25744545e-05	Whereas both parties have broadly negotiated the terms and condition	1
10	fb5640db-09	/Users/shash	[5.57819232e-02 7.09354281e-02 6.76150844e-02 -4.28848155e-02	NOW THEREFORE THIS MEMORANDUM OF UNDERSTANDING WITNES	5
11	5020b0a2-bl	/Users/shash	[0.02985247 -0.08037467 0.05027387 0.04141634 0.05706681 -	1.The Vendor has offered to sell the Schedule Property to the Purchas	1
12	fffc2863-6f2	/Users/shash	[-6.54819086e-02 -6.85731992e-02 4.19579484e-02 3.58905457e-02	2.It was mutually agreed that the sale consideration paid by the purch	6
13	629065ad-e	/Users/shash	[-0.06213733 -0.06531172 0.00523691 -0.01024853 0.0446715 -	a.The Purchaser has paid a advance of Sale Consideration of Rs. 5,00,	6
14	a39e5d9d-ae	/Users/shash	[-6.91814646e-02 -6.94538206e-02 6.04189001e-03 1.38984993e-02	b. The Purchaser has paid a advance of Sale Consideration of Rs. 2,05	6
15	795cb358-f4	/Users/shash	[-7.64077976e-02 -4.72601466e-02 3.30936760e-02 -1.18303066e-02	c. The Purchaser has paid the remaining Sale Consideration of Rs. 17	6
16	74d092f6-72	/Users/shash	[0.01720439 -0.03720717 0.05632317 0.05898408 0.03791082 -	3.The Vendor has today delivered vacant peaceful possession of the Sc	1
17	cfb579e4-01	/Users/shash	[4.56525199e-02 -7.60463700e-02 3.63529362e-02 7.40821362e-02	4.The Vendor represents and assures the Purchaser that he has a clear	1
18	b0c97630-4	/Users/shash	[0.02570123 -0.06510617 -0.00602266 0.06447491 0.07375335 -	5.The Vendor represents and assures the Purchaser that the schedule	1

HIERARCHICAL CLUSTERING

At this stage the algorithm takes the regular K-means clustering and goes deeper. It takes one cluster at a time and applies K-means again within that cluster. This creates layers of groups—big groups broken down into smaller ones.

