Srividya Majeti

Assignment 9

CS 532: Introduction to Web Science
Dr. Michael Nelson
Spring 2016

April 21, 2016

Contents

1

Question 1

Choose a blog or a newsfeed (or something similar with an Atom or RSS feed). Every student should do a unique feed, so please "claim" the feed on the class email list (first come, first served). It should be on a topic or topics of which you are qualified to provide classification training data. Find something with at least 100 entries (or items if RSS).

Create between four and eight different categories for the entries in the feed:

examples:

work, class, family, news, deals

liberal, conservative, moderate, libertarian

sports, local, financial, national, international, entertainment

metal, electronic, ambient, folk, hip-hop, pop

Download and process the pages of the feed as per the week 12 class slides.

Be sure to upload the raw data (Atom or RSS) to your github account.

Following are the steps I have taken to the solve the problem:

- The most hardest part in the assignment is to find the blog with atleast 100 entries. I tried searching different blogs of my favorite cricketer, actor, restaurants etc. But I could find a blog. I tried even paginating the blogs but they did not have more than 100 Entries.
- Finally I started searching newsfeed and I found an entertainment feed in 'Times of India'.
- It has about 172 items. I worked on 86 items manually and the remaining using the fisher classification.
- Based on their language, I categorized the items in the feed into 12 categories. The categories are as follows: 'Hindi', 'Telugu', 'English', 'Malayalam', 'Kannada', 'Marathi', 'Gujarathi', 'Punjabi', 'Bhojpuri', 'Tamil', 'Bengali' and 'others'.

2

Question 2

Manually classify the first 50 entries, and then classify (using the fisher classifier) the remaining 50 entries.

Create a table with the title, predicted category, actual category, and cprob() and fisherprob() for the actual category.

Following are the steps I have taken to solve the problem:

- In my blog I have 172 entries I classified the first 86 entries manually. To do this classification I used the 'docclass.py' and 'feedfliter.py' from 'PCI' book code 'chapter 3' and the code from 'document-filtering' powerpoint slide 26. These are listed in Listing 2.1 2.2 and 2.4 respectively.
- The output with 'Title' , 'Description', 'Predicted category' and 'Actual category' for each item is illustrated in the below Table 2.1

Table 2.1. Output with 'Title' , 'Description', 'Predicted category' and 'Actual category'

Title	Description	Predicted category	Actual category
Spoiler Alert: Saala Khadoos	http://timesofindia.indiatimes.com/entertainment/hindi/movie-reviews/Saala-Khadoos/movie-review/50775792.cms?tabtype=spoiler	None	Hindi
Pic: Katrina Kaif officially a part of the Kapoor Khaandaan	You know what happened this Christmas, apart from the usual celebrations and merriment? Katrina Kaif and Ranbir Kapoor posed together as a couple for the shutterbugs, something they seldom do.	Hindi	Hindi
'Dilwale' song Premika to release on December 25	If youve seen 'Dilwale,' you are obviously wondering what happened to the much-hyped song, Premika, which was missing from the film.	Hindi	Hindi
Spoiler Alert: Dilwale	http://timesofindia.indiatimes.com/entertainment/hindi/movie-reviews/Dilwale/movie-review/50229987.cms?tabtype=spoiler	Hindi	Hindi
Movie Review: Dilwale	http://timesofindia.indiatimes.com/entertainment/hindi/movie-reviews/Dilwale/movie-review/50229987.cms	Hindi	Hindi
Asha Bhosle not a follower of Radhe Maa	http://timesofindia.indiatimes.com/entertainment/hindi/bollywood/Radhe-Maas-Bollywood-connection/photostory/48585163.cms	Hindi	Hindi
Salman appeal put off for lack of translated documents	The Bombay HC on Wednesday adjourned to July 13 the appeal filed by actor Salman Khan against his conviction for culpable homicide not amounting to murder in the 2002 hit-and-run case in Bandra.	Hindi	Hindi
Spoiler Alert: Me - Climax revealed!	http://timesofindia.indiatimes.com/entertainment/hindi/movie-reviews/Me/movie-review/47811614.cms?tabtype=spoiler	Hindi	Hindi
Forty years on, only a handful of films have investigated the Emergency	Gulzars Aandhi was banned. Film Kissa Kursi Kas print was burnt. Kishore Kumars songs went off the air.	Hindi	Hindi
Kissa Kursi Ka: Government mulls returning prints or compensating filmmaker as kin	Forty years after controversial film, Kissa Kursi Kaa satire on former PM Indira Gandhi and the Emergency a was banned and its prints burnt, the information and broadcasting (I amp B) ministry has said that it is considering an appeal for either returning the prints or compensating the filmmakers son.	Hindi	Hindi
Mithun Chakraborty surrenders Rs 1.2cr received from Saradha to ED	During his last questioning by ED here in May, Chakraborty had provided the agency sleuths with a number of DVDs, CDs and scripts that he had got as part of being the brand ambassador of the Saradha group.	Hindi	Hindi
Sangeeta Bijlani: Sangeeta uncomfortable doing lovemaking-kissing scenes, quits comeback movie	Onirs Shab was being touted as Sangeeta Bijlanis	Hindi	Hindi
Whatever I am today is because of tV: Siddharth Shukla	Actor Siddharth Shukla, who will be making his filmy debut soon, credits his success to TV shows	Hindi	Hindi
Raqesh and Ridhi in Nach Baliye	The celebrity couple will be seen in the sixth season of the dance reality show	Hindi	Hindi

