Mining Diverse *Views* from Related Articles

Ravali Pochampally
Kamal Karlapalem
[IIIT Hyderabad]

Information Overload

"I'm sorry. It appears Mr. Mitchell won't be accepting any more information today."

- WWW
  - diverse content
  - 100+ articles on major topics
- Google News/Amazon
  - organized
  - (yet) too much text

Summarization

- Condenses information

✓ salient points
✓ length α (1/content)
✓ user-specified parameters

- Lacks Organization

✗ delineation of issues
✗ model diversity
✗ too long (?)

What is a view?

- A *view* intends to represent an issue pertaining to a set of related[1] articles
  - organized (multiple concise views)
  - information exploration
  - detailed snapshot

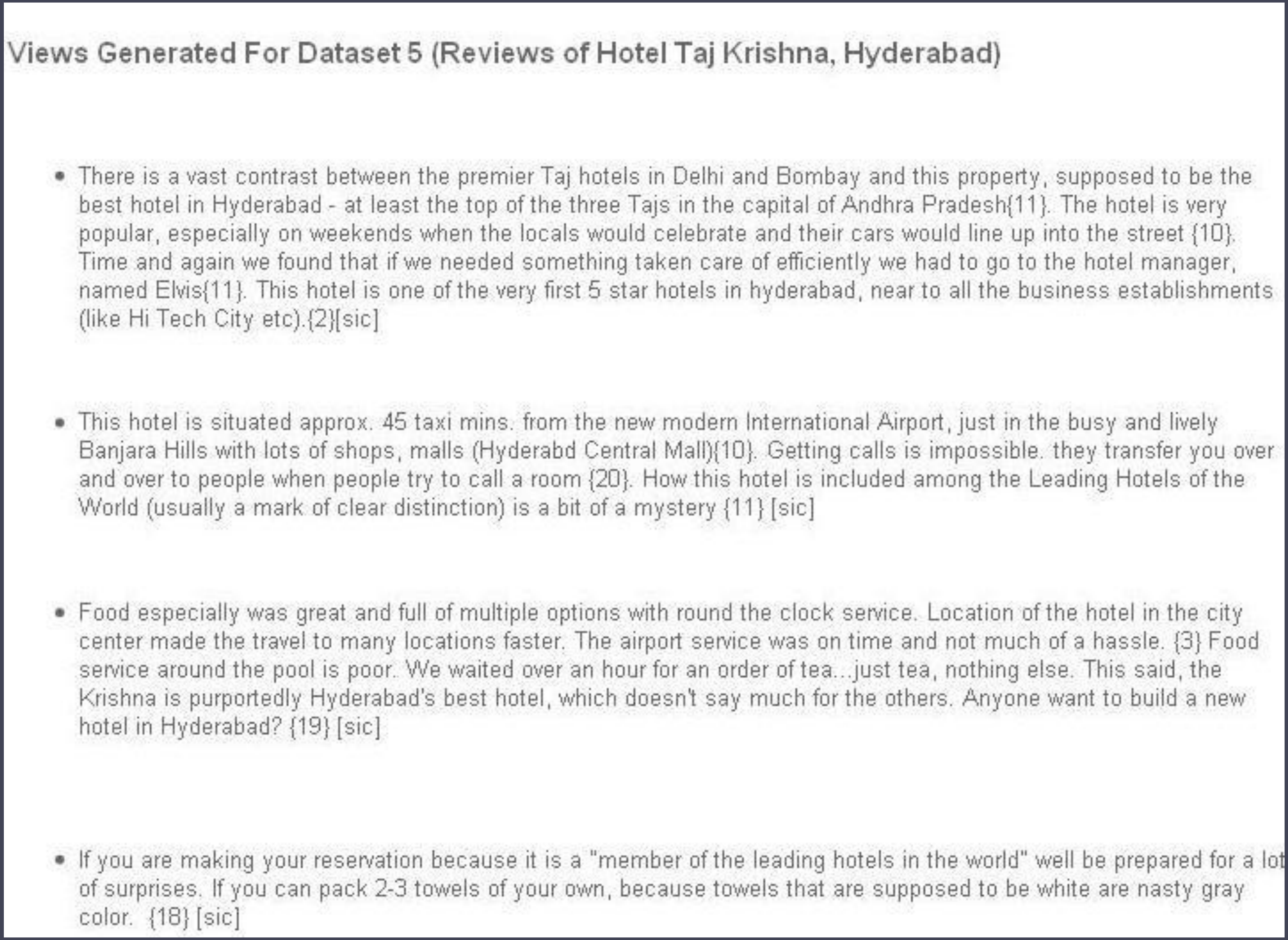
- Example[2] : review dataset (hotel)

- views	[positive, negative, food, facilities]
- summary	[unorganized]

[1] articles concerning a common topic (FIFA 2010, swine flu in India etc.)