	para_embedding	filenam	para_text	k_means_labels_primary	k_means_labels_secondary
12	[-6.54819086e-02 -6.85731992e-02 4.19579484e-02 3.58905457e-02	/Users/shash	2.It was mutually agreed that the sale consideration paid by the purchaser for absolute sale of th	6	0
13	[-0.06213733 -0.06531172 0.00523691 -0.01024853 0.0446715 -	/Users/shash	a.The Purchaser has paid a advance of Sale Consideration of Rs. 5,00,000/- (Rupees Five lakh or	6	2
14	[-6.91814646e-02 -6.94538206e-02 6.04189001e-03 1.38984993e-02	/Users/shash	b. The Purchaser has paid a advance of Sale Consideration of Rs. 2,05,250/- (Rupees Two lakh f	6	2
15	[-7.64077976e-02 -4.72601466e-02 3.30936760e-02 -1.18303066e-02	/Users/shash	c. The Purchaser has paid the remaining Sale Consideration of Rs. 17,44,750/- (Rupees Seventy	6	2
39	[-5.8393799e-02 -4.9062911e-03 -3.4850225e-02 2.9226022e-02	/Users/shash	Witnesses:-	6	1
58	[-6.54819086e-02 -6.85731992e-02 4.19579484e-02 3.58905457e-02	/Users/shash	2.It was mutually agreed that the sale consideration paid by the purchaser for absolute sale of th	6	0
59	[-7.11608157e-02 -6.50484711e-02 2.58592051e-03 -4.95111709e-03	/Users/shash	a.The Purchaser has paid a advance of Sale Consideration of Rs. 2,80,000/- (Rupees Two lakh E	6	2
60	[-6.35709018e-02 -6.37473688e-02 2.06792653e-02 -3.20656896e-02	/Users/shash	b. The Purchaser has paid a advance of Sale Consideration of Rs. 3,70,000/- (Rupees Three lakh	6	2
61	[-7.50726536e-02 -4.85080145e-02 -3.65608744e-02 -2.25272938e-03	/Users/shash	c. The Purchaser has paid the remaining Sale Consideration of Rs. 18,00,000/- (Rupees Eighteer	6	2
85	[-5.8393799e-02 -4.9062911e-03 -3.4850225e-02 2.9226022e-02	/Users/shash	Witnesses:-	6	1
104	[-6.54819086e-02 -6.85731992e-02 4.19579484e-02 3.58905457e-02	/Users/shash	2.It was mutually agreed that the sale consideration paid by the purchaser for absolute sale of th	6	0
105	[-0.06586938 -0.05528237 -0.05518338 -0.02255814 0.04290388 -	/Users/shash	a.The Purchaser has paid a advance of Sale Consideration of Rs. 10,00,000 (Rupees Ten Lakh or	6	2
106	[-7.0960961e-02 -3.8694058e-02 -2.3669016e-02 -4.1604275e-05	/Users/shash	b. The Purchaser has paid the remaining Sale Consideration of Rs. 14,50,000/- (Rupees Fourtee	6	2
130	[-5.8393799e-02 -4.9062911e-03 -3.4850225e-02 2.9226022e-02	/Users/shash	Witnesses:-	6	1
149	[-8.47568139e-02 -7.22237602e-02 5.94600774e-02 3.09729669e-02	/Users/shash	2.It was mutually agreed that the sale consideration paid by the purchaser for absolute sale of th	6	0
150	[-7.4631847e-02 -5.5141289e-02 1.9243389e-02 -3.0407557e-02	/Users/shash	a.The Purchaser has paid a advance of Sale Consideration of Rs. 1,50,000/- (Rupees One lakh fi	6	2
151	[-0.07286442 -0.06404237 0.04524062 -0.03623231 0.03612331 -	/Users/shash	b. The Purchaser has paid a advance of Sale Consideration of Rs. 5,50,000/- (Rupees Five lakh f	6	2
152	[-7.62812421e-02 -6.45299852e-02 5.06763831e-02 -2.65577454e-02	/Users/shash	c. The Purchaser has paid a advance of Sale Consideration of Rs. 25,000/- (Rupees Twenty five	6	2
153	[-7.96713606e-02 -5.16908132e-02 4.89488691e-02 -2.29855739e-02	/Users/shash	d. The Purchaser has paid the remaining Sale Consideration of Rs. 17,34,500/- (Rupees Seven	6	2
177	[-5.8393799e-02 -4.9062911e-03 -3.4850225e-02 2.9226022e-02	/Users/shash	Witnesses:-	6	1
195	[-5.76763898e-02 -7.42441043e-02 4.32399027e-02 6.91270381e-02	/Users/shash	2.It was mutually agreed that the sale consideration paid by the purchaser for absolute sale of th	6	0
196	[-0.06564724 -0.05664584 -0.02930501 -0.04176994 0.05831785 -	/Users/shash	a.The Purchaser has paid a advance of Sale Consideration of Rs. 5,00,000 (Rupees Five Lakh onl	6	2
197	[-6.57660738e-02 -5.58683164e-02 -3.31906080e-02 -5.78261772e-03	/Users/shash	b. The Purchaser has paid a advance of Sale Consideration of Rs. 2,50,000 (Rupees Two Lakh F	6	2
198	[-0.07165776 -0.01871461 -0.03581569 0.01095073 -0.00239511 -	/Users/shash	c. The Purchaser has paid the remaining Sale Consideration of Rs. 14,30,000/- (Rupees Fourtee	6	2
221	[-5.8393799e-02 -4.9062911e-03 -3.4850225e-02 2.9226022e-02	/Users/shash	Witnesses:-	6	1
228	[-3.32084447e-02 -4.63857166e-02 5.38475513e-02 5.26178954e-03	/Users/shash	SALE AGREEMENT	6	3
229	[-0.07181025 -0.02104024 0.02760532 0.05249951 0.00411572	/Users/shash	This Agreement of Sale has been made on Nineteenth Day of August Two Thousand Twenty Two (	6	3
241	[0.01959835 -0.05869689 0.03247628 0.01562327 0.04494688 -	/Users/shash	That the VENDOR, Sri.Shivshankar Gangadhar Dudhale alias Shivshankar Dudhale has offered to	6	0
242	[-0.01165177 0.04749333 -0.00606331 0.03229045 -0.05826059 -	/Users/shash	Now this agreement of sale witnesses as hereunder:	6	3

RULE BASED TEXT EXTRACTION

Here we have used pattern-matching rules to find things like PAN and Aadhar card numbers within the paragraphs. It also has a way to spot phrases like "VENDOR" or "PURCHASER" to tag paragraphs. This is where we finally get the information we are looking for. The clustering helps to focus on the correct parts of the document, and then these rules extract the final details.

