

**Finding Travellers’ Interest in Darwin by Using text mining**

**Student: Chao Liang (S304935)**

**Supervisor: Jerry Shen**

Contents

[Abstract 3](#_Toc515552100)

[1.Introduction 4](#_Toc515552101)

[2. Topic Detection and Mining User-Generated Content 6](#_Toc515552102)

[2.1. Mining User-Generated Content 6](#_Toc515552103)

[2.2 Text Mining and Topic Detection 7](#_Toc515552104)

[3.The Text-Mining Methodology 8](#_Toc515552105)

[3.1 The Text-Mining Apparatus 8](#_Toc515552106)

[3.2 Data Cleaning Approaches 10](#_Toc515552107)

[3.3 TF-IDF 10](#_Toc515552108)

[4. Empirical Application 12](#_Toc515552109)

[4.1 Darwin Travel Forum 12](#_Toc515552110)

[4.2 Collecting Raw Text 13](#_Toc515552111)

[4.3 Data cleaning process 14](#_Toc515552112)

[4.4 The Major Topics in the Darwin Travel Forum 23](#_Toc515552113)

[4.5 Robustness Check 26](#_Toc515552114)

[5. Conclusion 27](#_Toc515552115)

[6. References 28](#_Toc515552116)

# Abstract

As the world is increasingly globalised, tourism is becoming a booming industry. It is therefore important to understand the needs and interests of travellers. In this paper, we use text-mining approach to explore user-generated content and analyse what travellers want. Our objective is to user the user-generated data to get an insight of which of Darwin's most popular attractions, and what's the favourite topic for visitors to Darwin. The difficulty in analysing user-generated data is that the quality of user-generated content sometimes can be low. To overcome these difficulties, we apply some data cleaning approaches and improve the quality of user-generated content. We demonstrate these approach by comparing the analysis result that used data cleaning approaches with the analysis result without using data cleaning approaches in order to improve validity and get meaningful insights from the user-generated content.

Key words: text mining; user-generated content; data cleaning; tourism

# 1.Introduction

There is a huge amount of user information on the Internet that contains users' opinions and preferences about things. This information can be find on every corner of Internet, such as social media, blogs and forums. This kind of user-generated content can give the government and companies a choice to find out what tourists really want and use the information to develop tourist attractions and do business decision making (Urban and Hauser, 2004). By exploring user-generated content, the government and companies can have a deep understanding of what attract people to visit the particular tourist attraction.

In recently, more and more commercial research is starting to attach importance to text mining. However, the quality of these user-generated data remains poor, so it is not easy to analyse meaningful results. The poor quality of data causes several difficulties for data analysis process: First, the data from the Internet is so big that it's hard to screen out useful information. Second, some data are meaningless and unstructured, which make getting usable information from the dataset become more difficult. In this paper, we use data cleaning and text mining approaches to address these problems.

Our objective is to explore the user-generated context on the Internet to give the government and companies a deep understanding about which topic travellers are discussing about Darwin. In order to perform a reliable and useful result, we also use some data cleaning approaches to improve data quality. we first clean the data by using some data cleaning methods, like removing portions of the data that are not text. Stemming and lemmatization.

In order to get an insight of the dataset’s topic, we utilise term frequency-inverse document frequency(TFIDF) approach (Ramos, 2003). We use these techniques to explore the textual data and find meaningful knowledge to help the government and companies to do decision making about local tourism.

The government and companies can use these approaches as a tool to understand travellers’ interest so as to generate more efficient marketing strategies to attract more travellers. We compare the insight we obtained by using data cleaning approaches to those got from data without using any data cleaning approaches. The comparison suggests that the data that used data cleaning approaches perform better than the data without data cleaning process. In the same time, we also compare the effect of different approaches. We use user-generated data that gathered from Darwin travel forum of Tripadvisor and find the main topics that travellers are discussing by using TFIDF approach to help the government and companies to make development decision.

In what follows, we describe the research strategies and material are used. In part 3, we explain the approaches we used. In the part 4, we compare the effect of different data cleaning approach and demonstrate the use of TFIDF approach in topic detection of Darwin travel forum. We conclude with a summary of the potential and limitation of the TFIDF approach.

# 2. Topic Detection and Mining User-Generated Content

## 2.1. Mining User-Generated Content

One can get massive of data about users’ personal information and opinion about particular things by looking into user-generated content. Academic and commercial organizations are paying great attention on the development and research about user-generated content (Liu, Karahanna and Watson, 2011). Forums, blogs and social medias give people a good opportunity to express themselves. The amount of user-generated data has increased dramatically in recent years.