[2] http://sites.google.com/site/diverseviews/comparison

Example View [review dataset]



Views Generated For Dataset 5 (Reviews of Hotel Taj Krishna, Hyderabad)

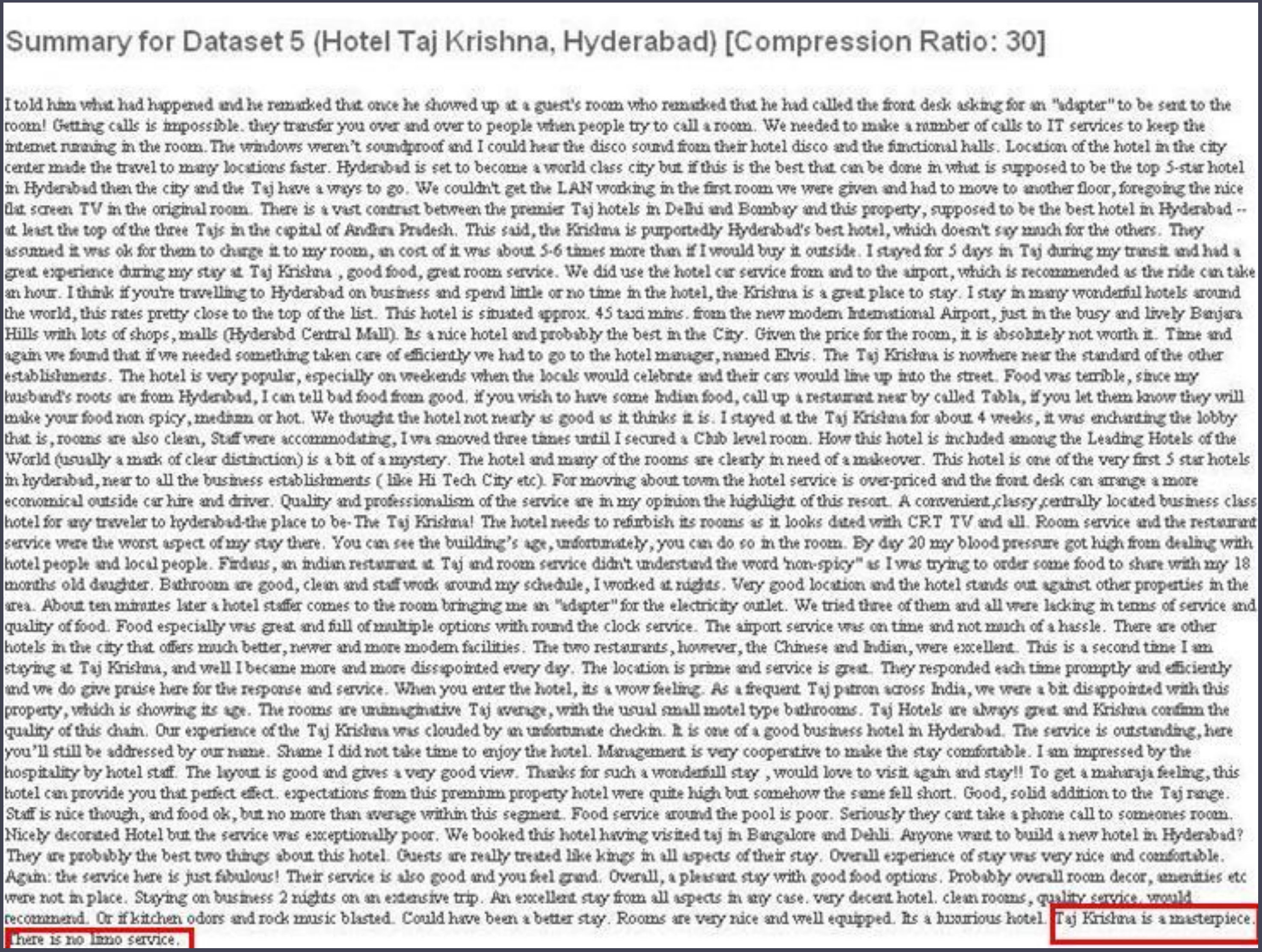
- There is a vast contrast between the premier Taj hotels in Delhi and Bombay and this property, supposed to be the best hotel in Hyderabad - at least the top of the three Tajs in the capital of Andhra Pradesh{11}. The hotel is very popular, especially on weekends when the locals would celebrate and their cars would line up into the street {10}. Time and again we found that if we needed something taken care of efficiently we had to go to the hotel manager, named Elvis{11}. This hotel is one of the very first 5 star hotels in hyderabad, near to all the business establishments (like Hi Tech City etc).{2}[sic]

