USING MASS MEDIA DATA TO ANALYSE THE GROWTH IN VARIOUS INDIAN DISTRICTS

Konark Verma 2018MCS2025
Kumari Rekha 2018MCS2144

Prof. Aaditeshwar Seth

MOTIVATION and PROBLEM STATEMENT

- A lot of research has been going around the world where people have been trying to analyse as well as predict the growth of various geographical locations using various kinds of data such as: Satellite Imagery data, Census data, Wikipedia data, and so on..

- Surprisingly, there hasn't been a lot of attention paid to mass media data, i.e. day-to-day news. Now, we know that there are a lot of news articles that are being written everyday and these could help us provide various insights in how different districts have grown through the time, and what kinds of news topics are much more commonly spoken about.

- So our goal is to use this particular Mass Media data to analyse the growth of various Indian districts.

A BRIEF ABOUT DATA

- For our problem we have used a corpus of news articles which had more than 5M news articles.
- A classification of districts based on the employment : Unemp districts, Agri districts and Non-Agri districts.
- Another classification of districts based on their pace of growth: Slow, Average and Fast growing districts.

UNDERSTANDING THE DATA

- This table represents the number of districts for employment type vs their pace of growth.
- e.g. The value 5 represents the Fast growing, Non-Agri districts.

- Here, for pace of growth we have used the 2019 prediction from ADI data.
- ADI Predicted values: 0-1 = Slow, 2 = Average, 3-4 = Fast

Number of Districts

	SLOW	AVG	FAST
UNEMP	124	75	29
AGRI	145	72	12
N-AGRI	99	32	5

This column represents the total number of unique articles for each collection.

These values are also represented in this table, as indicated by the arrow, for better understanding.

Number of Articles for each subclass in each collection

COLLECTIONS	TOTAL NUMBER OF ARTICLES	UNEMP DISTRICTS			AGRI DISTRICTS			NON AGRI DISTRICTS		
		Slow (124)	Avg (75)	Fast (29)	Slow (145)	Avg (72)	Fast (12)	Slow (99)	Avg (32)	Fast (5)
AGRICULTURE	80221	7506	5155	4884	14513	5429	735	53840	8955	443
DEVELOPMENT	15984	1426	848	1034	2077	642	124	11936	1571	137
ENVIRONMENT	100038	8720	5704	4711	12141	3885	650	75165	9444	1261
INDUSTRIALIZATION	111291	6466	4017	3898	8379	3002	325	94602	9001	1126
LIFESTYLE	234158	18829	12904	15173	19616	6198	925	187286	19622	2017

Note: Here the sum of articles in each subclass of a collection is more than the total number of articles, because an article can have multiple locations.

Here, 9001 means that there are 9001 number of articles for Average growing Non Agri Districts, for the industrialization collection.

APPROACH USING LDA

Dataset Preprocessing
(Stopword removal, Stemming, Blinding entities)

↓

Compute Coherence Value
For values between 5 to 25, we have computed the coherence value, to find the appropriate number of topics.

↓

Train the LDA Model using this value and obtain the clusters

↓

Obatined the seed keywords
From each cluster we obtained the top 50 keywords and then manually filtered out the keywords.

↓

Recluster the corpus with seed keywords

↓

Collected a set of 100 articles

Issues with this approach

1. During selection of no. of topics, because of the variation in the size of data collections, no. of topics varies more and hence comparison between two collections dont give explainable pattern difference.

2. During seed keyword selection, we are using top 50 keywords we got from first layer clustering. Because of limiting this to top 50 there are many overlapping keywords among topics, which creates overlapping cluster as well.

3. We are using ranking method as 100/n where n is the no. of topics we get. Because of variation in the size of collections, for one collection it gives more diverse articles and for other less diverse.

4. Finally we manually compare its results with the results of other experiments and found out that other experiments yielded better results.