Researchers and businessmen have found the value of user-generated content and try to use these data to understand people’s mind. Consumer-generated data can be converted into useful knowledge and information, which can inform the government or the company about people’ requirements and business opportunities (Girardin, Calabrese, Dal, Ratti and Blat, 2008). But there are some difficulties related with mining user-generated information. One of the major difficulty is that most of user-generated data are unstructured. It is hard to analyse the meaningful information from raw user-generated content (Fader and Winer, 2012). Many researchers have found that collection of user-generated content can be time-consuming and reliability of the data is low. Unlike many structured data researchers got from traditional data source, consumer-generated data can be very unreliable and meaningless. It is inability to use these data to find some insights about users’ opinion. Therefore, the data cleaning process is becoming more and more important in mining user-generated content. Although there are massive amounts of data on the Internet, we need to screen out the essence from these content of online text.

Our objective is to find what tourists are discussing about Darwin by analysing data from online forum. In order to achieve this objective, we need to know the tourism industry of Darwin and find the relationship between the tourist attractions. We use some text-mining techniques to help us to do it.

## 2.2 Text Mining and Topic Detection

Text mining is a process to extract the effective, novel, useful, understandable and valuable knowledge scattered in text files, and use this knowledge to better organize information (Tan, 1999). With the increase of text files stored inside the company or online, the academic and commercial world has found the opportunities of exploring textual data and started to use text-mining techniques to analyse these textual data. Text mining are used in many areas. For example, in hotel industry, they used text mining techniques to predict customers’ satisfaction level about the hotel (Lau, Lee and Ho, 2005). Our paper use text mining to find the main topic that travellers in Darwin discuss about and use this information to help the government and the companies in Darwin make better development decisions. We compare different user-generated content with different structure and quality so as to improve the reliability of our result and provide validation of our approach.

# 3.The Text-Mining Methodology

Our objective is to mine travellers’ comments and reviews about traveling in Darwin and find out the main topic travellers discuss about. To achieve this goal, we use some text-mining approaches that are developed to solve the difficulties we face in mining Darwin travel forum. In this paper we give the general overview.

## 3.1 The Text-Mining Apparatus

Extracting main topics from travellers’ discussions in Darwin travel forum involves four main steps:

Step1. Data collection: The web pages are crawled from Darwin travel forum in HTML format.

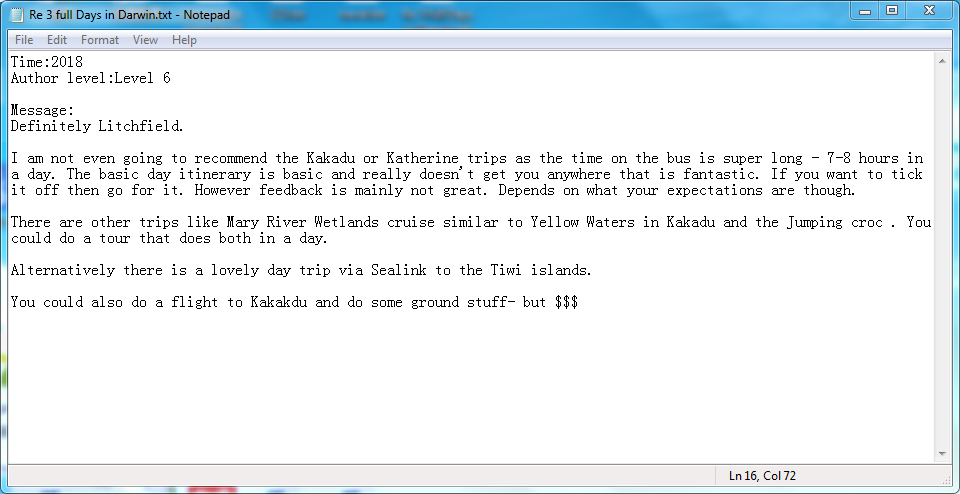
Step2. Data cleaning: There is a lot of unnecessary information in the web page, such as ads, navigation bars, HTML, JavaScript code, comments, etc., which are not meaningful for our analysis. They can be cleaned from the crawled files. If the text extraction is required, the text can be extracted by the use of tag label, density determination and other strategies.

Step3. Data chunking: We divided the text into several parts based on the date the post is created and message author’s level.

Step4. Identification of major topics: Main topics of the forum are extracted from user-generated content. By understanding main topics of the forum, the government and companies can develop tourist attractions more efficiently.

**Figure 1 A message crawled from the Forum**

**Tripadvisor.com.au**



Source: <https://www.tripadvisor.com.au/ShowTopic-g255066-i1010-k11543188-3_full_Days_in_Darwin-Darwin_Top_End_Northern_Territory.html>

Figure 1 shows a message crawled form the forum. The data we collected from this forum use in our empirical application. During the whole process, the most time-consuming part is data cleaning. Some contents of textual data are ambiguous. For instance, in figure 1, the author use “$$$” represent “money”. Without using appropriate data cleaning approaches, the raw textual data cannot get reliable information. The more details of data cleaning process are in the following part. There are several approaches to detect main topics, like topic models and TFIDF (AlSumait, Barbará, and Domeniconi, 2008). Our paper focus on using TFIDF to find main topics in the forum. By using TFIDF approach, we can understand what travellers feel interesting and help the government and companies in Darwin do decision making. We evaluate the accuracy of topic detection by using human reviewing random messages that come from test set.