- This hotel is situated approx. 45 taxi mins. from the new modern International Airport, just in the busy and lively Banjara Hills with lots of shops, malls (Hyderabd Central Mall){10}. Getting calls is impossible. they transfer you over and over to people when people try to call a room {20}. How this hotel is included among the Leading Hotels of the World (usually a mark of clear distinction) is a bit of a mystery {11} [sic]

- Food especially was great and full of multiple options with round the clock service. Location of the hotel in the city center made the travel to many locations faster. The airport service was on time and not much of a hassle. {3} Food service around the pool is poor. We waited over an hour for an order of tea...just tea, nothing else. This said, the Krishna is purportedly Hyderabad's best hotel, which doesn't say much for the others. Anyone want to build a new hotel in Hyderabad? {19} [sic]

- If you are making your reservation because it is a "member of the leading hotels in the world" well be prepared for a lot of surprises. If you can pack 2-3 towels of your own, because towels that are supposed to be white are nasty gray color. {18} [sic]

Example Summary [review dataset]



Summary for Dataset 5 (Hotel Taj Krishna, Hyderabad) [Compression Ratio: 30]

I told him what had happened and he remarked that once he showed up at a guest's room who remarked that he had called the front desk asking for an "adapter" to be sent to the room! Getting calls is impossible. they transfer you over and over to people when people try to call a room. We needed to make a number of calls to IT services to keep the internet running in the room. The windows weren't soundproof and I could hear the disco sound from their hotel disco and the functional halls. Location of the hotel in the city center made the travel to many locations faster. Hyderabad is set to become a world class city but if this is the best that can be done in what is supposed to be the top 5-star hotel in Hyderabad then the city and the Taj have a ways to go. We couldn't get the LAN working in the first room we were given and had to move to another floor, foregoing the nice flat screen TV in the original room. There is a vast contrast between the premier Taj hotels in Delhi and Bombay and this property, supposed to be the best hotel in Hyderabad -- at least the top of the three Tajs in the capital of Andhra Pradesh. This said, the Krishna is purportedly Hyderabad's best hotel, which doesn't say much for the others. They assumed it was ok for them to charge it to my room, an cost of it was about 5-6 times more than if I would buy it outside. I stayed for 5 days in Taj during my transit and had a great experience during my stay at Taj Krishna , good food, great room service. We did use the hotel car service from and to the airport, which is recommended as the ride can take an hour. I think if you're travelling to Hyderabad on business and spend little or no time in the hotel, the Krishna is a great place to stay. I stay in many wonderful hotels around the world, this rates pretty close to the top of the list. This hotel is situated approx. 45 taxi mins. from the new modern International Airport, just in the busy and lively Banjara Hills with lots of shops, malls (Hyderabd Central Mall). Its a nice hotel and probably the best in the City. Given the price for the room, it is absolutely not worth it. Time and again we found that if we needed something taken care of efficiently we had to go to the hotel manager, named Elvis. The Taj Krishna is nowhere near the standard of the other establishments. The hotel is very popular, especially on weekends when the locals would celebrate and their cars would line up into the street. Food was terrible, since my husband's roots are from Hyderabad, I can tell bad food from good. if you wish to have some Indian food, call up a restaurant near by called Tabla, if you let them know they will make your food non spicy, medium or hot. We thought the hotel not nearly as good as it thinks it is. I stayed at the Taj Krishna for about 4 weeks, it was enchanting the lobby that is, rooms are also clean, Staff were accommodating, I wa smoved