Title	Description	Predicted category	Actual category
'The Revenant' wins big in Britain's BAFTA awards	Survival drama 'The Revenant' was the top winner at Britain's biggest movie awards on Sunday, taking the best film prize and honours for leading actor Leonardo DiCaprio and director Alejandro Gonzalez Inarritu.	Hindi	English
Movie Review: Trumbo	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/Trumbo/movie-review/50961462.cms	English	English
Spoiler Alert: The Boy	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/The-Boy/movie-review/50758961.cms?tabtype=spoiler	English	English
Spoiler Alert: The Danish Girl	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/The-Danish-Girl/movie-review/50578161.cms?tabtype=spoiler	English	English
Spoiler Alert: The Hateful Eight	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/The-Hateful-Eight/movie-review/50579793.cms?tabtype=spoiler	English	English
Spoiler Alert: Our Brand Is Crisis	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/Our-Brand-Is-Crisis/movie-review/50499376.cms?tabtype=spoiler	English	English
Spoiler Alert: Daddy's Home	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/Daddys-Home/movie-review/50499975.cms?tabtype=spoiler	English	English
Spoiler Alert: Point Break	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/Point-Break/movie-review/50391153.cms?tabtype=spoiler	English	English
Spoiler Alert: Star Wars: The Force Awakens	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/Star-Wars-The-Force-Awakens/movie-review/50309447.cms?tabtype=spoiler	English	English
Spoiler Alert: In The Heart Of The Sea	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/In-The-Heart-Of-The-Sea/movie-review/50041881.cms?tabtype=spoiler	English	English
Spoiler Alert: The Hunger Games: Mockingjay - Part 2	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/The-Hunger-Games-Mockingjay-Part-2/movie-review/49933446.cms?tabtype=spoiler	English	English
Celebs who kept 'investigating' eyes on their spouses	Being a celeb means a constant public glare and thus most celeb relations become easy targets of scrutiny.	English	English
The Martian: Climax revealed	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/The-Martian/movie-review/49180156.cms?tabtype=spoiler	English	English
Bobbi Kristina Brown, daughter of Whitney Houston, passes away	The brief, chaotic life of Bobbi Kristina Brown was never really her own.	English	English
Spoiler Alert: Insidious: Chapter - Climax revealed!	http://timesofindia.indiatimes.com/entertainment/english/movie-reviews/Insidious-Chapter-3/movie-review/47816236.cms?tabtype=spoiler	English	English

Title	Description	Predicted category	Actual category
MOVIE REVIEW: 'The Boy Next Door'	http://timesofindia.indiatimes.com/entertainment/english/movie-review/the-boy-next-door/movie-review/46130208.cms	English	English
Big B to promote Shamitabh in Chennai	Dhanush will accompany the legendary actor during his promotions in Chennai this week	English	English
Suriya, in top 50 the sexiest Asian	Suriya is the only Tamil actor in the list of Top 50 Sexiest Asian Men	English	Tamil
Siddharth continues relief work in Cuddalore	After helping with relief work in Chennai, Siddharth has shifted base to Cuddalore	Tamil	Tamil
Kollywood actor Siddharth's home flooded in Chennai	Chennai rains has got lot of commoners into trouble, but even celebrities were hugely affected by the rains. Kollywood and Bollywood actor Siddharth took to Twitter and shared a picture of his flooded residence.	Tamil	Tamil
Actor Vivek loses 13-year-old son to dengue, brain fever	Tamil actor Viveks son Prasanna Kumar, 13, died in a city hospital on Thursday of complications arising out of dengue and brain fever. He was undergoing treatment at the hospital for more than a month.	Hindi	Tamil
What are the collection figures of 10 Endrathukulla and Naanum Rowdy Dhaan?	There are two versions on the collections of Vikrams 10 Endrathukulla	English	Tamil
OMG: Vishal attacked at Nadigar Sangam elections	The actor was attached during the elections held this morning	English	Tamil
Income tax department raids Tamil film actor Vijay's house	Income Tax officials raided the premises of Tamil actor Vijay and producers and director of Tamil movie Puli on Wednesday morning in connection with alleged tax evasion.	Hindi	Tamil
Tamil actor Krishna booked under charges of dowry and harassment	Tamil cinema actor Krishna has been booked under dowry harassment case by all women police station, Thudiyalur here on Thursday.	Tamil	Tamil
Simhaa will be seen in an action avatar in his next	Simhaa will team up with director Vaanan in his next	English	Tamil
Director Saravanan injured in car accident	Engeyum Eppodhum fame director Saravanan was injured in a car accident today.	Tamil	Tamil
Kamal Haasan's 'Indian' sequel on the cards!	Shankar-directed 'Indian' featured Kamal Haasan in the title role.	English	Tamil
Jyotika-Suriya may act together next year	Suriya says he may act alongside his actress-wife Jyotika in a Tamil project soon.	Tamil	Tamil
Kishore roped in for Kamal Haasans next	He has a key role in the movie and will be introduced in the beginning itself.	Tamil	Tamil
Shruti Haasan plays a taxi driver in Thala 56	Shruti Haasan, plays the role of a taxi driver and not Ajith.	Tamil	Tamil
Dhanush's Maari teaser	The team had recently wrapped up the shooting of the film.	Tamil	Tamil
Raghava Lawrence to direct Rajinis next?	The buzz is that the actor might soon rope in the Superstar for his next	Tamil	Tamil