	filenam	para_text	k_means_labels_primary	p_name	p_pan	p_aadhar	v_name	v_pan	v_aadhar
2	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
4	/Users/shash	Smt. SWATI, D/o. Sri. Narasimha Bhat	4	SWATI	['DANPSXXXXN', 'AIOPCXXXXE']	['648X XXXX X193', '958X XXXX X829']			
48	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
50	/Users/shash	Sri. VEERABHADRA SWAMY K.M S/o. L	4	VEERABHADRA SWAMY K.M	['ABEPVXXXXK']	['370X XXXX X129']			
94	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
96	/Users/shash	Sri. MURALIDHAR S/o. Late K.B.Shiva	4	MURALIDHAR	['AEZPMXXXXK']	['291X XXXX X696']			
139	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
141	/Users/shash	Sri. RUDRAPRAKASH, S/o. S.C.Chandı	4	RUDRAPRAKASH	['AALPRXXXXJ']	['319X XXXX X220']			
186	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
188	/Users/shash	Smt. KUSUMA.D W/o. Manjunatha He	4	KUSUMA.D	['GIDPKXXXXQ']	['355X XXXX X500']			
309	/Users/shash	Sri. M. ANANTHAMURTHY (PAN No. AS	4	M. ANANTHAMURTHY	['ASYPAXXXXK']	['877X XXXX X416']			
346	/Users/shash	S/o.Sri.Gangadhar Dudhale, aged abo	4						
349	/Users/shash	Sri. M.ANANTHAMURTHY (PAN No. AS	4	M.ANANTHAMURTHY	['ASYPAXXXXK']	['877X XXXX X416']			
427	/Users/shash	SRI. KISHORE KUMAR.D S/o. Sri. Deva	4	KISHORE KUMAR.D	['AHNPDXXXXP']	['607X XXXX X216']			
461	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
463	/Users/shash	SRI. CHANDRASHEKAR.K (PAN No. AD	4	CHANDRASHEKAR.K	['ADXPCXXXXH']	['415X XXXX X025']			
508	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
510	/Users/shash	SRI. LOHITH KUMAR S (PAN No. DOQF	4	LOHITH KUMAR S	['DOQPSXXXXG']	['618X XXXX X895']			
555	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
557	/Users/shash	SMT. L.T. SHILPA (PAN No. BWEPS275	4	L.T. SHILPA	['BWEPSXXXXL']	['233X XXXX X437']			
603	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
605	/Users/shash	SRI.S.G.KRISHNA MURTHY (PAN No. A	4	S.G.KRISHNA MURTHY	['ARQPSXXXXP']	['911X XXXX X694']			
652	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
654	/Users/shash	SRI. VINOD KUMAR G (PAN No. ANYPV	4						
655	/Users/shash	SMT. M ANUSHREE (PAN No. BQRPA8	4	M ANUSHREE	['BQRPAXXXXL']	['720X XXXX X808']			
702	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
704	/Users/shash	SRI. VINOD KUMAR G (PAN No. ANYPV	4						
705	/Users/shash	SMT. M ANUSHREE (PAN No. BQRPA8	4	M ANUSHREE	['BQRPAXXXXL']	['720X XXXX X808']			
749	/Users/shash	Sri. SHIVSHANKAR GANGADHAR DUD	4				SHIVSHANKAR (	['AISPDXXXXG']	['349X XXXX X818']
751	/Users/shash	SRI. LOKESHA M (PAN No. AHAPL4126	4	LOKESHA M	['AHAPLXXXXH']	['678X XXXX X763']			

TRAINING KNN MODEL BASED ON K-MEANS OUTPUTS

Since we have used a dataset that is both private and not an established dataset, there are no standard testing techniques that can be applied on the results. So, in this research we have tested the clustering by training a KNN model based on the paragraph embeddings and the obtained K-means labels. The KNN model classifies the embeddings based on the differences in the K-means labels.