three times until I secured a Club level room. How this hotel is included among the Leading Hotels of the World (usually a mark of clear distinction) is a bit of a mystery. The hotel and many of the rooms are clearly in need of a makeover. This hotel is one of the very first 5 star hotels in hyderabad, near to all the business establishments (like Hi Tech City etc). For moving about town the hotel service is over-priced and the front desk can arrange a more economical outside car hire and driver. Quality and professionalism of the service are in my opinion the highlight of this resort. A convenient,classy,centrally located business class hotel for any traveler to hyderabad-the place to be-The Taj Krishna! The hotel needs to refurbish its rooms as it looks dated with CRT TV and all. Room service and the restaurant service were the worst aspect of my stay there. You can see the building's age, unfortunately, you can do so in the room. By day 20 my blood pressure got high from dealing with hotel people and local people. Firdaus, an indian restaurant at Taj and room service didn't understand the word 'non-spicy" as I was trying to order some food to share with my 18 months old daughter. Bathroom are good, clean and staff work around my schedule, I worked at nights. Very good location and the hotel stands out against other properties in the area. About ten minutes later a hotel staffer comes to the room bringing me an "adapter" for the electricity outlet. We tried three of them and all were lacking in terms of service and quality of food. Food especially was great and full of multiple options with round the clock service. The airport service was on time and not much of a hassle. There are other hotels in the city that offers much better, newer and more modern facilities. The two restaurants, however, the Chinese and Indian, were excellent. This is a second time I am staying at Taj Krishna, and well I became more and more dissapointed every day. The location is prime and service is great. They responded each time promptly and efficiently and we do give praise here for the response and service. When you enter the hotel, its a wow feeling. As a frequent Taj patron across India, we were a bit disappointed with this property, which is showing its age. The rooms are unimaginative Taj average, with the usual small motel type bathrooms. Taj Hotels are always great and Krishna confirm the quality of this chain. Our experience of the Taj Krishna was clouded by an unfortumate checkin. It is one of a good business hotel in Hyderabad. The service is outstanding, here you'll still be addressed by our name. Shame I did not take time to enjoy the hotel. Management is very cooperative to make the stay comfortable. I am impressed by the hospitality by hotel staff. The layout is good and gives a very good view. Thanks for such a wonderfull stay , would love to visit again and stay!! To get a maharaja feeling, this hotel can provide you that perfect effect. expectations from this premium property hotel were quite high but somehow the same fell short. Good, solid addition to the Taj range. Staff is nice though, and food ok, but no more than average within this segment. Food service around the pool is poor. Seriously they cant take a phone call to someones room. Nicely decorated Hotel but the service was exceptionally poor. We booked this hotel having visited taj in Bangalore and Dehli. Anyone want to build a new hotel in Hyderabad? They are probably the best two things about this hotel. Guests are really treated like kings in all aspects of their stay. Overall experience of stay was very nice and comfortable. Again: the service here is just fabulous! Their service is also good and you feel grand. Overall, a pleasant stay with good food options. Probably overall room decor, amenities etc were not in place. Staying on business 2 nights on an extensive trip. An excellent stay from all aspects in any case. very decent hotel. clean rooms, quality service, would recommend. Or if kitchen odors and rock music blasted. Could have been a better stay. Rooms are very nice and well equipped. Its a luxurious hotel. Taj Krishna is a masterpiece. There is no limo service.