CREATING THE DATASETS

Collections:

- Agriculture
- Development
- Environment
- Industrialization
- Lifestyle

In these Collections, for each article we have:

- Article_Id
- Article_Title
- Article_Text
- Article_Date
- Location_names

Mapping of 2011 districts to 2001 districts
(640 to 593 districts)

Mapping of Location Names to District Ids

Mapping of District Ids to their corresponding Employment Labels

Pace of Growth:
2019 Change Predictions based on ADI
0-1 : Slow, 2 : Average, 3-4 : Fast

Mapping of District Ids to their corresponding Industry Type

Dataset for each collection:

For each collection we form the following dataset and it contains the following fields:

1. **Article_Id**
2. **Article_Title**
3. **Article_Text**
4. **Processed Text ***
5. **Article_Date**
6. **District_Id**
7. **Employment Type**
8. **Pace of growth**
9. **Industry Type**

Now in our dataset, Article_Id and District_Id together can form the primary key.

* **Note**: To obtain the Processed Text, we have concatenated the article title and article text, and then applied the **stop-word removal, entity blinding,** and **stemming**.

TAGGED DOCUMENTS

For each dataset corresponding to each collection we do the following:

Dataset

Tagged Documents

DATA	TAGS
Processed Text	Article_Id District_Id Employment Type Pace of growth

Here, for each article we provide the processed text as data, and Article_Id, District_Id, Employment Type, Pace of growth as tags.

DOC-TAG-2-VEC MODEL

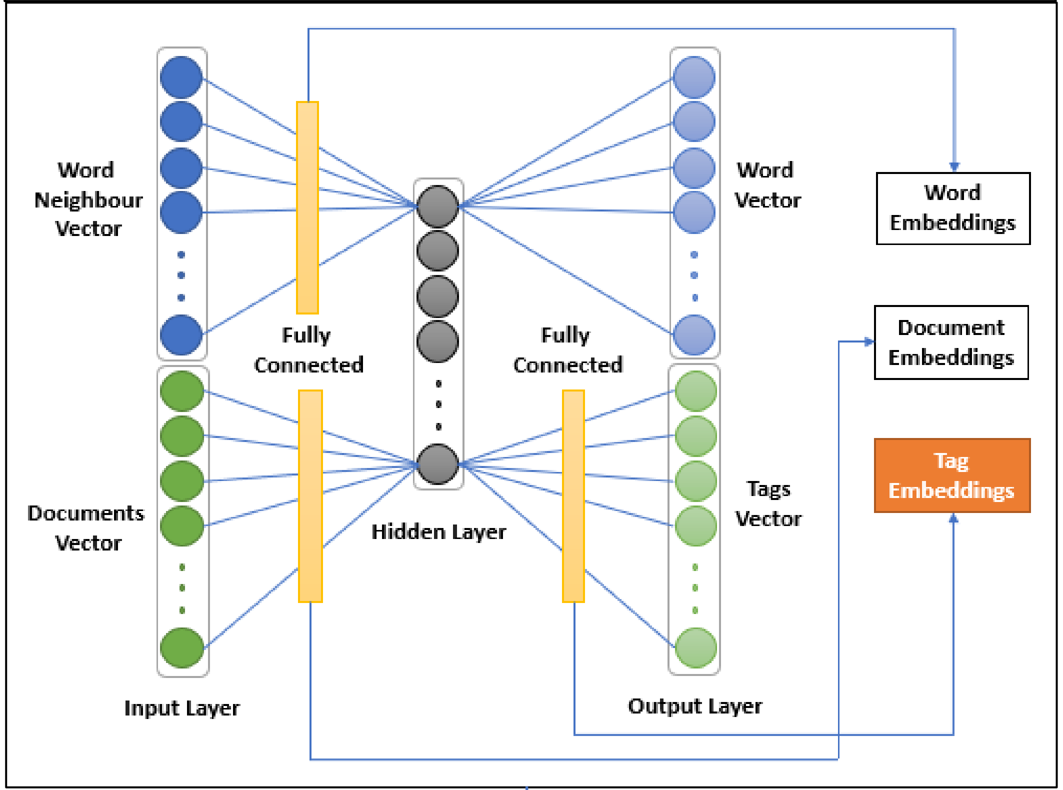
Note: At this point we have both datasets, as well as Tagged documents. Tagged Documents are used to train the DT2V Models, while datasets are used for further analysis.