	id	para_text	para_embedding	filename	k_means_labels	kneigh_output	
0	0af78b	ABSOLUTE SALE DEED SITE NO : 38	[-0.00454662 -0.01964579 -0.01830843 0.00396701	/Users/shash	5	5	TRUE
1	79505:	This Deed of Sale of the Scheduled property is made on this 15th day o	[-6.57165572e-02 -7.12047964e-02 1.25249820e-02	/Users/shash	1	1	TRUE
2	bc17d	Sri. SHIVSHANKAR GANGADHAR DUDHALE ALIAS SHIVSHANKAR DUDI	[-0.04715949 0.06256256 -0.04148636 -0.05242401 -	/Users/shash	3	3	TRUE
3	288f4c	AND	[-1.41516859e-02 6.94209593e-04 5.53115122e-02	/Users/shash	3	3	TRUE
4	94570:	SRI. SRINIVASA. H.V, S/o. Venkataramana Shetty (PAN No. ACRPH368:	[-3.96114178e-02 3.01767532e-02 -2.94073410e-02 -	/Users/shash	3	3	TRUE
5	7c445(	WHEREAS, M/s.Janani Developers and Builders represented by its Parti	[0.00057216 -0.03169177 0.00496114 0.01897375	/Users/shash	2	2	TRUE
6	06208	WHEREAS, the Vendor along with M/s. Janani Developers and Builders	[-1.08633954e-02 3.81260812e-02 1.05968462e-02 -	/Users/shash	3	3	TRUE
7	ce084	have obtained sanction for approval of layout plan in the meeting held	[0.04819353 -0.05979848 -0.02082599 -0.01462638	/Users/shash	2	2	TRUE
8	1b752	WHEREAS,Sri.Shivshankar Gangadhar Dudhale alias Shivshankar Dudl	[2.82057803e-02 -1.56345777e-02 -9.37668083e-04 -	/Users/shash	2	2	TRUE
9	a9c5d(	And the khata of the schedule property bearing Site No. 38 registered i	[5.70971072e-02 -1.72874369e-02 -1.08917430e-02 -	/Users/shash	2	2	TRUE
10	e319fe	Whereas both parties have broadly negotiated the terms and condition	[3.11471485e-02 -8.09351653e-02 4.24961969e-02 -	/Users/shash	4	4	TRUE
11	c624e7	NOW THEREFORE THIS MEMORANDUM OF UNDERSTANDING WITNES	[5.57819232e-02 7.09354281e-02 6.76150844e-02 -	/Users/shash	3	3	TRUE
12	89fe3e	1.The Vendor has offered to sell the Schedule Property to the Purchase	[0.02985247 -0.08037467 0.05027387 0.04141634	/Users/shash	4	4	TRUE
13	8095d	2.It was mutually agreed that the sale consideration paid by the purch	[-0.04716174 -0.07005536 0.03025175 0.06936544	/Users/shash	3	3	TRUE
14	af5926	a.A Sum of Rs.7,00,000/- (Rupees Seven Lakh Only) received by way o	[-1.7208535e-02 -9.4100581e-03 -4.9518000e-02 -5.8682691e-	/Users/shash	3	3	TRUE
15	9c6e0	b. The purchaser has availed a loan facility from ICICI Bank of Rs. 21,0	[-5.49808815e-02 -6.60994127e-02 1.51910575e-03	/Users/shash	3	3	TRUE
16	bce07	3.The Vendor has today delivered vacant peaceful possession of the Sc	[2.57883444e-02 -4.81919684e-02 4.10872661e-02	/Users/shash	4	4	TRUE
17	389f42	4.The Vendor represents and assures the Purchaser that he has a clear	[4.56525199e-02 -7.60463700e-02 3.63529362e-02	/Users/shash	4	4	TRUE
18	94c92f	5. The Vendor represents and assures the Purchaser that the schedule	[0.012776 -0.07368623 -0.00196322 0.05800066 0.07684522 -	/Users/shash	4	4	TRUE
19	98b7d	The Vendor represents and assures the Purchaser that the Schedule Pr	[-0.02290511 -0.07828131 0.04100724 0.07158407	/Users/shash	1	1	TRUE
20	aab05(	The Vendor represents and assures the Purchaser that in regard to the	[0.05043019 -0.04909044 0.03455427 -0.00309411	/Users/shash	4	4	TRUE
21	1e213(	The Vendor represents and assures the Purchaser that in regard to the	[7.16565847e-02 -1.25931986e-02 -1.61267184e-02	/Users/shash	4	4	TRUE
22	fcca9c	execution of this Sale Deed, the vendor has no objection for the Purch	[0.01194699 -0.06641395 0.06066549 0.01421271	/Users/shash	1	4	FALSE
23	994a0(	The Vendor represents and assures the Purchaser that he has not ente	[3.54715027e-02 -4.37837951e-02 2.80000847e-02	/Users/shash	4	4	TRUE
24	3c7c3b	The Vendor has no objection for the said transfer of Khata, apart from t	[0.05764056 0.03822595 -0.02795313 0.01504602	/Users/shash	0	0	TRUE

RESULTS

The results obtained are compared with the actual K-means labelling and an approximate accuracy score is calculated manually. This is done by calculating the number of rows with data and then calculating the percentage of accurate results obtained.

RESULT	98.519

THANK YOU