Outline

- Related Work
- Problem

- Extraction of views
  - Ranking

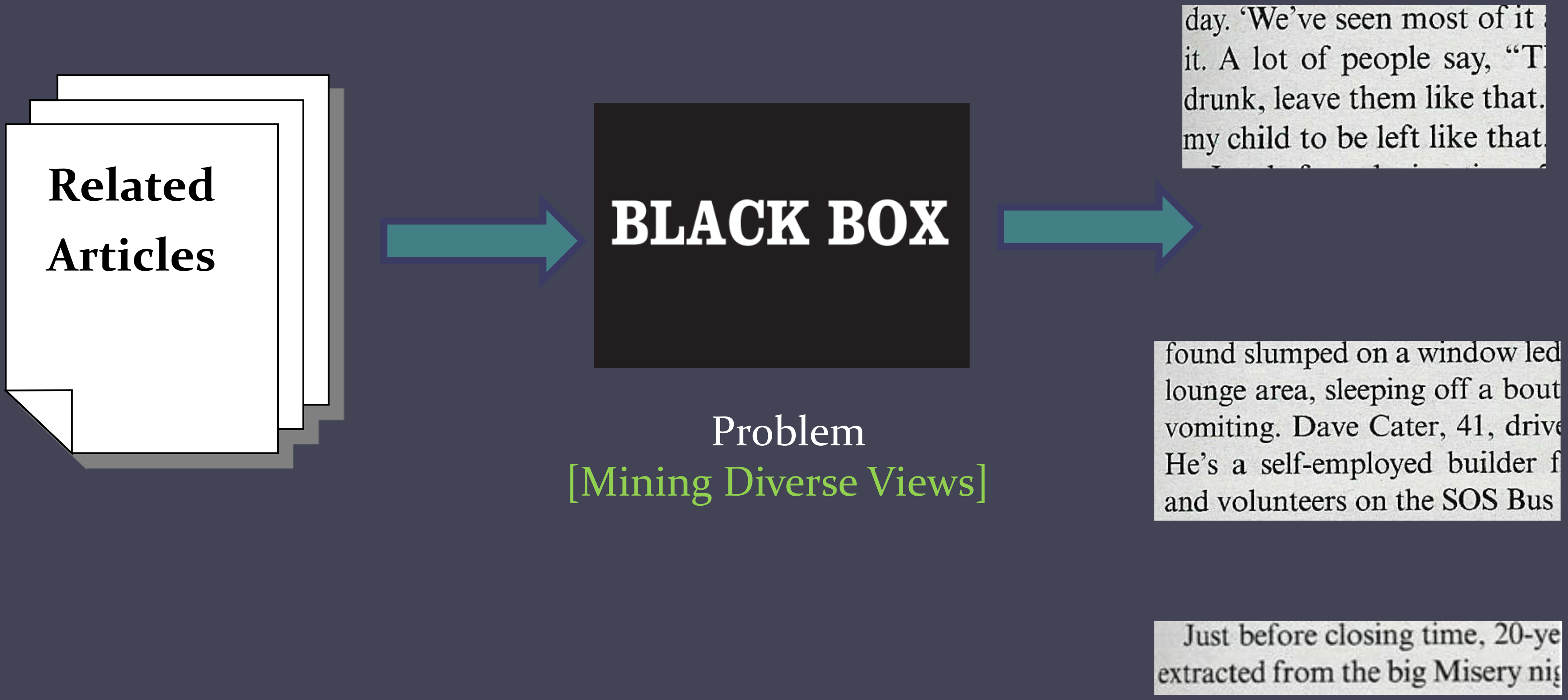
- Results
  - Discussion

Related Work

- Allison et. al
  - idea of multiple view-points
  - framework
- Tombros et. al
  - clustering of top-ranking sentences
- TextTiling [1]
  - divide text into multi-paragraph units
  - unit represents a sub-topic

[1] M. A. Hearst 1997

Problem



Related Articles

BLACK BOX

day. 'We've seen most of it
it. A lot of people say, "T
drunk, leave them like that.
my child to be left like that.

found slumped on a window led
lounge area, sleeping off a bout
vomiting. Dave Cater, 41, drive
He's a self-employed builder f
and volunteers on the SOS Bus