Title	Description	Predicted category	Actual category
62nd National Film Awards: Tamil movies bag eight honours	Several Tamil movies bagged honours at the 62nd National Film Awards announced on Tuesday.	Tamil	Tamil
Madras high court stays release of Vikram-starrer 'I'	I, featuring Vikram and Amy Jackson and directed by Shankar, was scheduled for release on the Pongal day, January 15.	Tamil	Tamil
I came all the way from Hollywood looking for a job in India: Arnold	Hollywood superstar Arnold Schwarzenegger said that he was taken back to his younger days as a bodybuilder at the audio launch of Shankars I.	Tamil	Tamil
I came all the way from Hollywood looking for a job in India: Arnold	Hollywood superstar Arnold Schwarzenegger said that he was taken back to his younger days as a bodybuilder at the audio launch of Shankars I.	Tamil	Tamil
61st Filmfare Awards (South) Tamil winners list 2013	Director Ram's 'Thangameengal' swept the awards at the Filmfare.	Tamil	Tamil
Pawan Kalyan returns as 'Sardaar Gabbar Singh'	Telugu superstar Pawan Kalyan is set to enthrall viewers with Sardaar Gabbar Singh. Helmed by KS Ravindra, the action-packed entertainer will release in Hindi and Telugu next month.	Tamil	Telugu
Tollywood stars send relief supplies to flood-hit Chennai	Ten heroes of Tollywood will be available at three malls in the city on December 6 from 4 pm to 7 pm to collect donations for flood-hit Chennai.	Tamil	Telugu
Actor Mahesh Babu adopts Siddapuram village of backward Mahbubnagar	Tollywoods big hero Mahesh Babu has adopted a Telangana village to participate in its development. He has decided to adopt Siddhapuram village of Kottur mandal in backward Mahbubnagar district.	English	Telugu
Bhale Bhale Magadivoi brought Lavanya	Lavanya Tripathi just cant seem to stop herself from breaking into a guffaw when talking about her upcoming rom-com Bhale Bhale Magadivoi	English	Telugu
Baahubalis song leaked?	A song from the forthcoming film Baahubali, Tollywoods biggest film ever, has reportedly been leaked.	Hindi	Telugu
Only after my grandfather passed away did I realise how much my life revolved around him: Rana	The actor talks about his memories about his grandfather and how he realised that his life revolved around his grandfather	Tamil	Telugu
Yoga and walking helped me to shed my weight: Raashi Khanna	The actress talks about her new avatar and how she managed to finally shed weight	Telugu	Telugu
Is Mahesh Babu playing himself in Koratala Siva's film?	The film's teaser has led credence to rumours about how Mahesh Babu's real life could have inspired the plot of his upcoming film	Telugu	Telugu
It would be stupid of me to try and emulate my dad: Goutham	In conversation with Hyderabad Times, Brahmanandams son Goutham shares his journey from flab to fab	Telugu	Telugu
Allari Naresh gets married to Virupa	Actor Allari Naresh tied the knot with Virupa at a traditional ceremony in Hyderabad on Friday night	None	Telugu

