Milestone 4

1. We found smileys were strong indicators of sentiments. In addition, Twitter Search API does support such search keywords as “:)” and “:(“, and there are abundant tweets with smileys available. Current application is based on xxx tweets in total.

2. Inspired by the Alexander Pak and Patrick Paroubek’s paper[1] about sentiment analysis based on Twitter as a corpus, we retrieved 2-grams from two sets of tweets with same amount of positive – with :) – and negative – with :(– tweets, respectively. The occurrences of 2-grams in either set in essence suggest the possibilities of corresponding sentiment.

   $\mathbf{P(\,s \mid g\,) \sim P(\,g \mid s\,)}$ – g: gram; s: sentiment

3. We first purged tweets by removing punctuations, stopwords, @user, and URLs; we preserved #topics, since it might sometimes also indicators of certain sentiment. Then we leveraged **Hadoop MapReduce** tweets into 2-grams with occurrence counts. $P(g \mid s)$ = occurrences / # of tweet (we assume each 2-gram appeared in one tweet only once).

4. To remove noise, Pak and Paroubek’s paper introduces two concepts, namely ***entropy*** and ***salient***. By computing these two values for mutual grams, we could sift out those grams prevalent in both sets. These two values are different when there are three sentiments (positive / negative / objective), but they became equivalent as we just coped with positive and negative tweets.

5. To score and classify tweets, we sum up possibilities of 2-grams in one tweet, and a **threshold** is set to assign sentiments. When the score of one tweet is lower than the threshold, we exploited **weights** of 2-grams indicated by entropy and salient.

6. Why 2-grams? Only one word brings too much noise, 3-grams did not distinguish many tweets when the source tweets collection was small.

7. **Future work:** 1) try more source tweets; 2) try 3-grams; 3) utilize LSA to reduce grams into clusters.

[1] Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, et al., Eds.)*Computer*, *2010*(10), 1320-1326. European Language Resources Association (ELRA). doi:10.1371/journal.pone.0026624