TRAINING DOC-TAG-2-VEC MODEL

For each tagged document corresponding to each collection we do the following:

Tagged Documents

DOC-TAG-2-VEC MODEL



Vector embeddings for each document and tags

Some points to note:

1. **Parameters** for training the model:
   - Epochs: 100
   - Learning rate: 0.1
   - Minimum learning rate: 0.001
   - Vector size: 50Min occurrences for a word: 3
   - Distributed bag of words (D-BOW)
   - Window size: 3
2. In our model, the learning rate linearly falls down till the value of min-learning-rate.
3. Each article is considered as a separate document, along with appropriate tags.
4. Tags used: Article Id, District Id, Employment type, Pace of growth.
5. Here, we are getting the both the direct document embeddings as well as tag embeddings for each document. We are ignoring the direct document embeddings and are using the **tag embeddings** for each document.

OBTAINING THE MOST EXPLAINABLE ARTICLES

For each set of document vector embeddings corresponding to their collection we do the following:

Vector embeddings for all documents

Divide these to their corresponding subclasses

Subclass: fast-growing-agri, slow-growing-non-agri, etc..
So, for each collection we divide their vector embeddings into the 9 subclasses.

Cluster the documents of each subclass

We use **hierarchical clustering**, with **Dunn index** to obtain clusters with **min-intra-cluster distance** and **max-inter-cluster distance**. The number of clusters formed are between 50 and 80.

Obtain the global centroid for each subclass

We find the centroid using the **median** of all the vector representations of documents, for each subclass, for each collection.
Here, the reason we choose median is to get rid of the outliers.

Obtain the centroid for all clusters of each subclass

We find the centroid of each cluster using the **mean** of all the vector representations of documents corresponding to that cluster, for each subclass, for each collection.

Pick the top 25 clusters for each subclass

For each subclass of each collection we choose those 25 clusters whose **cluster centroid is closest to the global centroid** of corresponding subclass.
Reason for picking only 25 clusters is to avoid the outliers, as hierarchical clustering can generate few errors.

Collect the Article_Id of 100 documents for each subclass

From each cluster obtain the 4 documents that are closest to their corresponding cluster centroid. Thus collecting the 100 documents for each subclass, for each collection.

Now, we use our dataset to obtain title and processed text for these 100 documents, for each subclass, for each collection.

MOST EXPLAINABLE ARTICLES

Finally, we have **Article Id, Article Title, Processed Text, Global-Centroid** for the top 100 documents for each subclass, for each collection which will be used for further analysis.

Note: Here to compare the vector representations we use cosine similarity. So the word closest means the max cosine similarity.

ANALYSIS

For each collection we do the following:

OBTAIN THE MOST EXPLAINABLE ARTICLES

Now, we have **Article Id, Article Title, Processed Text, Global-Centroid** for the top 100 documents for each subclass, of each collection.

Keywords Extraction

We find out the **100 most explainable keywords** using **TF-IDF** on the **Processed Text**, and do this for all the subclasses of all the collections.

Similarity Matrix

We use cosine-similarity to obtain the similarity between the global centroids of all the subclasses, and form a 9x9 matrix. We do this for all the collections.

A global centroid of a subclass can be thought as a point in 50 dimensional vector space, which represents, that particular subclass.

Analysis using keywords

We use these keywords belonging to different subclasses to do the following:

1. Obtain those words that occur for all the subclasses of a collection.
2. Obtain those words that are unique to a particular subclass(i.e. the words that occur in a particular subclass but not in any other subclass).
3. For a pattern (subclass-A vs subclass-B) we find out those words that occur for both these subclasses (i.e. both A,B).
4. For a pattern (subclass-A vs subclass-B) we find out those words that occur only for subclass A, and only for subclass B.

We do this for all the collections.

LIST OF PATTERNS

Using the tables shown on the first slide of this presentation, we obtain a list of patterns for which we would like to study or analyze the data.

FINAL RESULTS

ANALYSIS ACROSS PATTERNS

Finally we perform our analysis, using the similarity matrix, and keywords results, across all the patterns and for all the collections.