Title	Description	Predicted category	Actual category
Whos the most desirable woman?	That beautiful face which can send a million hearts aflutter; those bedroom eyes that scream come hither, that smile that leaves you utterly disarmed... The woman who spells desirable.	Telugu	Malayalam
Who is the most desirable man?	The heartthrobs, the chocolate boys, the stars, the movers and the shakers - ladies, come take your pick.	Malayalam	Malayalam
Priyadarshan directs Kohli, Gayle and Watson	The filmmaker shot an advertisement with the cricketers recently in Bangalore	Malayalam	malayalam
Priyadarshan to direct Prithviraj in his next?	The film's script will be written by 1983 scribe Bipin Chandran	malayalam	Malayalam
Several dialogues came to me in my dream: Vineeth	The actor-director took to his social networking page to share his excitement about the review of Jacobinte Swargarajyam	Malayalam	Malayalam
Amal Neerad, Siddique team up for the first time	The duo will work together for the upcoming Dulquer Salmaan movie	Malayalam	Malayalam
Dulquer's 100 Days of Love to be remade in Telugu	The movie has Dulquer Salmaan and Nithya Menen in the lead	Malayalam	Malayalam
Dulquer to film in Pala for his next	The movie will be a love story, directed by Amal Neerad	Malayalam	Malayalam
Don Max reveals why he cast Meera Jasmine	The actress will play a stylish cop in his film	Malayalam	Malayalam
Walayar Paramasivam will start filming in 2016	Dileep made the announcement earlier this week on social media	Malayalam	Malayalam
Priyadarshan says Charlie, Premam inspired him	The director talks about how his upcoming movie Oppam will be his foray into the so-called new-gen filmmaking	Malayalam	Malayalam
Samuthirakani in a cameo in Karinkunnam Sixes	The film has Manju Warrier playing a volleyball coach	Malayalam	Malayalam
Vedhika joins Welcome to Central Jail	The film has Dileep in the lead	Malayalam	Malayalam
Anoop Menon to pen a love story based in Malaysia	The Malaysian tourism department is involved with the project	Malayalam	Malayalam
Madonna will make it big in M-town: Dileep	Dileep and Madonna co-star in Siddique-Lal duo's King Liar	Malayalam	Malayalam
Dont want to regret our choice of movies: Vineeth	The actor-director talks about what his friends' circle wants to accomplish with their movies	MMalayalam	Malayalam
Biju Menon to play Innocents driver	The duo will team up for Jos Thomas' Vellakaduva	Malayalam	Malayalam
White release pushed by two weeks	The movie has Mammootty and Huma Qureshi in the lead	Malayalam	Malayalam
Dont compare King Liar with Two Countries: Dileep	The film is directed by Lal and scripted by Siddique	Malayalam	Malayalam
James and Alice wraps up filming	The film has Prithviraj and Vedhika in the lead	Malayalam	Malayalam
Censor cuts irkes Rajeshwari Poornachandra Tejaswi	The movie has got good response from movie goers.	Malayalam	Kannada

Title	**Description**	**Predicted category**	**Actual category**
'Chandi Kori' team all geared up for 100 days bash on Jan 4	Bolli Movies Tulu film 'Chandi Kori' that broke-even in 52 days of its release clocked its 100th day on January 2.	Malayalam	Kannada
After Darshan, Urvashi Rautela picks Salman?	The Mr Airavathaa heroine was seen at Salman's 50th birthday bash	Kannada	Kannada
Kannada actor Shivarajkumar suffers heart attack	Kannada actor Shivaraj Kumar, son of late matinee idol Dr Rajkumar has been hospitalised after he complained of a severe chest pain while working out at a gym on Tuesday.	Kannada	Kannada

- Futhermore, I classified the remaining items using fisher classifier and got the cprob() and fisherprob(). The code for this is illustrated in Listing 2.3.
- The output table with 'Title', 'Predicted category', 'Actual category', 'cprob' and 'fisherprob' is illustrated below in Table 2.8.

Table 2.8. Output with 'Title', 'Predicted category', 'Actual category', 'cprob' and 'fisherprob'