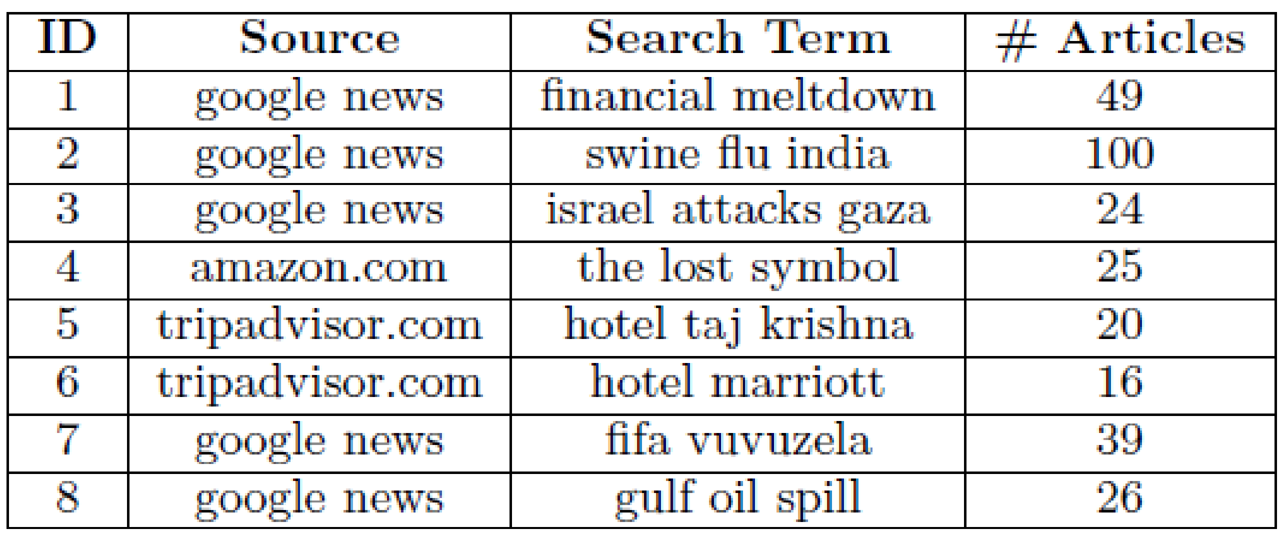
Just before closing time, 20-ye
extracted from the big Misery nig

Problem

[Mining Diverse Views]

Ranked set of views

Datasets*



ID	Source	Search Term	# Articles
1	google news	financial meltdown	49
2	google news	swine flu india	100
3	google news	israel attacks gaza	24
4	amazon.com	the lost symbol	25
5	tripadvisor.com	hotel taj krishna	20
6	tripadvisor.com	hotel marriott	16
7	google news	fifa vuvuzela	39
8	google news	gulf oil spill	26

* http://sites.google.com/site/diverseviews/datasets

Data preparation

- Datasets
  - sources: google news, amazon.com, tripadvisor.com
  - crawling + parsing [html and rss]

- Data cleaning & pre-processing
  - stopwords, stemming and duplicates
  - word-frequency, TF-IDF [1]

[1] http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

Top-ranking sentences

- Main idea
  - We score each sentence in our dataset and extract the top-ranked ones. These sentences are used to generate views [Pruning]
  - We assign a Importance (I) score to each sentence
  - Importance I_k of a sentence S_k belonging to article d_j of length r is

$$I_k = \frac{\Pi_r \mathrm{T}_{i,j}}{\mathrm{r}}$$

$$\mathrm{T}_{i,j} = \text{TF-IDF of } \mathrm{w}_i \, \epsilon \, (S_k \, \Lambda \, d_j)$$

Sentences → views

- A measure of similarity is required to extract views from sentences
  - *Semantic similarity* : likeness of meaning

- Mihalcea et. al
  - specificity of a word can be determined by its *idf*
  - we use
    - ➢ word-to-word similarity &
    - ➢ specificity - to calculate semantic similarity

Semantic similarity

- Semantic similarity between sentences S_i and S_j where w represents a word in a sentence is

$$sim(S_i, S_j) = \frac{1}{2}\left(\frac{\sum_{w \in \{S_i\}}(maxSim(w,S_j) * idf(w))}{\sum_{w \in \{S_i\}} idf(w)} + \frac{\sum_{w \in \{S_j\}}(maxSim(w,S_i) * idf(w))}{\sum_{w \in \{S_j\}} idf(w)}\right)$$

(symmetric relation & range ∈ [0,1])

- Need to define *maxSim(w, S_j)*
  - Wordnet : sets of cognitive synonyms (synsets)
  - *wup*[1] : based on path length between synsets

[1] Z. Wu 1994

Sentence clustering

- Clustering is used to extract views from the set of important sentences

- Hierarchical Agglomerative Clustering (HAC) was used
  - upper triangle [symmetric matrix]
  - no restrictions on # of clusters
  - terminate clustering when scoring parameter converges