ANALYSIS

Initial List of Patterns:

1. Unemployment districts that are slow growing
   (vs) Unemployment districts that are fast growing.

2. Agriculture districts that are slow growing
   (vs) Agriculture districts that are fast growing.

3. Non Agri districts that are slow growing
   (vs) Non Agri districts that are fast growing.

4. Unemployment districts that are slow growing
   (vs) Agriculture districts that are fast growing.

5. Agriculture districts that are slow growing
   (vs) Non-Agri districts that are fast growing.

6. Unemployment districts that are slow growing
   (vs) Non-Agri districts that are fast growing.

Some general analysis using keywords:

- For **Agriculture collection**, the following words**: 'monsoon', 'rain', 'water', 'farmer', 'village', 'land', 'rainfall', 'crop'** do come up on top **for all the subclasses**. On reading some of the articles manually, we see that there are a lot of talks regarding the weather, crops and farmers.

- For **Development collection**, the following words**: 'project', 'scheme', 'develop', 'central', 'program'** do come up on top **for all the subclasses**. On reading some of the articles manually, we see that there are a lot of articles about the central government policies, schemes and projects.

- For **Industrialization collection**, the following words**: 'project', 'develop', 'power'** do come up on top **for all the subclasses**. On reading some of the articles manually, we see that there are a quite some articles about the electricity, and ways of generating electricity from different sources.

- For **Lifestyle collection**, the following words**: 'health', 'hospital', 'policies'** do come up on top **for all the subclasses**. On reading some of the articles manually, we see that there are many articles about the hospitals and human-health.

PATTERN-2: SLOW GROWING AGRI DISTRICTS vs FAST GROWING AGRI DISTRICTS

COLLECTION	GLOBAL CENTROID SIMILARITY	KEYWORDS THAT OCCUR IN BOTH SLOW-GROWING AND FAST-GROWING AGRI DISTRICTS	KEYWORDS THAT OCCUR MUCH MORE OFTEN IN SLOW-GROWING AGRI DISTRICTS	KEYWORDS THAT OCCUR MUCH MORE OFTEN IN FAST-GROWING AGRI DISTRICTS	COMMENTS
AGRICULTURE	0.35	loan, rainfall, agriculture, rain, farmer, protest, monsoon, water, loan, crop, farm, village, demand	system, variety, suicide, plant, mani, irrigation, grape, dam, wheat, hectare, hailstorm, season, flood, cultivate, paddy	forecast, field, problem, power, consumption, tariff, protest, waiver, electricity, growth, corrupt, sector, develop	• In both subclasses, there are a lot of talk about rainfall, climate changes, farmer protests, bank loans. • In slow-growing Agri districts there are much more articles about cultivation of crops, floods, irrigation projects. • In fast-growing Agri districts there are more articles about loan and electricity waivers.
DEVELOPMENT	0.22	air, scheme, connect, market, route, airport, train, skill, employ, invest, industry, project, flight, technology, develop	msme, assist, worker, manage, power, farmer, agriculture, irrigation, land, crop, cultivate, plant	policy, student, business, fund, service, job, company, investor, heath, infrastructure, digital, data, research	• In both subclasses, there are a lot of talk about schemes regarding airports and technology related developments. • In slow-growing Agri districts, articles are more about agricultural activities and related issues. • In fast-growing Agri districts, articles talk more about health, investments and digital development.