Title	Predicted category	Actual category	cprob	fisherprob	string
Sequels galore to storm Sandalwood screens	kannada	kannada	0	0.5966	sandalwood screens
Shubra Aiyappa at a cocktail party at Sheraton Grand in Bengaluru	malayalam	kannada	0	0.5966	shubra aiyappa
Shubra Aiyappa at a cocktail party at Sheraton Grand in Bengaluru	kannada	kannada	0	0.7428	sheratn grand
Shubra Aiyappa at a cocktail party at Sheraton Grand in Bengaluru	kannada	kannada	1	0.8333	bengaluru
When values get the better of society	malayalam	kannada	0	0.5	values
When opposites attract	kannada	kannada	0	0.5	opposites
Who will replace Chandan in Lakshmi Baramma?	kannada	kannada	0	0.5	chandan
Award winners at the 7th BIFFES 2014	kannada	kannada	0	0.3849	award winners
Watch Sindhu Loknath sizzle	kannada	kannada	0	0.5966	sindhu loknath
Trailer: Mr and Mrs Ramachari	malayalam	kannada	0	0.4481	mr and mrs ramachari
Priyanka Kothari is Bullet Rani	kannada	kannada	0	0.5966	bullet rani
Trailer of Yogaraj Bhat's Vaastu Prakara	kannada	kannada	0	0.5	yogaraj
Watch: Trailer of Namo Bhootaathma	Malayalam	kannada	0	0.5	bhootaathma
Watch: The teaser trailer of Godhi Banna Sadharna Mykattu	kannada	kannada	0	0.698	godhi banna sadgarna mykattu
Shivarajkumar's Belli cut by 13 minutes	malayalam	kannada	0	0.5	shivrajkumar
Kannada film Sachin Tendulkar Alla to be remade in Telugu	kannada	kannada	0	0.5966	sachin tendulkar
Churni's film moves Bangla blogger in Oz	kannada	bengali	0	0.5	churni
FTII workshop film still speaks to viewers	Malayalam	bengali	0	0.5	ftii
Celebrated movie gets short shrift	Hindi	bengali	0	0.5	shrift
She was always there for us	Tamil	bengali	0	0.5966	sharmila tagore

Title	Predicted category	Actual category	cprob	fisherprob	string
SikhNet To Launch Free Online Animated Story 'Kaur'	Malayalam	punjabi	0	0.5	kaur
I am ready to learn Marathi: Indraneil Sengupta	Malayalam	marathi	0	0.3849	indraneil sengupta
Marathi film leaked online, producer lodges complaint	Telugu	marathi	0	0.5966	dagadi chawl
Court actress Geetanjali Kulkarni: I appreciate experimental films	marathi	marathi	0	0.5	geetanjali
Whats brewing Adinath?	marathi	marathi	0	0.5	adinath
New Ganpati song from Bikers Adda	marathi	marathi	0	0.5966	new ganpato
Mohan Joshi is Tanvi's mentor	marathi	marathi	1	0.9	marathi
Sena slams BJP minister for watering down 'prime-time' rule	marathi	marathi	0	0.698	sena slams bjp minister
Hollywood enters Marathi industry	marathi	marathi	0	0.5	hollywood
I regret not listening to Aamir Khan	marathi	marathi	0	0.5966	aamir khan
Is Adinath a joru ka ghulam?	kannada	marathi	0	0.25	adinath
Mahesh Manjrekar back in business?	marathi	marathi	0	0.5014	mahesh manjrekar
Sonali goes rehearsing on bicycle	malayalam	marathi	0	0.5	sonali
Madhur Bhandarkars Lohri wishes	marathi	marathi	0	0.5	madhur
Sonali to portray Geeta Dutt	marathi	marathi	1	0.75	sonali
John Travolta bumps into Riteish	Malayalam	marathi	0	0.5966	john travolta
Sadashiv Amrapurkar was born as Ganesh Kumar Narbode	marathi	marathi	0	0.5	sadashiv
Actress ready for a butt selfie?	marathi	marathi	0.3148	0.3333	actress
Sunny Leone and Adinath Kothare together?	marathi	marathi	0	0.5966	sunny leone
Shocking revelations: Actress admits being sexually abused	marathi	marathi	0	0.9188	nikita gokhale

Title	Predicted category	Actual category	cprob	fisherprob	string
Celebs strive for quality family time for their first Gudi Padwa post marriage	marathi	marathi	0	0.25	celebs
Holi Aail Baa: These Bhojpuri songs will add more colour to your Holi	Malayalam	bhojpuri	0	0.6552	holi aail baa
Havent kissed my co-star: Rani Chatterjee	kannada	bhojpuri	0	0.7428	rani chatterjee
Hogi Pyar Ki Jeet goes on floors	bhojpuri	bhojpuri	0	0.6552	hogi pyar ki jeet
Injured Shubhi Sharma out of Hogi Pyar Ki Jeet	bhojpuri	bhojupuri	0	0.5966	shubhi sharma
Poonam Dubey to star in Hum Hai Jodi No. 1	bhojupuri	bhojpuri	0	0.5966	poonam dubey
Khesari Lal Yadavs next goes on floors	bhojpuri	bhojpuri	0	0.7239	khesari lal
Veteran Gujarati actress Padmarani passes away	marathi	gujarati	0	0.5	padmarani
Aaj Jane Ki Zid Na Karo staged in Vadodara	kannada	gujarati	0	0.5	vadodara
Gujarati theatre needs better writers: Rajeev Mehta	marathi	gujarati	0	0.3849	rajeev mehta
Gujarati celebs who love to cook	Malayalam	gujarati	0	0.125	gujarati
Plays need to be marketed better: Anupam Kher	gujarati	gujarati	0	0.5966	anupam kher
Bhoomi, Samir, Mana bond over music	gujarati	gujarati	0	0.5	music
Bas Ek Chance is an urban Gujarati film: Kirtan Patel	Malayalam	gujarati	0	0.5966	kirtan patel
War of the roses in Aa Toh Prem Chhe	gujarati	gujarati	0	0.5966	war of roses
The only way to survive in this industry is to want to be in it badly	malayalam	gujarati	0	0.1998	survive industry
Music in Gujarati films a largely untapped market: Parthiv Gohil	gujarati	gujarati	1	0.75	music
Have fond memories of working in Gujarati films: Upasna Singh Bharadwaj	gujarati	gujarati	0	0.476	upasna singh bharadwaj

