qqqqq

qqq and qqq

Faculty of Natural Sciences, University of Tampere, Finland
qqq@uta.fi

Abstract. qqq

1 Introduction

qqq

2 Preliminaries

Sentiment Analysis Sentiment analysis refers to the problem of assigning a quantitative positive/negative mood to a short text [1]. To extract the sentiment strength from a text, we use the SentiStrength tool [2]. SentiStrength assigns both a positive and a negative score (since users can express both types of sentiments at the same time). The score of a positive sentiment strength ranges from +1 (not positive) to +5 (extremely positive). Similarly, negative sentiment strength ranges from -1 (not negative) to -5 (extremely negative).

3 Measures

Let E denote a finite set of *hotels*, and C denote a collection of reviews for the hotels in E, covering the time period $T = [t_s, t_e]$ (where t_s, t_e are two different time points with $t_s < t_e$). Let also U be the total set of users posted these reviews.

Popularity Let $e \in E$ be a given hotel, and $T_i \subseteq T$ be a given time period. Let also $C_i \subseteq C$ be the collection of reviews posted during T_i. The popularity, with respect to the number of submitted reviews, of e during T_i equals to the percentage of reviews submitted for e during that period. Formally:

$$popularity_c(e, T_i) = \frac{|C_{e,i}|}{|C_i|} \tag{1}$$

where $C_{e,i} \subseteq C_i$ denotes the set of reviews that refer to e during T_i.

The above measure does not consider the number of different users discussing about a hotel. This means that a hotel can be very popular even if it is reviewed by only a few users (but in a large number of reviews). A more fine-grained

indication of popularity is given by the amount of different users discussing about the hotel. In that case, if $u_c \in U$ denotes the user who posted the review c, the popularity of a hotel $e \in E$ during T_i can be defined as the percentage of different *users* discussing about e during that period, i.e.:

$$popularity_u(e, T_i) = \frac{|\cup_{c \in C_{e,i}} u_c|}{|\cup_{c \in C_i} u_c|} \tag{2}$$

Sentimentality For a review $c \in C$, let $s_c^+ \in [1, 5]$ be the review's positive sentiment score and $s_c^- \in [-5, -1]$ be the review's negative sentiment score, computed based on SentiStrength. The sentimentality of c is given by $\phi_c = s_c^+ + s_c^-$, i.e., $\phi_c \in [-4, 4]$. In turn, the sentimentality of a hotel e in a time period T_i is defined as the average sentimentality of reviews submitted for e during T_i. Formally:

$$sentimentality(e, T_i) = \frac{\sum_{c \in C_{e,i}} \phi_c}{|C_{e,i}|} \tag{3}$$

Controversiality A hotel e can be considered controversial in a time period T_i, if there exist many positive and negative reviews for it. Specifically, let $C_{e,i}^+$ be the set of reviews that refer to e during T_i with strong positive sentimentality, i.e., $C_{e,i}^+ = \{c \in C_{e,i} \mid \phi_c \geq \delta\}$, where $\delta \in [0, 4]$ is a strong sentimentality threshold (e.g., $\delta = 2.0$). Likewise, let $C_{e,i}^-$ be those with strong negative sentimentality, i.e., $C_{e,i}^- = \{c \in C_{e,i} \mid \phi_c \leq -\delta\}$. Then, the controversiality of a hotel is defined as follows:

$$controversiality(e, T_i) = \frac{|C_{e,i}^+| + |C_{e,i}^-|}{|C_{e,i}|} \cdot \frac{min(|C_{e,i}^+|,\ |C_{e,i}^-|)}{max(|C_{e,i}^+|,\ |C_{e,i}^-|)} \tag{4}$$

Intuitively, a value close to 1 means that the probability of the hotel being controversial is high, since there is a big percentage of reviews with strong sentimentality (first part of the formula), and also there are both many reviews with strong positive sentimentality and many reviews with strong negative sentimentality (second part of the formula).

4 Territory Measures

Let L denote a territory consisting of a set of hotels, $L \subseteq E$. Then, the popularity of L is defined as:

$$popularity_c(L, T_i) = \frac{\sum_{e \in L} popularity_c(e, T_i)}{|L|}.$$

Similarly:

$$sentimentality(L, T_i) = \frac{\sum_{e \in L} sentimentality(e, T_i)}{|L|}.$$

$$controversiality(L, T_i) = \frac{\sum_{e \in L} controversiality(e, T_i)}{|L|}.$$

5 Conclusion

qqq

References

1. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2007.
2. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment in short strength detection informal text. *JASIST*, 61(12):2544–2558, 2010.