Real-time summarization of social-media streams

Final project

Web Search
Big Picture

Stream of Tweets
Interest profiles
Daily email digest

Interest profile 1
Interest profile 2
... 
Interest profile N

Twitter icons
Project

• At the end of each day, your system must create a daily digest of Tweets that best cover the events related to an interest profile.

  - The coverage of a topic is measured in terms of information relevance, novelty and diversity.
Stream of Twitter data

• The evaluation data cover the days from August 2, 2016 00:00:00 UTC to August 11, 2016 23:59:59 UTC. You will be provided with a sample of Tweets from this period.

• The relevance judgments are also provided together with the dataset.
Some of the most promising advances in cancer involve these treatments. Here are answers to some basic questions about this rapidly evolving field.
Tweet documents: user field

"user":{"id":3078843052,
"id_str":"3078843052",
"name":"Alexandra Mcghee",
"screen_name":"nelwynhallman",
"location":null,
"url":null,
"description":null,
"protected":false,
"verified":false,
"followers_count":197,
"friends_count":195,
"listed_count":15,
"favourites_count":0,
"statuses_count":4406,
"created_at":"Sun Mar 08 05:40:28 +0000 2015",
"utc_offset":-25200,
"time_zone":"Pacific Time (US & Canada)",
"geo_enabled":false,
"lang":"en",
"contributors_enabled":false,
"is_translator":false,
"profile_background_color":"C0DEED",
"profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png",
"profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme1/bg.png",
"profile_background_tile":false,
"profile_link_color":"0084B4",
"profile_sidebar_border_color":"C0DEED",
"profile_sidebar_fill_color":"DDEEF6",
"profile_text_color":"333333",
"profile_use_background_image":true,
Interests profiles

{
  "topid": "MB246",
  "title": "Greek international debt crisis",
  "description": "Find information related to the crisis surrounding the Greek debt to international creditors, and the consequences of their possible withdrawal from the European Union."
,
  "narrative": "Given the continuing crisis over the Greek debt to international creditors, such as the International Monetary Fund (IMF), European Central Bank (ECB), and the European Commission, the user is interested in information on how this debt is being handled, including the possible withdrawal of Greece from the euro zone, and the consequences of such a move."
}
Relevant Tweets per topic per day

Figure 3: Heatmap of the distribution of all relevant and highly-relevant tweets: interest profiles in columns, days of the evaluation in rows.
Train/test interest profiles

• The interest profiles should be divided into train/test as follows:

  • **Train interest profiles:** use the even queries (e.g., RTS2, RTS4, RTS6, …) to tune and develop your system.

  • **Test interest profiles:** evaluate your system on the odd queries (RTS1, RTS3, RTS5, …)
Daily digest

• Your system must identify a batch of up to 100 ranked tweets per day per interest profile.

• The runs should be formatted as a plain text file, where each line has the following fields:

  YYYYMMDD topic_id Q0 tweet_id rank score runtag

• Basically, this is just the standard TREC format prepended with a date in format YYYYMMDD indicating the date the result was generated.
Relevance

• Some documents are more relevant than others.
  • Documents have different levels of relevance.

• The position of a document in the rank is also important to the user.
  • Relevant documents ranked top count more.
DCG: Incomplete multi-level relevance

- The Discounted Cumulative Gain measure, considers the notion of multi-level relevance:

\[
DCG_m \propto 2^{rel_i} - 1 \quad rel_i = \{0,1,2,3,...\}
\]

- The DCG measure, also considers the position where the document is on the rank:

\[
DCG_m = \sum_{i=1}^{m} \frac{2^{rel_i} - 1}{\log_2(1 + i)} \quad rel_i = \{0,1,2,3,...\}
\]

- The normalized metric measures the deviation from the optimal sort order:

\[
nDCG_m = \frac{DCG_m}{bestDCG_m}
\]

Diverse daily digest

• Many documents contain the same information, thus **may add no value** to the search results if similar documents were retrieved first.

• Similar documents must be detected and possibly removed from the results list (or grouped into a common sub-topic).
Diversity and novelty

• Diversity and novelty are difficult to evaluate.

• There is no *de facto* method to measure it.