- We treat clusters as views discussing similar content

Ranking of views

- Focus on average pair-wise similarity between the sentences in a view

- Cohesion

$$C = \frac{\Sigma_{i,j \in V} S_{i,j}}{\mathrm{len}(v)}$$

$$S_{i,j} = sim(T_i, T_j)$$

V = set of sentences (T_i) in the view

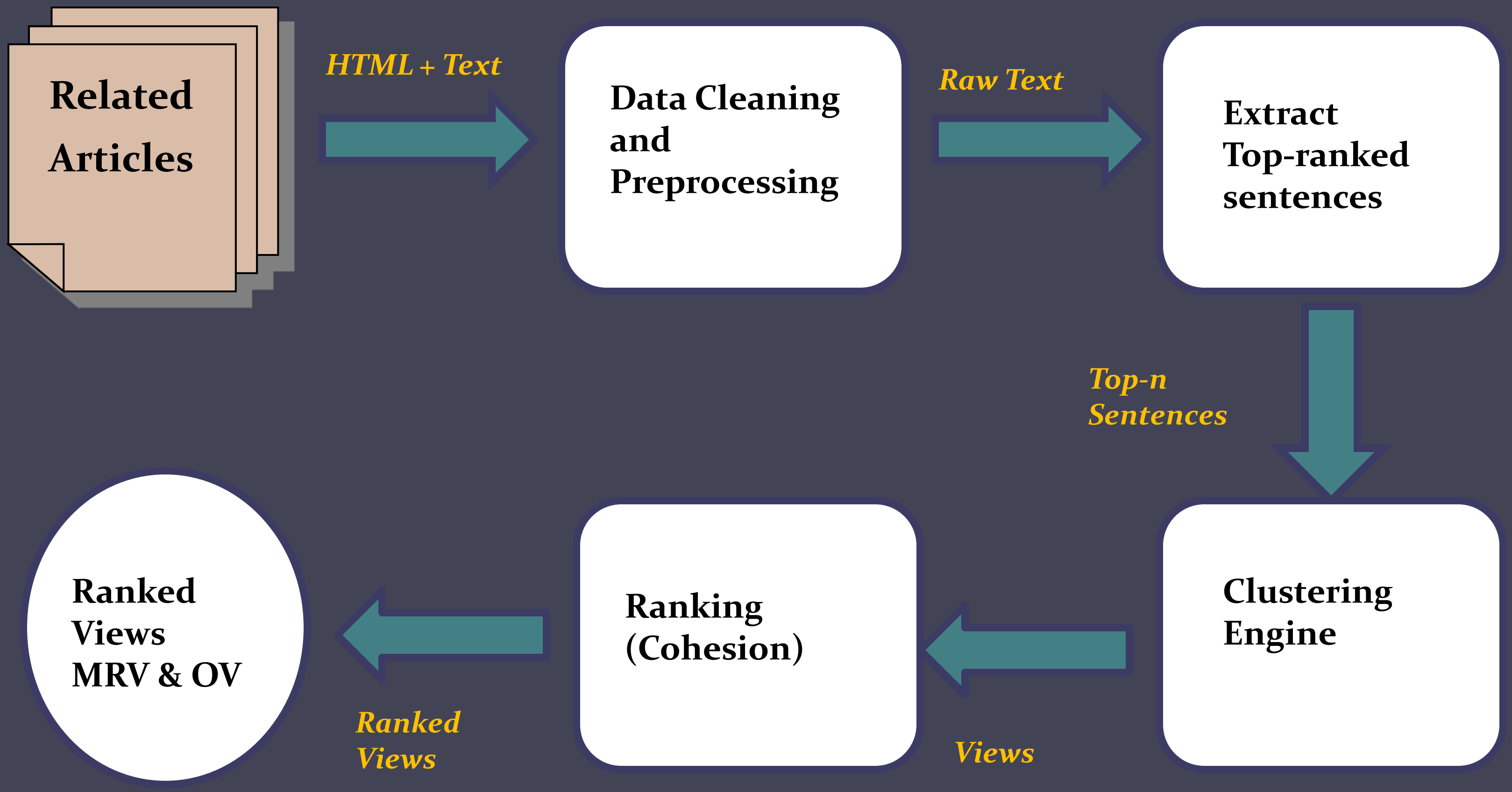
len(v) = # of sentences in the view

MRV & OV

- Most relevant view (MRV)
  - preference to views discussing similar content [greater cohesion]
  - top-ranked view

- Outlier view (OV)
  - single sentence
  - low semantic similarity with other sentences
  - Cohesion = 0 [ordered by importance]

Framework


Related Articles
HTML + Text
Data Cleaning and Preprocessing
Raw Text
Extract Top-ranked sentences
Top-n Sentences
Clustering Engine
Views
Ranking (Cohesion)
Ranked Views
Ranked Views MRV & OV