## 3.2 Data Cleaning Approaches

During text mining process, we use five data cleaning approaches:

Approach 1. Removing the non-text part of textual data: This method is aimed at the corpus data collected by the crawler. As there are many HTML tags in the crawler content, it needs to be removed. Small amounts of non-text content can be deleted directly with Python's regular expression, while complex ones can be removed with Beautifulsoup framework. There are also special non-alpha characters that can be deleted with Python's regular expression (Jonathan, 2018).

Approach 2. Spelling correction: Because there may be spelling errors in English text, spell checking is generally required. If we are sure that the text we are analysing has no spelling problems, we can proceed to the next step.

Approach 3. Stemming and lemmatization: Both have something in common, which is to find the original form of the word. But stemming can get a non-word term. For example, the stem of "analysing" might get "analys", instead of “analyse”. By contrast, lemmatization only process those words that can be converted into original type (George, and Whitlock, 2006).

Approach 4. Changing to lowercase: Since English words are case sensitive, it is generally necessary to convert all words to lowercase.

Approach 5. Using stop words: There are a lot of invalid word in the English text, such as "a", "to", and also some punctuation, which we don't want to be introduced in the text analysis, so we need to take out. These words are the stop words.

## 3.3 TF-IDF

Term frequency (TF) refers to the number of times a given word appears in the file. This number is usually normalized (usually the word frequency divided by the number of words in the article) to prevent it from favouring long files. Because the same word may have a higher word frequency in a long document than a short one, regardless of its importance (Hans, Mikhael and Derwin, 2016). For the word ti in a particular document (The numerator is the number of times the word appears in the file, and the denominator is the sum of all the words in the file.), its importance can be expressed as:

However, it is important to note that some common words do not have much effect on the file, but some words with less frequency can express the topic of the file, so only use of TF is inappropriate. The weight is needed. The stronger the ability of a word to predict a topic is, the greater its weight is, and vice versa. In all the files, some words appear only in a few of files, but these words have a great effect on the topic of the file (Aizawa, 2003). the weight of these words should be greatly designed. IDF is doing this kind of work. The main idea of IDF is that the smaller the number of documents containing the word is, the larger the IDF is. It indicates that the word has good classification ability. The IDF of a particular word can be obtained by dividing the total number of files by the number of files containing the word, and then taking the logarithm of the resulting quotient.

In order to get more precise information, many researchers try to combine TF and IDF together. TF-IDF (term frequency - inverse-document frequency) is a commonly used weighted technique for information retrieval and information exploration (Ramos, 2003). The importance of a word increases with the number of times it appears in the document, but decreases inversely with the frequency it appears in the set of documents.

In the section four, we test the reliability of our analysis and describe the application of TFIDF approach.

# 4. Empirical Application

By using the text-mining process we described in part three, we convert unstructured user-generated content into reliable dataset. We use TFIDF approach to find the main topics that travellers are discussing about Darwin. We demonstrate this process using empirical application: traveller forum about tourist attractions in Darwin.

## 4.1 Darwin Travel Forum

The first step of text mining process is crawling textual data from website. We crawled data on advises about traveling in Darwin from Darwin travel forum on Tripadvisor.com.au. on May 20,2018. The Darwin travel forum included 1,606 posts consisting of nearly 20 thousand sentences created by almost 2 hundred unique users from 2004 to 2018. We focus on analysing the major topics of these posts. Within the post, we pay attention on TFIDF value of the key words. This dataset contains abundant information from real travellers in Darwin about their opinion about their travel destinations. However, the messages are created by both professionals and laymen, which makes finding major topics of these posts become more challenging.