PATTERN-2: SLOW GROWING AGRI DISTRICTS vs FAST GROWING AGRI DISTRICTS

COLLECTION	GLOBAL CENTROID SIMILARITY	KEYWORDS THAT OCCUR IN BOTH SLOW-GROWING AND FAST-GROWING AGRI DISTRICTS	KEYWORDS THAT OCCUR MUCH MORE OFTEN IN SLOW-GROWING AGRI DISTRICTS	KEYWORDS THAT OCCUR MUCH MORE OFTEN IN FAST-GROWING AGRI DISTRICTS	COMMENTS
ENVIRONMENT	0.51	air, system, student, school, manage, river, waste, land, court, pollution, environment, industry, plant, pollution, develop	medic, tree, survey, agriculture, farmer, conserve, activist, water, village, drain, tourism, forest, proposal	illegal, consult, suicide, construct, poor, electricity, vehicle, clearance, public, education, transport	• In both subclasses, there are a lot of talk about pollution, waste and other development activities. • In slow-growing Agri districts there are much more articles about agricultural activities and forest conservation. • In fast-growing Agri districts there are more articles about forest clearances and transportation.
INDUSTRIALIZATION	0.57	manage, power, real, estate, market, business, opportunity, invest, industry, technology, develop	passenger, air, survey, river, hike, waste, land, grass, pollution, water, city, village, demand, steel, sanitisation	policy, term, fuel, airline, domestic, investor, money, spectrum, bank, growth, firm, global	• In both subclasses, there are a lot of talk about business and industry development. • In slow-growing Agri districts there are much more articles about pollution and waste related problems. • In fast-growing Agri districts there are more articles about investment and monetary policies.
LIFESTYLE	0.53	medic, air, college, service, pollution, tourist, water, environment, develop, health	worker, care, problem, student, traffic, agriculture, vehicle, organisation, business, doctor, festival	theatre, show, cbi, culture, market, travel, violence, protest, social, film, industry, invest, terror, crime	• In both subclasses, there are a lot of talk about protests, tourism, hospital services. • In slow-growing Agri districts there are much more articles about festivals, facilities, policies. • In fast-growing Agri districts there are more articles about violence, protests, crimes.

PATTERN-2: SLOW GROWING AGRI DISTRICTS vs FAST GROWING AGRI DISTRICTS

EXAMPLES OF TITLES OF ARTICLES FOR BOTH SUBCLASSES

COLLECTION	SLOW GROWING AGRI	FAST GROWING AGRI
AGRICULTURE	• Farmers wait for subsidised farm equipment • Mallu promises relief to flood-hit farmers of Madhira • Raichur to get irrigation water after August 5 • Plan alternative crops: Collector • Central team assesses drought in Arsikere	• Farmers with land along streams lost almost all the crop • Rain fails to push up dam levels • Agri staff stop work to protest corruption • High-level committee assesses crop damage in Ariyalur district • Monsoon ends with 12% shortfall
DEVELOPMENT	• 2 lakh labourers to get employment under NREGS • Drought-hit farmers pour out their woes • Agricultural implements donated • ‘Tribal areas are well connected with NREGS' • ‘Focus on development works in grama sabha meetings'	• UDAN’s first flight: City connects with Porbander • Maharashtra nod for regional plan in eight districts • Forest Department launches afforestation drive in Ariyalur • Yogi govt plans airport terminal in Chitrakoot • Farmers pine hopes on north east monsoon
ENVIRONMENT	• Elephants kill two in Davangere district • Tribal people resume struggle for land • Farmers seek an end to monkey raids • Crop-raiding tusker captured in Dharmapuri • Naxal gunned down in Sukma weapons recovered	• Pranhita sanctuary only on paper • Tiger kills 1 in Gadchiroli • Forest department: 89% of 10-year-old plantations are unsuccessful • Mulak takes up cudgels for wildlife • Forest department evaluation confirms plantation scams
INDUSTRIALIZATION	• Ponnam dares KCR to inspect Karimnagar town • Rural sanitation takes a beating • Scouts take out cycle rally • Power crisis to get worse as coal stocks plunge • SC notice to Rajasthan on illegal mining	• Chaos at environmental hearing for JSW project • Illegal sand mining continues unabated • No visible improvement in road infrastructure in Erode • Signals from Pak mobile companies reach bordering areas • Drought hits freight movement
LIFESTYLE	• Flamingo festival in tourism calendar • Bandh shuts down the city • Fervour marks Lord Ganesha immersion • Out of forests and into mines • Universal Health Coverage programme inaugurated	• Second phase of pulse polio campaign tomorrow • Rs 80 lakh public wealth destroyed in public violence • Govt zeros in on 6 Red-hit districts to tackle Maoists • Chela gets award along with guru • Clashes tensions during Holi relieves in Rajasthan