Title	Predicted category	Actual category	cprob	fisherprob	string
Good content is important for audience now: Naishadh Purani	gujarati	gujarati	0	0.5966	naishadg purani
Sunburn Reload: EDM Fest floors Barodians!	gujarati	gujarati	0	0.5966	sunburn reload
I had a good time in Vadodara: Vijender Singh	gujarati	gujarati	0	0.5655	vijendra singh
Fugly stars groove at Vadodara!	gujarati	gujarati	1	0.875	vadodara
Vadodara has historical character in abundance: Tushar Unadkat	gujarati	gujarati	0	0.5	vadodraa
Chandan Rathod and Preet Mulani's romantic sojourn	gujarati	gujarati	0	0.3849	chandan rathod
My first girlfriend was a Gujarati: Vishal Malhotra	gujarati	gujarati	0	0.5	girlfriend
My entire life revolves around my mom: Tanvi Vyas	gujarati	gujarati	0	0.5014	tanvi vyas
Lions of Gujarat to be released in Gujarati and Hindi	marathi	gujarati	0	0.5966	lions of gujarath
Luxury car's magical March in Gujarat	marathi	others	0	0.3849	luxry car
A happening Xmas bash in Hyderabad	kannada	others	0	0.5966	happenung xmas
Hyderabad gets a taste of Subramaniam gharana	others	others	0	0.5	others
A lavish wedding celebration in Hyderabad	None	others	0	0.5966	wedding celebration
When metalheads had a blast in Hyderabad	others	others	0	0.5	metalheads
We love handspun fabrics, say the P3Ps of Hyderabad	others	others	0	0.5966	love handspun
Sherry Javeri hosts a dinner party for friends in Hyderabad	others	others	0	0.5966	sherry javeri
Mallika Sarabhai's multi-media dance-drama in Hyderabad	others	others	0	0.5966	mallika sarabhai

Title	Predicted category	Actual category	cprob	fisherprob	string
P3Ps get baking in Hyderabad	others	others	0	0.5	p3p
A perfect sundowner to end the weekend in Hyderabad	others	others	0	0.5	perfect
TR Sahwney Motors opens its seventh showroom in Ghaziabad	others	others	0	0.5966	sahwney motors
Singer Karthik performs with his band Arka at a party at ITC Grand Chola, Chennai	others	others	0	0.5966	singer karthik
Manju Warrier and Ranjith at a book launch in Kochi	others	others	0	0.5966	manju warrier
Ladies have a stylish day out at Westin, Hyderabad	others	others	0	0.5	westin
Sunny Leone sizzles in Hyderabad's New Year party	others	others	0	0.5966	sunny leone
Santa comes in an auto at a Christmas event in Westin, Hyderabad	others	others	0	0.5	santa
Isha Sharvani dressed to the at the audio launch of Iyobinte Pusthakam, held in Kochi	others	others	0	0.5966	isha sharvani
Actress Rachana Narayanankutty spotted at the pooja function of Thilothama held in a prominent hotel in Trivandrum	others	others	0	0.5	rachana
Actress Tarushi at the promotional event of her film Study Tour, in Trivandrum	others	others	0	0.7428	actress tarushi