• The goal is to measure *how diverse and novel is the information* contained in the retrieved documents.
  • Assessment focus is not at the level of the documents.
Nuggets or information facts

• A **nugget** is an information fact
  • **Documents** contain many nuggets.
  • The same **nugget** can be present in many different documents.

• The goal is to retrieve a ranked list with many different nuggets at the top of the list

• Repeated nuggets will have a decreasing importance
The $\alpha$-nDCG metric for diversity and novelty

• The relevance of a document is determined by its nuggets

$$\sum_{j=1}^{m} N(d_i, n_j).$$

and by the nuggets that occurred previously in the ranked results

$$r_{j,k-1} = \sum_{i=1}^{k-1} N(d_i, n_j),$$

• A popular metric is the $\alpha$-nDCG, where each document at position $k$ is judged by its nuggets

$$G[k] = \sum_{j=1}^{m} N(d_k, n_j)\alpha^{r_{j,k-1}}, \quad \alpha = 0.5$$
Example

• Top results for query “Norwegian Cruise Lines”

<table>
<thead>
<tr>
<th>Document Title</th>
<th>85.1</th>
<th>85.2</th>
<th>85.3</th>
<th>85.4</th>
<th>85.5</th>
<th>85.6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Carnival Re-Enters Norway Bidding</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>b. NORWEIGAN CRUISE LINE SAYS...</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>c. Carnival, Star Increase NCL Stake</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>d. Carnival, Star Solidify Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>e. HOUSTON CRUISE INDUSTRY GETS...</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>f. TRAVELERS WIN IN CRUISE...</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>g. ARMCHAIR QUARTERBACKS NEED...</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>h. EUROPE, CHRISTMAS ON SALE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>i. TRAVEL DEALS AND DISCOUNTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>j. HAVE IT YOUR WAY ON THIS SHIP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

\[ r_{j,k-1} = \sum_{i=1}^{k-1} N(d_i, n_j), \]

\[ G[k] = \sum_{j=1}^{m} N(d_k, n_j)a^{r_{j,k-1}}, \]

• The relevance of each document is: \( G = (2, \frac{1}{2}, \frac{1}{4}, 0, 2, \frac{1}{2}, 1, \frac{1}{4}, ...). \)

• What would be the ideal ordering?

\( a-c-g-b-f-c-h-i-j-d \quad G' = (2, 2, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, ...). \)
Relevant Tweets per topic per day

Figure 3: Heatmap of the distribution of all relevant and highly-relevant tweets: interest profiles in columns, days of the evaluation in rows.

Figure 4: Heatmap of the distribution of the first tweet in each semantic cluster: interest profiles in columns, days of the evaluation in rows.
Task 1: Daily digest

• Implement an incremental indexing where you store a separate index for each day. Each index contains the Tweets of that day and all previous days.

• **Baseline 1 [5%]:** Build the daily email digest as a retrieval task with the LMD retrieval model. Calibrate the $\mu$ parameter.

• **Baseline 2 [5%]:** Build the daily email digest as a retrieval task with the BM25 retrieval model. Calibrate the $b$ and $k_1$ parameters.

• **Baseline 3 [20%]:** Use pseudo-relevance feedback to expand the initial query and improve the daily summaries. Justify the selection of the retrieval models. Calibrate the parameters.

• **Evaluation [5%]:** Compare and analyse the performance of the different retrieval models.
Task 2: Diverse daily digests

- **Baseline 4 [15%]:** To address the problem of information diversity, i.e. groups of tweets that are too similar in content, implement a strategy with KMeans to remove redundant tweets from your search results.

- **Baseline 5 [15%]:** Following the MinHash algorithm, implement a near-duplicate detection method with the Jaccard coefficient to remove redundant tweets from your search results.

- **Evaluation [15%]:** Do a success and failure analysis of your results. Identify the elements in your system that contribute positively and negatively. Discuss your insights.
Task 3: Rank fusion

• **Combining multiple fields [10%]:** using the reciprocal rank fusion method, implement a search model to combine multiple ranking models.

• **Evaluation [10%]:** Compare and analyse the performance of your system with multiple fields.

• **Discussion of possible improvements [5%]**
Report

• Introduction
• Methods and algorithms
• Implementation: What are your ideas? What makes your project unique?
• Evaluation
  • Dataset description
  • Baselines
  • Results and discussion
• Possible improvements
• References
Questions?

Web Search