PATTERN-5: SLOW GROWING AGRI DISTRICTS vs FAST GROWING NON-AGRI DISTRICTS

COLLECTION	GLOBAL CENTROID SIMILARITY	KEYWORDS THAT OCCUR IN BOTH SLOW-GROWING AGRI AND FAST-GROWING NON-AGRI DISTRICTS	KEYWORDS THAT OCCUR MUCH MORE OFTEN IN SLOW-GROWING AGRI DISTRICTS	KEYWORDS THAT OCCUR MUCH MORE OFTEN IN FAST-GROWING NON-AGRI DISTRICTS	COMMENTS
AGRICULTURE	-0.34	loan, rainfall, suicide, agriculture, dam, rain, farmer, monsoon, crop, wheat, price, flood, forest, farm, water, bank, loss, weather, harvest, village, demand	supply, variety, pond, scheme, market, stock, grape, shower, hailstorm, season, strike, winery, sow, horticulture, tanker, temperature, onion,	kill, field, sanction, death, power, college, river, gang, construct, relief, hospital, protest, court, attack, rape, expressway, life, police, tiger, product, claim, cabinet, develop	• In both subclasses, there are a lot of talk about rainfall, floods and farmer suicides. • In slow-growing Agri districts there are much more articles about different cultivation practices, schemes to resolve farmers issues, strikes of farmers, situation of market, effect of temperature. • In fast-growing non-agri districts there are much more articles about sanction of funds, power, construction, hospital, field of crops.
DEVELOPMENT	-0.21	worker, scheme, construct, connect, airport, farmer, tribal, highway, train, land, employ, women, project, implement, water, program, village, develop	air, power, college, estate, agriculture, prison, market, subsidy, skill, crop, enterprise, textile, bird, institute, youth, invest, horticulture, industry, flight, park, smart, entrepreneurship, storage, city, plant, technology, entrepreneur	complaint, school, sanction, student, die, girl, fund, yojana, job, court, road, bridge, mnrega, maoist, child, commission, police, beti, wage, pmgsy, campaign, engine, household, law, bank, monitor, labour, health, demand	• In both subclasses, there are a lot of talk about schemes regarding village development schemes and technology related developments. • In slow-growing Agri districts, articles are more about horticulture, production of crops, implementation of technology, skills of farmers, facilities to farmers. • In fast-growing non-agri districts, articles talk more about students , IT jobs, air pollution, bank and health facilities,

PATTERN-5: SLOW GROWING AGRI DISTRICTS vs FAST GROWING NON-AGRI DISTRICTS

COLLECTION	GLOBAL CENTROID SIMILARITY	KEYWORDS THAT OCCUR IN BOTH SLOW-GROWING AGRI AND FAST-GROWING NON-AGRI DISTRICTS	KEYWORDS THAT OCCUR MUCH MORE OFTEN IN SLOW-GROWING AGRI DISTRICTS	KEYWORDS THAT OCCUR MUCH MORE OFTEN IN FAST-GROWING NON-AGRI DISTRICTS	COMMENTS
ENVIRONMENT	-0.29	river, farmer, land, court, bird, area, forest, police, project, water, tourism, reside, village, develop	medic, air, plastic, hospital, conserve, quality, activist, mayor, clean, noise, plant, green, municipal, waste, commission, industry, drain, school, student, NGO, treatment, puja, dispose, cover, swachh	fire, kill, death, poach, electricity, blackbuck, relief, crop, maoist, arrest, guard, tourist, rain, camp, recover, protest, flood, attack, tiger, sanctuary, college, crocodile, poacher, wild, park, herd, tribal, wildlife	• In both subclasses, there are a lot of talk about forest clearance wildlifes and tourism. • In slow-growing Agri districts there are much more articles about programms for wildlifes, Forest clearance, plantation, industries, • In fast-growing non-agri districts there are more articles about relief factors, wildlife, mining, tourism, agriculture fire.
INDUSTRIALIZATION	0.01	school, student, power, river, estate, construct, well, toilet, train, land, company, sanity, mine, water, reside, village, plant, develop	air, survey, connect, hike, shop, market, airport, waste, pollute, tax, textile, garbage, institute, children, commission, invest, industry, product, engine, manufacture, wine, technology, demand, steel	illegal, fire, worker, kill, vehicle, rain, tribal, protest, murder, rescue, raid, flood, weaver, arrest, forest, dengue, attack, rape, agency, police, victim, rebel, coal	• In both subclasses, there are a lot of talk about illegal mining and vllagers issues • In slow-growing Agri districts there are much more articles about industry, textile, studies, pollution, investment. • In fast-growing Non-Agri districts there are more articles about forest, illegal mining , rape, cop duties, temperature.
LIFESTYLE	-0.12	medic, vehicle, service, hospital, doctor, flood, police, patient, water, village, health, develop	air, kumbh, farmer, pollute, children, tourist, worker, problem, traffic, agriculture, dam, institute, research, disease, product, college, toilet, municipal, mela, women, tourism, screen, program, school, student, drive, fever, treatment, train, irctc	fire, kill, death, die, violence, land, elect, strike, arrest, power, poll, rain, highway, protest, spot, law, force, reside, murder, rape, vote, dispute, secure, curfew, court, crime, youth, victim, army, life	• In both subclasses, there are a lot of talk about health awareness and crimes like rape, violence. • In slow-growing Agri districts there are much more articles about pollution, tourism, woman and children, treatment, health program, dam. • In fast-growing Non-Agri districts there are more articles about strikes, minorities success, murder, sexual harassment, raids.

