Infrastructural Damage Information Retrieval from Microblogs

Name : Gopal Kumar
Roll No : 1401CS17
Guide : Dr. Joydeep Chandra

Introduction

- **AIM :** To design and evaluate Information Retrieval (IR) systems that retrieve information from microblogs during disaster situations pertaining to infrastructural damage.
- Microblogging sites are important sources of situational information during disaster situations and are increasingly being used for aiding post-disaster relief operations.

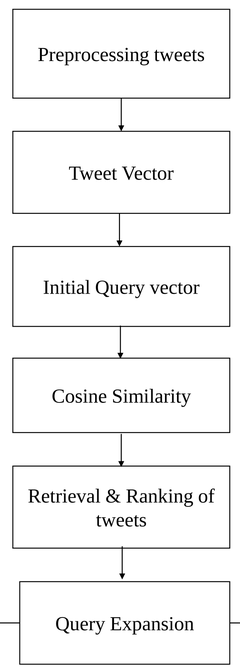
Methodology

Information retrieval is being done by structuring four models out of which Model 1 and Model 2 are already proposed[1][2] and the rest models are the novel works of this project.

A query is formed consisting of terms relevant to an information need, and tweets containing the query-terms are retrieved. The retrieved tweets are also ranked based on some measure of their relevance to the query. Two different IR Models based on different techniques for query expansion are shown here:

Model 1
Rocchio Expansion

For each distinct term in the k = 10 top-ranked tweets, compute the tf ×idf Rocchio scores, where tf is the frequency of the term among the 10 top-ranked tweets, and idf is the inverse document frequency of the term over the entire dataset. The top p = 5 terms are selected.



Model 2
Word2vec Expansion

Compute the cosine similarity of the query-vector with the term-vector of every distinct term in the dataset, and select those p = 5 terms for which the term vector has the highest cosine similarity with the query-vector.

Binary Independence Model

We can rank documents by their odds of relevance (ignore the common denominator)

$$O(R|\vec{x},\vec{q}) = \frac{P(R=1|\vec{x},\vec{q})}{P(R=0|\vec{x},\vec{q})} = \frac{\frac{P(R=1|\vec{q})P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|\vec{q})}}{\frac{P(R=0|\vec{q})P(\vec{x}|R=0,\vec{q})}{P(\vec{x}|\vec{q})}} = \frac{P(R=1|\vec{q})}{P(R=0|\vec{q})} \cdot \frac{P(\vec{x}|R=1,\vec{q})}{P(\vec{x}|R=0,\vec{q})}$$

Probability of a term appearing in a document relevant to the query $p_t = P(x_t = 1|R = 1,\vec{q})$
Probability of a term appearing in a nonrelevant document $u_t = P(x_t = 1|R = 0,\vec{q})$

$$O(R|\vec{q},\vec{x}) = O(R|\vec{q}) \cdot \prod_{t:x_t=q_t=1} \frac{p_t}{u_t} \cdot \prod_{t:x_t=0,q_t=1} \frac{1-p_t}{1-u_t} = O(R|\vec{q}) \cdot \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \cdot \prod_{t:q_t=1} \frac{1-p_t}{1-u_t}$$

That means that this right product is a constant for a particular query. We can rank documents by the logarithm of the middle term, called the Retrieval Status Value (RSV)

$$RSV_d = \log \prod_{t:x_t=q_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{t:x_t=q_t=1} \log \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{t:x_t=q_t=1} \log \frac{p_t}{(1-p_t)} + \log \frac{1-u_t}{u_t}$$

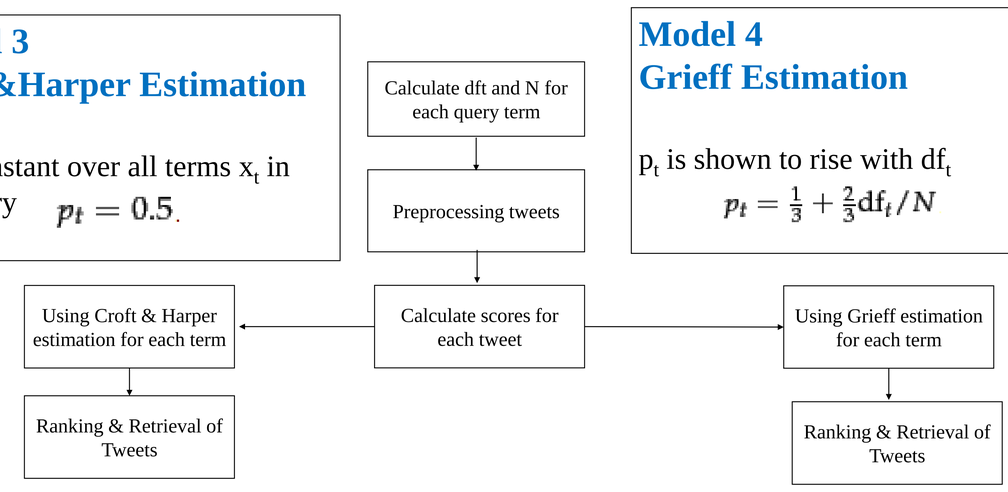
For each term t, number of documents that contain t = df_t
Relevant documents are a very small percentage of the collection, it is plausible to approximate statistics for nonrelevant documents by statistics from the whole collection.
So, Probability of term occurrence in nonrelevant documents for a query, $u_t = df_t/N$

$$\log[(1-u_t)/u_t] = \log[(N-df_t)/df_t] \approx \log N/df_t$$

$$RSV_d = \sum_{t:x_t=q_t=1} \log \frac{p_t}{(1-p_t)} + \log N/df_t$$

Model 3
Croft&Harper Estimation

p_t is constant over all terms x_t in the query $p_t = 0.5$



Model 4
Grieff Estimation

p_t is shown to rise with df_t

$$p_t = \frac{1}{3} + \frac{2}{3}df_t/N$$

Results

Comparison graphs of Precision, Recall and F-score of all the four models cutoff at k=100.

Conclusions

- Model 3 (Croft & Harper estimation) performs the best with best precision as well as good recall value. The harmonic mean of the both F-Value of the model is also better than the other models at Cutoff (k) = 100.
- The second next better model can be Model 2 (Word2Vec Expansion) or Model 4 (Grieff Estimation) with comparative scores.
- Model 1 does not give satisfactory results proven by a low precision and recall value.

Reference

1. M. Basu, K. Ghosh, S. Das, R. Dey, S. Bandyopadhyay, and S. Ghosh. 2017. Identifying Post-Disaster Resource Needs and Availabilities from Microblogs. In Proc. ASONAM.
2. M. Basu, A. Roy, K. Ghosh, S. Bandyopadhyay, and S. Ghosh. 2017. MicroblogRetrieval in a Disaster Situation: A New Test Collection for Evaluation. In Proc. Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP) co-located with European Conference on Information Retrieval. 22–31.