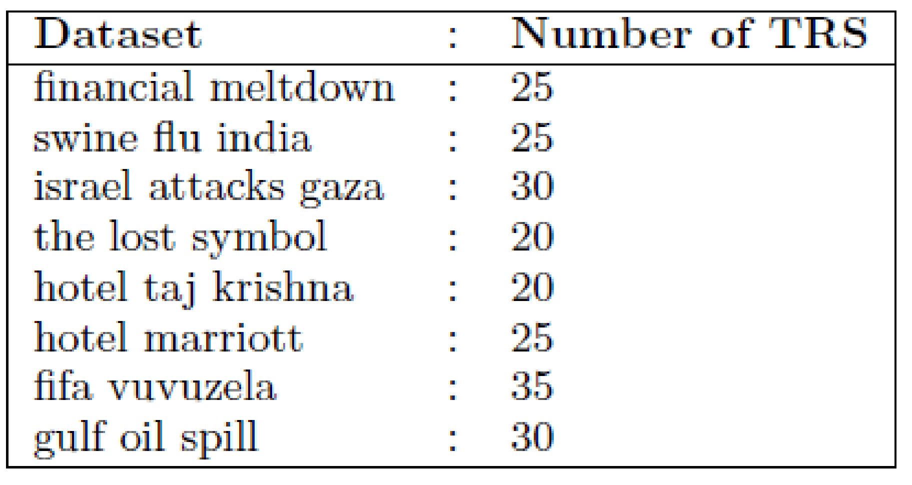
Results

- Number of top-ranking sentences (*n*) vs. *cohesion*
  - *n* which can maximize cohesion
  - median cohesion >= mean [outliers]
  - $20 <= n <= 35$
  - incremental clustering [1]

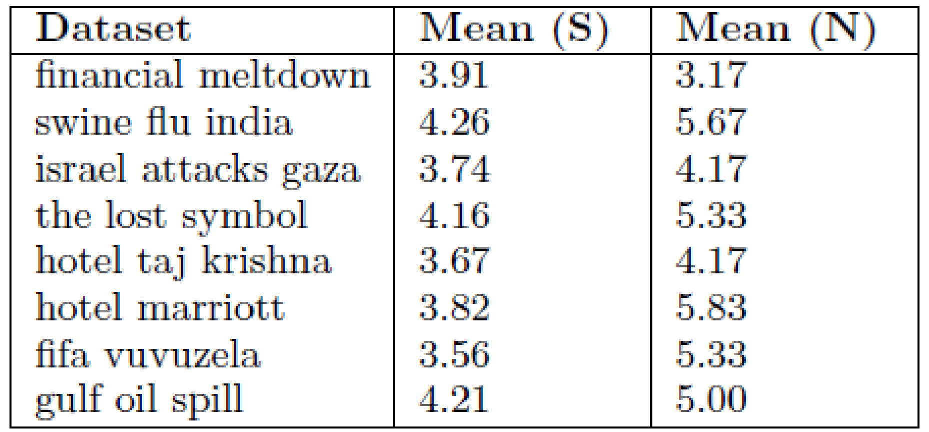
- More top-ranking sentences need not necessarily lead to views with better cohesion

[1] **M. Charikar STOC 1997**

Statistics



Dataset	:	Number of TRS
financial meltdown	:	25
swine flu india	:	25
israel attacks gaza	:	30
the lost symbol	:	20
hotel taj krishna	:	20
hotel marriott	:	25
fifa vuvuzela	:	35
gulf oil spill	:	30



Dataset	Mean (S)	Mean (N)
financial meltdown	3.91	3.17
swine flu india	4.26	5.67
israel attacks gaza	3.74	4.17
the lost symbol	4.16	5.33
hotel taj krishna	3.67	4.17
hotel marriott	3.82	5.83
fifa vuvuzela	3.56	5.33
gulf oil spill	4.21	5.00

Conclusion

- IR model as an alternate to summarization
  - ✓ multiple diverse views
  - ✓ easily navigable
  - ✓ browse top *x* views
  - ✓ detailed (yet organized) snapshot of a ToI
  - ✓ clustering at sentence/phrase level*

- Future work
  - polarity of a view
  - user feedback
    - Implicit [clicks, time-spent]
    - Explicit [user-ratings]

* **as opposed to document clustering**

Thanks!