**Table 1 Characteristics of the Darwin Travel Forum on Tripadvisor.com.au**

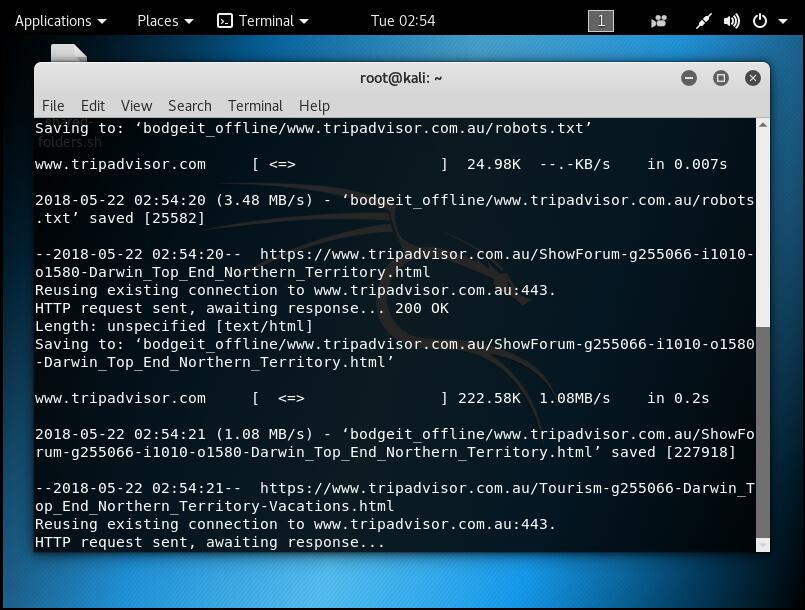
|  |  |
| --- | --- |
| No. of posts | 1,606 |
| No. of messages | 5,268 |
| No. of sentences | 19,878 |
| No. of users | 178 |

## 4.2 Collecting Raw Text

According to data analytics lifecycle, collection of dataset is the first step. We find the research objective and choose the right data source. In this paper, we collect user-generated content from Darwin travel forum by using Web Crawler Technology (Thelwall, 2001).

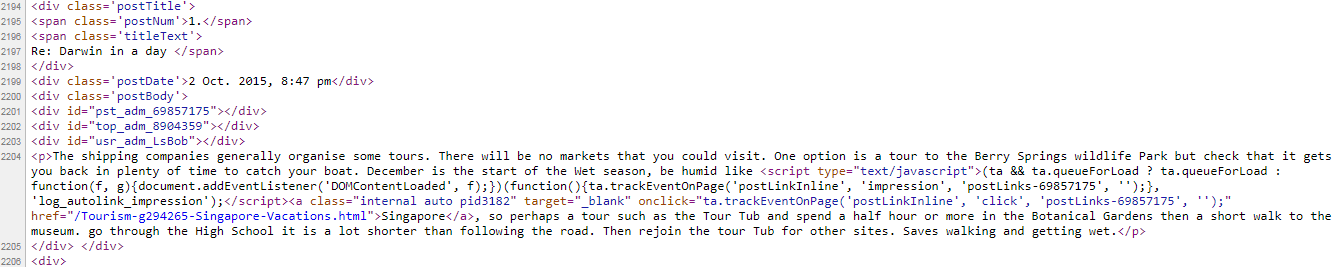
The crawl process:

**Figure 2 Using Curl and Wget to crawl webpages on Tripadvisor.com.au**



The data we collected is HTML format. The user-generated content we need is shown next.

**Figure 3 user’s reply for a post**



## 4.3 Data cleaning process

Before we begin to analyse the user-generated data, we clean the raw textual data to make sure the quality and usability of the dataset. Text pre-processing is an indispensable part of text mining. First of all, the format of raw data is usually HTML format. We need to remove unnecessary non-text parts. In many cases, our pre-processing should include spell checking, such as "Helo". We can't find the mistake in the analysis. So it needs to be corrected before processing (Schierle, Schulz and Ackermann, 2008). Thirdly, stemming and lemmatization are also very important. This is mainly due to the singular, plural and various tenses in English, resulting in a different form of word. For example, "wolf" and "wolves", they actually describe the same object. The fourth point is changing words to lowercase. As English words are written in uppercase and lowercase, we expect the word "World" and "world" to be the same word. Therefore, it is generally necessary to convert all words into lowercase. We also use the stop-words. There are many meaningless words in English text, such as "on", "the”, which we do not want to introduce in text analysis, so we need to remove these words. The TF-IDF value we calculate in this part is from the set of messages in 2018. F value is the number of times term appears in the whole dataset.

**Table 2 The F of Raw materials**

|  |  |  |
| --- | --- | --- |
| Return Position | Term | F |
| 1 | '' | 26210495 |
| 2 | , | 11906133 |
| 3 | : | 6741674 |
| 4 | > | 6638154 |
| 5 | < | 6620528 |
| 6 | ) | 6235751 |
| 7 | ( | 6235113 |
| 8 | ' | 5207681 |
| 9 | ; | 5188418 |
| 10 | } | 3281376 |

From Table 2, we can see that 10 most frequently appeared terms are some punctuations. They are not related with travellers’ messages and opinions, so it is important to remove non-textual parts from the dataset.

Because the textual data we need is inside the HTML pages, we use Beautiful Soup framework to remove non-textual parts. Beautiful Soup framework can allow us to track HTML tags and extract text from a specific HTML tag (Yih, Goodman and Carvalho, 2006).