Code Listing

```
from sqlite3 import dbapi2 as sqlite
import re
import math

def getwords(doc):
  splitter=re.compile('\\W*')
  print doc
  words=[s.lower() for s in splitter.split(doc)
         if len(s)>2 and len(s)<20]

  return dict([(w,1) for w in words])

class classifier:
  def __init__(self,getfeatures,filename=None):
    self.fc={}
    self.cc={}
    self.getfeatures=getfeatures

  def setdb(self,dbfile):
    self.con=sqlite.connect(dbfile)
    self.con.execute('create table if not exists fc(feature,
        category,count)')
    self.con.execute('create table if not exists cc(category
        ,count)')


  def incf(self,f,cat):
    count=self.fcount(f,cat)
    if count==0:
      self.con.execute("insert into fc values ('%s','%s',1)"
                       % (f,cat))
    else:
      self.con.execute(
        "update fc set count=%d where feature='%s' and
            category='%s'"
        % (count+1,f,cat))

  def fcount(self,f,cat):
    res=self.con.execute(
      'select count from fc where feature="%s" and category
          ="%s"'
      %(f,cat)).fetchone()
    if res==None: return 0
    else: return float(res[0])

  def incc(self,cat):
    count=self.catcount(cat)
```

```
44      if count==0:
45        self.con.execute("insert into cc values ('%s',1)" % (
              cat))
46      else:
47        self.con.execute("update cc set count=%d where
              category='%s'"
48                          % (count+1,cat))
49
50    def catcount(self,cat):
51      res=self.con.execute('select count from cc where
            category="%s"'
52                           %(cat)).fetchone()
53      if res==None: return 0
54      else: return float(res[0])
55
56    def categories(self):
57      cur=self.con.execute('select category from cc');
58      return [d[0] for d in cur]
59
60    def totalcount(self):
61      res=self.con.execute('select sum(count) from cc').
            fetchone();
62      if res==None: return 0
63      return res[0]
64
65
66    def train(self,item,cat):
67      features=self.getfeatures(item)
68      for f in features:
69        self.incf(f,cat)
70
71      self.incc(cat)
72      self.con.commit()
73
74    def fprob(self,f,cat):
75      if self.catcount(cat)==0: return 0
76
77      return self.fcount(f,cat)/self.catcount(cat)
78
79    def weightedprob(self,f,cat,prf,weight=1.0,ap=0.5):
80      basicprob=prf(f,cat)
81
82      totals=sum([self.fcount(f,c) for c in self.categories()
            ])
83
84      bp=((weight*ap)+(totals*basicprob))/(weight+totals)
85      return bp
86
87
```

```
88
89
90  class naivebayes(classifier):
91
92    def __init__(self,getfeatures):
93      classifier.__init__(self,getfeatures)
94      self.thresholds={}
95
96    def docprob(self,item,cat):
97      features=self.getfeatures(item)
98
99      p=1
100     for f in features: p*=self.weightedprob(f,cat,self.fprob
            )
101     return p
102
103   def prob(self,item,cat):
104     catprob=self.catcount(cat)/self.totalcount()
105     docprob=self.docprob(item,cat)
106     return docprob*catprob
107
108   def setthreshold(self,cat,t):
109     self.thresholds[cat]=t
110
111   def getthreshold(self,cat):
112     if cat not in self.thresholds: return 1.0
113     return self.thresholds[cat]
114
115   def classify(self,item,default=None):
116     probs={}
117     max=0.0
118     for cat in self.categories():
119       probs[cat]=self.prob(item,cat)
120       if probs[cat]>max:
121         max=probs[cat]
122         best=cat
123
124     for cat in probs:
125       if cat==best: continue
126       if probs[cat]*self.getthreshold(best)>probs[best]:
              return default
127     return best
128
129 class fisherclassifier(classifier):
130   def cprob(self,f,cat):
131     clf=self.fprob(f,cat)
132     if clf==0: return 0
133
```

```
134     freqsum=sum([self.fprob(f,c) for c in self.categories()
            ])
135
136     p=clf/(freqsum)
137
138     return p
139   def fisherprob(self,item,cat):
140     p=1
141     features=self.getfeatures(item)
142     for f in features:
143       p*=(self.weightedprob(f,cat,self.cprob))
144
145     fscore=-2*math.log(p)
146
147     return self.invchi2(fscore,len(features)*2)
148   def invchi2(self,chi, df):
149     m = chi / 2.0
150     sum = term = math.exp(-m)
151     for i in range(1, df//2):
152         term *= m / i
153         sum += term
154     return min(sum, 1.0)
155   def __init__(self,getfeatures):
156     classifier.__init__(self,getfeatures)
157     self.minimums={}
158
159   def setminimum(self,cat,min):
160     self.minimums[cat]=min
161
162   def getminimum(self,cat):
163     if cat not in self.minimums: return 0
164     return self.minimums[cat]
165   def classify(self,item,default=None):
166     best=default
167     max=0.0
168     for c in self.categories():
169       p=self.fisherprob(item,c)
170       if p>self.getminimum(c) and p>max:
171         best=c
172         max=p
173     return best
```