PATTERN-5: SLOW GROWING AGRI DISTRICTS vs FAST GROWING NON-AGRI DISTRICTS

EXAMPLES OF TITLES OF ARTICLES FOR BOTH SUBCLASSES

COLLECTION	SLOW GROWING AGRI	FAST GROWING NON-AGRI
AGRICULTURE	• ‘Assess crop loss, provide compensation to farmers’ • Plan alternative crops: Collector • BJP Kisan Morcha stages dharna in front of Collectorate • Paddy purchase going on smoothly • Dry spell in Andhra forces tenant farmers to take a summer break	• Load rejig in peak hours may ease power cuts • Haryana sanctions DIF of over Rs. 1.93 crore • Haryana clears Rs. 255 crore road-widening project in NCR • Health department fighting dengue with dud gun • 10L acre wheat damaged in Haryana
DEVELOPMENT	• State increases target for horticulture crops • 1, 500 ha of jowar to be produced for Anna Bhagya scheme • Biometric machine installation at govt offices at snail’s pace • Farmers grow green fodder to tackle shortage • Five lakh ‘Agathi’ seedlings to be distributed to farmers	• New challenges in IT job prospects • Students baffled by CBSE results delay • Nod to new border road agency • World Health Organisation to South-East Asian countries: Accelerate efforts to address air pollution • World Bank scouts for innovative social projects
ENVIRONMENT	• AForest clearance for Palamuru-RR LI • Non-teak species to get greater priority in growing plantations from next year • Govt acts tough with tendu contractors • Uma allays ryots’ fear over Pattiseema project • Social forestry wing to help farmers take commercial route	• Punjab, Haryana told to check agricultural fires • DMRC launches e-rickshaw service in Ghaziabad • Manesar’s wild side: A peek at a leopard family in Aravali forest • SC talks tough on illegal Aravali mines • Tourist hotspot on the anvil
INDUSTRIALIZATION	• Industry prepared to pay more for power • WB studies biometric, Aadhaar-enabled services • Textile parks will be established in all taluks, says Anjaneya • Civic workers' strike raises a stink • Chief Minister invites U.S. investment in aerospace sector	• With summer on, demand for gensets goes up in city • Unitech promoters sent to five-day police custody for FD scheme probe • Bus rape spooks working women • Illegal mining ruined Aravalis in Haryana, Rajasthan' • Aravalis a forest? Survey to decide
LIFESTYLE	• Telangana to get tourism boost with heliports • Woman, infant die after nurse botches up delivery • State plans two tertiary centres for cancer treatment • Universal Health Coverage programme inaugurated • Bandh shuts down City	• Trade unions' strike paralyzes region • Free coaching for minorities and women at Jamia • Raids on to nab 3 key members of Kaushal gang • 7 more arrested for murder of Congress leader Vikas Chaudhary • Two girls questioned: ‘Knew about Crazy Sumit video, didn’t know he’ll upload it’

DISTRICT LEVEL ANALYSIS USING TAGS EMBEDDINGS

OBTAIN THE DISTRICT VECTORS

First we obtain the district vectors from tag embeddings of the models that we had trained. We do this for all the collections.