The code that follows shows how to remove non-textual pars form HTML pages of Darwin travel forum.

htmlpage = htmlfile.read()  
  
soup = BeautifulSoup(htmlpage, "html.parser")  
print(soup.title.string)  
  
  
  
level = soup.findAll(name='div', attrs={"class": re.compile(r'^levelBadge badge')})  
post = soup.findAll(name='div', attrs={"class": "postBody"})  
date = soup.findAll(name='div', attrs={"class": "postDate"})[0].text  
date = date[date.find(",")-4:date.find(",")]  
it += 1  
print(it)  
for a in range(len(post)):  
 if a > 0:  
 t=post[a].select('p')  
 if len(post)-len(level)>1:  
 break  
 if len(post)-len(level)==1:  
 l=level[a-1].attrs["class"]  
 else:  
 l=level[a].attrs["class"]  
 n = ""  
 for c in l:  
 n = n+c  
 filepath = os.path.join(localpath, date)  
 if os.path.exists(filepath) == False:  
 os.makedirs(filepath)  
 fp = open(filepath + n + '.txt', 'a', encoding='utf-8')  
 for b in range(len(t)):  
 section\_text = t[b].text  
 sentence = re.sub("[\s+\.\!\/\_,$%^\*(+\"\')]+|[+——()?【】“”！，。？、~@#￥%……&\*（）]+", " ", section\_text)  
 remove\_punctuation\_map = dict((ord(char), None) for char in string.punctuation)  
 sentence = sentence.translate(remove\_punctuation\_map)  
 # sentence= bytes(sentence, 'utf-8')  
 # encode\_type = chardet.detect(sentence)  
 # if encode\_type['encoding'] != None:  
 # sentence = sentence.decode(encode\_type['encoding'])  
 # sentence=str(sentence)  
 fp.write(sentence)  
 fp.write(","+"\n")

Once we remove all the non-textual parts, we can focus on textual data analyse and try to find the insights from the data.

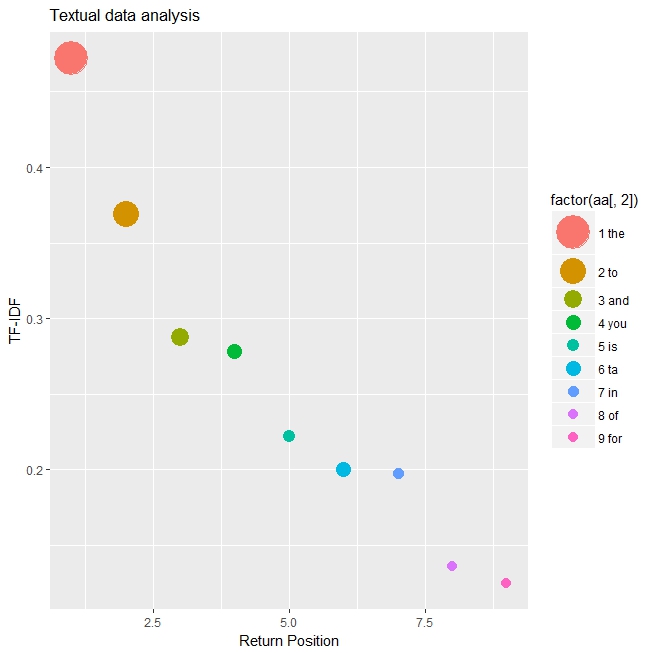
The code that follows shows how to calculate TF-IDF values.

path=path+"\\"  
corpus = []   
for ff in filelist:  
 fname = path + ff  
 f = open(fname, 'r+',encoding='utf-8')  
 content = f.read()  
 f.close()  
 corpus.append(content)  
  
vectorizer = CountVectorizer()  
transformer = TfidfTransformer()  
tfidf = transformer.fit\_transform(vectorizer.fit\_transform(corpus))  
  
word = vectorizer.get\_feature\_names()   
weight = tfidf.toarray()   
  
sFilePath = 'D:/tfidffile'  
if not os.path.exists(sFilePath):  
 os.mkdir(sFilePath)

**Table 3 The F and TF-IDF of textual data.**

|  |  |  |  |
| --- | --- | --- | --- |
| Return Position | Term | F | TF-IDF |
| 1 | the | 23309 | 0.472166025 |
| 2 | to | 17407 | 0.368669029 |
| 3 | and | 12123 | 0.287652248 |
| 4 | you | 10167 | 0.27801271 |
| 5 | is | 7734 | 0.222362757 |
| 6 | ta | 9865 | 0.200599594 |
| 7 | in | 7320 | 0.197429316 |
| 8 | of | 6509 | 0.136256328 |
| 9 | for | 6179 | 0.124901634 |

**Figure 4 Bubble Diagram for Textual Data (The X-axis represents the rank of the word, the Y-axis represents the TF-IDF value, and the bubble size represents the F value)**