Listing 2.1. 'docclass.py' from PCI book code

Code Listing

```
1  import feedparser
2  import re
3  import math
4  import docclass
5
6
7  # Takes a filename of URL of a blog feed and classifies the
       entries
8  def read(feed,classifier):
9
10     splitExpression = re.compile(r"<[^>]+>")
11     itemCount = 0
12     print "first 50 entry classification"
13     # Get feed entries and loop over them
14     f=feedparser.parse(feed)
15
16     for item in f['items'][0:86]:
17         itemCount = itemCount+1
18         print itemCount
19         print '-----'
20
21         # Print the contents of the item
22         title = item['title'].encode('utf-8')
23         print 'Title: '+title
24
25         description = item['description'].encode('utf-8')
26         print 'Description:' +description
27
28         # Combine all the text to create one item for the
               classifier
29         fulltext = '%s\n%s' % (item['title'], item['summary'
               ])
30         fulltext = fulltext.replace("'","")
31         predictedString = str(classifier.classify(fulltext))
32
33         # Print the best guess at the current category
34         print "predicted category",predictedString
35
36         # Ask the user to specify the correct category and
               train on that
37
38         actual =raw_input('Enter the actual category it
               belongs to: ')
39         classifier.train(fulltext,actual)
```

Listing 2.2. 'feedfilter.py' from PCI book code

Code Listing

```
import feedparser
import re
import math
import docclass


# Takes a filename of URL of a blog feed and classifies the
    entries
def read(feed,classifier):

    print "entering into another mode"
    itemCount = 86
    print "other 50 item classification"
    # Get feed entries and loop over them
    f=feedparser.parse(feed)
    for item in f['items'][0:10]:
        itemCount = itemCount+1
        print  itemCount
        print '-----'

        # Print the contents of the item
        title = item['title'].encode('utf-8')
        print 'Title: \t'+title

        description = item['description'].encode('utf-8')
        print 'Description:' +description

        # Combine all the text to create one item for the
            classifier
        fulltext = '%s\n%s' % (item['title'], item['summary'
            ])
        fulltext = fulltext.replace("'","")
        predicted= str(classifier.classify(fulltext))

        # Print the best guess at the current category
        print "predicted category",predicted

        # Ask the user to specify the correct category and
            train on that

        actual =raw_input('Enter the actual category it
            belongs to: ')
        feature = raw_input('Enter a string to classifier:')


        cprobabilty = round(classifier.cprob(feature,
            predicted),4)
```

```
42          print 'cprobabilty:',cprobabilty
43
44          fisherprobabilty = round(classifier.fisherprob(
                feature,predicted),4)
45          print 'fisherprobabilty:',fisherprobabilty
46          classifier.train(fulltext,actual)
```

Listing 2.3. 'feedfilter.py' from PCI book code

Code Listing

```
1 import docclass
2 import feedfilter
3
4 def main():
5     cl=docclass.fisherclassifier(docclass.getwords)
6     cl.setdb('smajeti.db')
7     print "testing the program"
8     feedfilter.read('toiEntertainment.xml',cl)
```

Listing 2.4. 'classifyEntries.py' from 'Document-filtering' powerpoint slide 26

3

Question 3

Assess the performance of your classifier in each of your categories by computing precision, recall, and F-measure.

Following are the steps I have taken to solve the problem:

- To find precision and recall values I need 'True positive', 'True negative', 'False positive' and 'False negative'.
- The values for 'True positive', 'True negative', 'False positive' and 'False negative' for each category is illustrated in Table 3.1.

Table 3.1. Example

	hindi	telugu	english	kannada	marathi	malayalam	others	punjabi	bhojpuri	gujarati	tamil	bengali	None
FP	1	1	0	5	4	14	0	0	0	0	1	0	1
FN	0	0	0	5	5	0	3	1	2	7	0	4	0
TP	0	0	0	11	15	0	16	0	4	13	0	0	0
TN	85	85	86	65	62	72	67	85	80	66	85	82	85

- The Formulas for 'Precision', 'Recall' and 'F-Measure' are as follows:
  Precision = TP/(TP+FP)
  Recall= TP/(TP+FN)
  F-Measure = 2TP/(2TP+FP+FN)
- The values for 'Precision', 'Recall' and 'F-Measure' for each category is illustrated in Table 3.2.

Table 3.2. Example

Category	Precision	Recall	F-measure
hindi	0	0	0
telugu	0	0	0
english	0	0	0
kannada	0.6875	0.6875	0.6875
marathi	0.7894	0.75	0.7692
malayalam	0	0	0
others	1	0.8421	0.9142
punjabi	0	0	0
bhojpuri	1	0.6667	0.8
gujarati	1	0.65	0.7878
tamil	0	0	0
bengali	0	0	0
None	0	0	0

References

1. Python code feedfilter.py from PCI book. Igraph Tutorial. `https://github.com/uolter/PCI/blob/master/chapter6/feedfilter.py`, 2007, Toby Segaran
2. Python code docclass.py from PCI book. Download GraphML for Karate Club. `https://github.com/uolter/PCI/blob/master/chapter6/docclass.py`, 2007, Toby Segaran