250-dimensional representation for each district

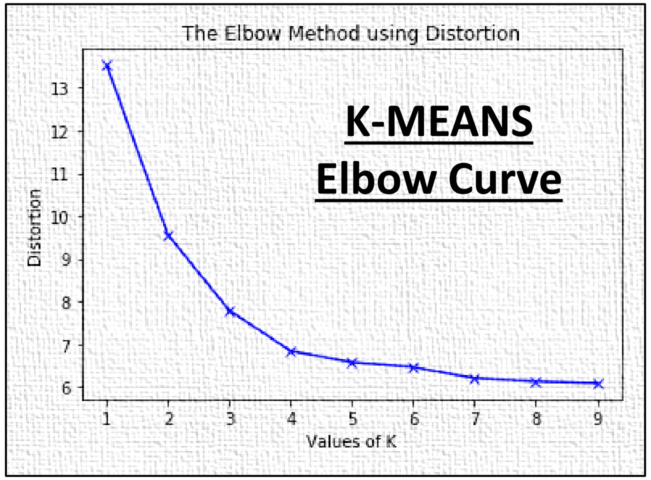
Each district, for a collection has a 50 dimensional tag embedding. Since we have 5 different collections, so by concatenating all these vectors we form a 250 dimensional vector.

Data Visualization using t-SNE

We then use t-SNE to convert this 250 dimensional vector to a 2-dimensional vector, for data visualization.

Clustering using K-Means

We use K-Means to cluster these district vectors. First, we start by trying different values for k, and plotting the elbow curve.



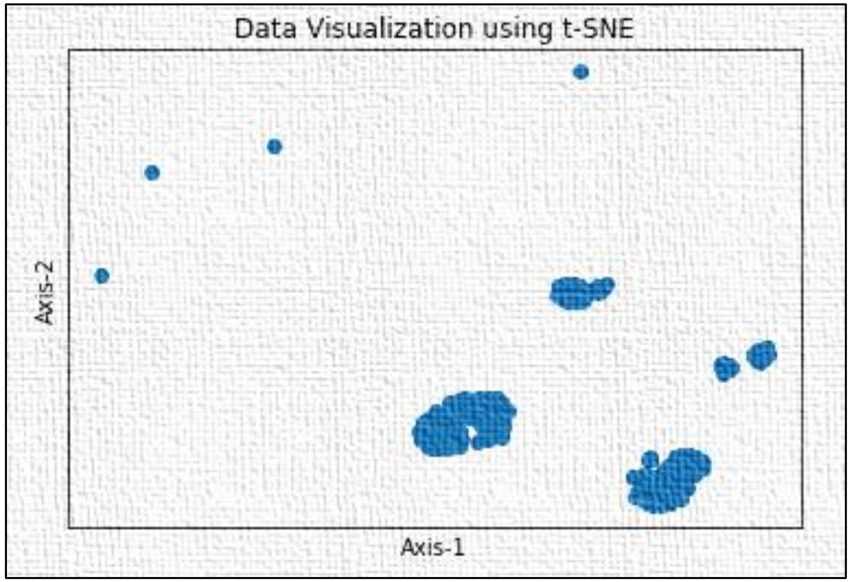
Using this elbow curve we choose the value for k as 4.

The clusters formed have the following number of districts in them:

Clusters-1 : 62
Clusters-2 : 82
Clusters-3 : 216
Clusters-4 : 176

T-SNE Plot

This plot shows us a 2-dim representation of districts



CONCLUSIONS

- We built a system which takes a collection of articles as input and filters out the most explainable articles from those.
- We then came up with a list of interesting patterns and used these to compare how different subclasses of district are similar as well as dissimilar to each other.
- We also generated the explainable keywords that could provide us better insights for this task.

FUTURE SCOPE

- We can add more data in the future and see how our system performs.
- We can further do time based analysis of subclasses to see how these have changed through the time.

THANK YOU

Konark Verma 2018MCS2025
Kumari Rekha 2018MCS2144

Prof. Aaditeshwar Seth