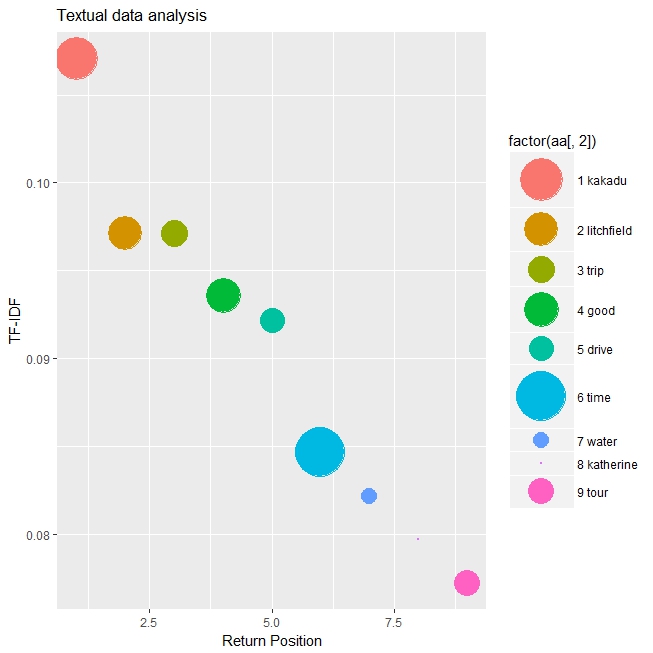


As Table 3 and Figure 4 show first nine terms with highest TF-IDF. These terms are meaningless and useless to our main topic analysis. Therefore, we need to introduce stop words to remove those meaningless words. We use scikit-learn framework for feature processing and introduce an array as a stop-word table with the stop\_words parameter (Agarwal, Xie, Vovsha, Rambow, and Passonneau, 2011).

**Table 4 The F and TF-IDF of textual data that removed stop words**

|  |  |  |  |
| --- | --- | --- | --- |
| Return Position | Term | F | TF-IDF |
| 1 | kakadu | 1495 | 0.107088513 |
| 2 | litchfield | 1164 | 0.097126791 |
| 3 | trip | 928 | 0.097126791 |
| 4 | good | 1200 | 0.093577846 |
| 5 | drive | 844 | 0.09214593 |
| 6 | time | 1713 | 0.084674638 |
| 7 | water | 542 | 0.082184208 |
| 8 | katherine | 22 | 0.079693777 |
| 9 | tour | 880 | 0.077203347 |

**Figure 5 Bubble Diagram for Textual Data that removed stop words**



As we can see from Table 4 and Figure 5, we can extract meaningful words from the textual data that removed stop words. But it is also should be noted that every word has various tenses and the singular and plural. English words also have lowercase and uppercase and there are some spell errors. For example, "like", "liking" and "likes", they actually describe the same thing. Therefore, we transfer different forms of the word into the one form. In this paper, we compare the usability about Stemming and Lemmatization in finding the original form of the word.

**Table 5 The F and TF-IDF of textual data that use Stemming**

|  |  |  |  |
| --- | --- | --- | --- |
| Return Position | Term | F | TF-IDF |
| 1 | day | 2561 | 0.182099445 |
| 2 | darwin | 3384 | 0.145970788 |
| 3 | tour | 1609 | 0.114115652 |
| 4 | drive | 1307 | 0.109259667 |
| 5 | trip | 1104 | 0.104403682 |
| 6 | kakadu | 1607 | 0.101975689 |
| 7 | litchfield | 1217 | 0.094691712 |
| 8 | water | 791 | 0.087407734 |
| 9 | time | 1882 | 0.082551749 |
| 10 | car | 1063 | 0.080123756 |
| 11 | katherin | 851 | 0.077695763 |
| 12 | park | 1108 | 0.067983793 |
| 13 | croc | 430 | 0.063127808 |
| 14 | bu | 445 | 0.053415837 |
| 15 | night | 979 | 0.053415837 |

**Table 6 The F and TF-IDF of textual data that use Lemmatization**

|  |  |  |  |
| --- | --- | --- | --- |
| Return Position | Term | F | TF-IDF |
| 1 | day | 2561 | 0.185989191 |
| 2 | darwin | 3384 | 0.149088806 |
| 3 | tour | 1567 | 0.116553227 |
| 4 | trip | 1102 | 0.106633803 |
| 5 | kakadu | 1607 | 0.104153947 |
| 6 | litchfield | 1217 | 0.09671438 |
| 7 | drive | 905 | 0.091754668 |
| 8 | good | 1242 | 0.090850991 |
| 9 | time | 1863 | 0.0843151 |
| 10 | water | 787 | 0.0843151 |
| 11 | car | 1063 | 0.081835244 |
| 12 | katherine | 848 | 0.079355388 |
| 13 | park | 1065 | 0.069435965 |
| 14 | bus | 546 | 0.061996397 |
| 15 | road | 953 | 0.059516541 |

