**[This is something we have done now, and we will devote much more time next ten days to finish our target as soon as possible since we are finding intern jobs last two weeks.]**

**Data Collection:**

Two corresponding json file about reviews: review.json, tip.json.

[Review.json is the key file we used which has user\_id,business\_id,text,useful,funny,cool.]

[tip.json is a dataset of reviews with short length which has “user\_id,business\_id,text“ , and we may use the model trained in the review.json to give a judgement of tip.json].

**Data Uploaded:**

Both of them have been uploaded into the HDFS.

**Word2Vec and tf-idf code website we found：**

<https://spark.apache.org/docs/latest/ml-features.html#word2vec>

<https://spark.apache.org/docs/latest/ml-features.html#tf-idf>