北 京 邮 电 大 学

BBC6521 Project 毕业设计 2016/17

Mid-term Progress Report

中期进展情况报告

学院 School	International School	专业 Programme	Telecommunication Engineering with Management	班级 Class	2013215110
学生姓名 Student Name	Zheng Weihong	BUPT 学号 BUPT Student No.	2013213217	QM 学号 QM Student No.	130801991
设计（论文）编号 Project No.	IC_3217	电子邮件 Email	2013213217@bupt.edu.cn		
设计（论文）题目 Project Title	Fine-grained Sentiment Analysis of Product Comments Based on Product Features				

毕业设计（论文）进展情况，字数一般不少于 1000 字
The progress on the project. Total number of words is no less than 1000

目标任务: Targets set at project initiation:

(must be the same as “What I expect to have working at the mid-term oral” in the Spec)

Finish the crawler program and collect data from a representative product.

Preprocess the data to satisfy requirements of further operations.

Finish preliminary feature and opinion extraction.

是否完成目标 Targets met?	[YES/NO] YES

目前已完成任务 Finished Work:

A crawler program under the frame of scrapy has been developed, and the technique named xpath is used to resolve web pages. About 8000 Chinese comments of Kindle from Amazon have been collected and been stored in the format of json.

When preprocessing the data from Amazon, I found the fact that those comments have less emotional tendency about the Kindle product but more about how they love reading and what books they like, that is not suitable for sentiment analysis. So another crawl program has been developed to collect the comments of a T-shirt in Taobao website, the amount of the comments is about 5000.

The data preprocess has been finished to a certain degree, it should be repeated again if effecting the result of extracting features and review words. The process includes word segmentation, eliminating stopwords and Part-of-speech tagging. Following are the details:

1. Because the comment is in Chinese, the NLTK cannot be used in data preprocess as it is designed to analyse English, so Jieba(an open source package in Github) is used here to preprocess the Chinese comments.
2. In word segmentation step, I use the Sogou thesaurus as the dictionary. And the API of Jieba(jieba.cut()) uses dynamic programming to find the most probable words combination based on the word frequency in the dictionary, for unknown words, HMM-based model is used with the Viterbi algorithm.

3. In step of eliminating stopwords , the list of HIT stop words has been used as the stopword dictionary to delete stopwords in Chinese.
4. The Part-of-speech tagging has been done under the universal Chinese standard, the objective of this step is for the convenience of extraction of features and review words.

I am now adopting easy ways to deal with preliminary feature and opinion extraction, it would be evaluated and completed afterwards. Followings are finished works until now:

1. For feature extraction, referring to "Fine-grained Sentiment Analysis Of Online Reviews", I manually set a dictionary for feature words I can take into consider, such as "质量", "材质" and "款式" (means "quality", "material" and "style" respectively). And I tag those feature words with my own special tag "f", then I used API from Jieba named jieba.analyse.extract_tags() to extract features based on TF-IDF algorithm, one of its parameters, allowPOS, can help filter words with specific tags. By now, features can be extracted. However, there may exits some other expressions for a same feature and the manual dictionary cannot cover all those cases, therefore the dictionary should be extended. This is a main problem having occurred, the solution is listed afterwards in this report.
2. For opinion words extraction, referring to some papers, most of opinion words are adjectives and Sogou thesaurus have provided tags for each word, so with the tagged words, those opinion words can be extracted easily by jieba.analyse.extract_tags() as well. However, the experiment result showed that many other words, not only the wanted opinion word, have also been extracted with tag "a", so the extraction method should be improved and the improvement is still in the process.

尚需完成的任务 Work to do:

1. Evaluate the feature and opinion words extraction and complete it. The extraction algorithm should be better to cover the features and opinions as much as possible and reduce the rate of wrong extraction.
2. Build up relationship between features and opinion words. Each feature should be related to its own opinion word, so that bigrams can be formed and the analysis of emotion tendency for each feature can be continued.
3. Analyze the outcome and write a draft report. Analysis of the outcome means calculating the customers' sentiment to every features of the product such that the whole image of the product can be produced. To achieve this, the opinion words would be grouped into positive and negative in advance.
4. Complete the report.

能否按期完成设计（论文） Can finish the project on time or not:	[YES/NO] YES

存在问题 Problems:

1. Different expression of features:

In customers' comments, some may directly indicate their opinions to specific features, such as "the style is good", in these cases, the feature "style" can be easily extracted. However, for the feature "style", someone may say "it looks handsome", so the feature words cannot be extracted. This problem would influence the whole analysis of customers' sentiment.

2. Building relationship between features and opinion words:

I have tried to use NLTK to form bigrams, but the result was not very ideal. NLTK can form bigrams by calculating Chi-square value or information entropy. Perhaps it is designed to handle English so the bigrams appear to fail, many unwanted bigrams were produced such as ("质量","帅"）("quality", "handsome"), on the other hand, some desired bigrams were missed.

拟采取的办法 Solutions:

1. Referring to “Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis” published by Wenhao Zhang, features can be grouped as explicit and implicit, for explicit features, the morpheme-based method can be used to extract similar feature words. And for implicit features, a method named PMI is used to find those features. Methods in this paper are temporarily regarded as the solution to features extraction, and more papers would be read for finding useful methods.

2. My idea to solve the relationship problem is to build up an opinion words dictionary for each feature. The build of dictionaries can refer to “Study on Chinese Text Sentiment Classification”, where some rules of part of speech are used. And then, bigrams can be established without calculating any statistic value, but by the easy permutation and combination for words in each comment. Afterward, a filter would be applied to select the bigrams fulfilling the condition that its opinion word is in the dictionary of its feature. Such that the establishment of bigrams can be correct.

最终论文结构 Structure of the final report:

Title

Abstract

1. Introduction

Introduce the background of this project and what works have been done by others before.

2. Problem definition and dataset

2.1 Problem definition: Define what problems should be handled in this project, and to handle those problems, what techniques were used.

2.2 Dataset: Introduce what is the dataset, where is it from and how the dataset is used to finish this project.

3. Data process

3.1 Data preprocess

3.1.1 words segment

3.1.2 Part of speech tagging

3.1.3 eliminating stop words

3.2 Establishment of bigrams

3.2.1 Features extraction

3.2.2 Opinion words extraction

3.2.3 Estanbish bigrams

4. Sentiment analysis

Introduce the method of calculating the sentiment.

5. Project result & Conclusion

6. Reference

日期 Date:	**Mar. 5th , 2017**