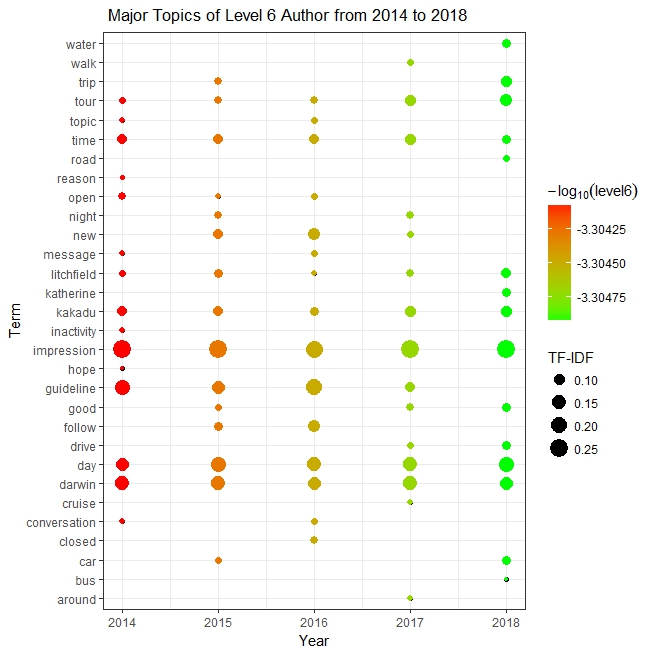
By observing Table 5 and Table 6, we can see that many words’ TF-IDF value increased after Stemming and Lemmatization, which make us can find deep insights from this dataset and understand travellers’ worry and concern about Darwin. The drawback of Stemming is that some words are converted into some meaningless strings, like the thirteenth word and the fourteenth word in Table 5. The reason is because the function of Stemming is to shorten similar words, regardless of their structure and spelling. By contrast, Lemmatization only transfer words into their original form and did not break the structure of the words.

Comparatively speaking, Stemming is a simple and lightweight form of word merging. The result of the final word is not necessarily practical. Lemmatization is relatively complex, and the result is the original form of the word, which can carry a certain meaning and has more research and application value than Stemming (Balakrishnan and Lloyd-Yemoh, 2014). Therefore, in this paper, we use Lemmatization to optimize our dataset.

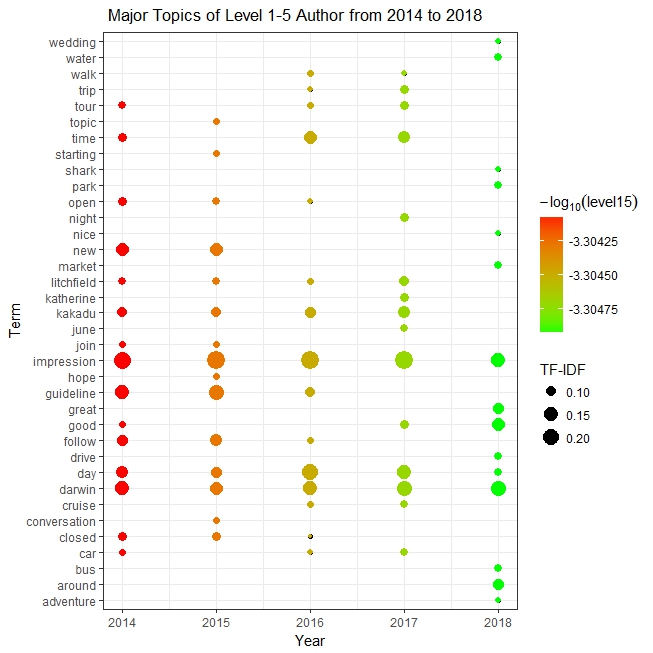
## 4.4 The Major Topics in the Darwin Travel Forum

We begin with the analysis of the TF-IDF of different words in the user-generated content. The more frequently travellers mention these words, the more these words are related with the major topics of this forum. We separate user-generated content into several documents according to the creation year from 2014 to 2018 and the author’s level. Because most of posts are created by level 6 author, so we separate posts to two parts, level 6 and level 1-5.

**Figure 6 Five years on posts of level 6 authors**



**Figure 7 Five years on posts of level 1-5 authors**



From Figure 6 and Figure 7, we can see that both professional travellers and amateur travellers discuss a lot about the impression about Darwin and also pay close attention on the time, namely how many days they will spend on Darwin. In term of tourist attractions, Kakadu and Litchfield are most popular destination for both professional travellers and amateur travellers. Katherine start attracting amateur travellers from 2017, while Katherine become very popular around professional travellers from 2018. But all three tourist attractions cannot attract new travellers’ attention since 2018. Compared with professional tourists, amateurs asked a lot about when the place is open or closed because they are less familiar with Darwin. It is also should note that both professionals and amateurs are looking for guideline before 2018, which means that there are not enough instructions about traveling in Darwin. From our opinion, the government can try to create an application to help new travellers who come to Darwin at first time. The transportation is another main topic that travellers are concerned about. “drive”, “car” and are frequently mentioned in recent years. “bus” are mentioned since 2018, which means that travellers start thinking public transportation as a way of travel. In recent years, travels also start talking about “water” and “cruise”. It shows an opportunity to exploit the tourism of the sea and islands.

## 4.5 Robustness Check

How many messages should be included in the Dataset? The Darwin travel forum we downloaded has 5,268 messages. However, it is possible that we may have sparser data for other researches. In order to check the sensitivity of the result to the dataset size, we use 1/8, ¼, ½ of the total dataset to calculate TF-IDF of the user-generated content and find the insights.

**Table 7 TF-IDF value for five terms from different sizes of dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Term Size | 1/8 dataset | 1/4 dataset | 1/2 dataset | 1 dataset |
| impression | 0.208456664 | 0.274828706 | 0.270867314 | 0.256144491 |
| day | 0.185294812 | 0.130870812 | 0.183233771 | 0.18123431 |
| darwin | 0.138971109 | 0.157044975 | 0.159333714 | 0.154653278 |
| tour | 0.150552035 | 0.091609569 | 0.099583571 | 0.113573501 |
| trip | 0.104228332 | 0.085066028 | 0.087633543 | 0.103907671 |

From Table 7, we can see that even with only 1/8 dataset, the key words of main topics still get high TF-IDF value and remain in the top ten. And other analysis results are extremely similar to the result of full dataset. As expected, some special words, like tourist attractions’ name, are relatively sensitive to dataset size. This analysis shows that the main topic analysis using TF-IDF is robust to the size of dataset.

# 5. Conclusion

In this paper, we propose a hypothesis that the government and companies in Darwin can find and understand travellers’ needs from the Web and use the information to help do decision making. We use text-mining approach and TF-IDF to solve the difficulties and use data cleaning techniques to improve reliability of user-generated content. We demonstrate the value of these approaches by empirical application. We find out tourist attractions that travellers are interested in and travellers’ concerns from analysing the dataset. These insights can help the government and local companies to manage infrastructures and find business opportunities.

It is also should be noted that some terms may not have high F value, but they have high TF-IDF value. This means that these words are highly related with the main topic. For further research, we can try add some weights to some specific words, so as to get deeper insights from the traveller-generated content. And we also can download data from other forums and compare their main topics so as to increase the reliability of our analysis result.

# 6. References

Urban, G. L., & Hauser, J. R. (2004). “Listening in” to find and explore new combinations of customer needs. Journal of Marketing, 68(2), 72-87.

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, pp. 133-142).

Liu, Q., Karahanna, E., & Watson, R. (2011). Unveiling user-generated content: Designing websites to best present customer reviews. Business Horizons, 54(3), 231-240.

Fader, P., & Winer, R. (2012). Introduction to the Special Issue on the Emergence and Impact of User-Generated Content. Marketing Science, 31(3), 369-371.

Girardin, F., Calabrese, F., Dal Fiore, F., Ratti, C., & Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. IEEE Pervasive computing, 7(4).

Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases (Vol. 8, pp. 65-70).

Lau, K. N., Lee, K. H., & Ho, Y. (2005). Text mining for the hotel industry. Cornell Hotel and Restaurant Administration Quarterly, 46(3), 344-362.

AlSumait, L., Barbará, D., & Domeniconi, C. (2008, December). On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 3-12). IEEE.

Jonathan E. Germann. (2018). Approaching the largest ‘API’: Extracting information from the Internet with Python. Code4Lib Journal, (39), Code4Lib Journal, 01 February 2018, Issue 39.

George, R., & Whitlock, John. (2006). Scaling the Technology Opportunity Analysis Text Data Mining Methodology: Data Extraction, Cleaning, Online Analytical Processing Analysis, and Reporting of Large Multi-source Datasets, ProQuest Dissertations and Theses.

Hans Christian, Mikhael Pramodana Agus, & Derwin Suhartono. (2016). Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). ComTech, 7(4), 285-294.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), 45-65.

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, pp. 133-142).

Thelwall, M. (2001). A web crawler design for data mining. Journal of Information Science, 27(5), 319-325.

Schierle, M., Schulz, S., & Ackermann, M. (2008). From spelling correction to text cleaning–using context information. In Data Analysis, Machine Learning and Applications (pp. 397-404). Springer, Berlin, Heidelberg.

Yih, W. T., Goodman, J., & Carvalho, V. R. (2006, May). Finding advertising keywords on web pages. In Proceedings of the 15th international conference on World Wide Web (pp. 213-222). ACM.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38). Association for Computational Linguistics.

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances. Lecture Notes on Software Engineering, 2(3), 262.