Topic Modelling of Patient Opinion

A minor thesis submitted in partial fulfilment of the requirements for the degree of

Masters of Computer Science

Bin Lu

School of Computer Science and Information Technology

Science, Engineering, and Technology Portfolio,

Royal Melbourne Institute of Technology

Melbourne, Victoria, Australia

October 26, 2014

Declaration

This thesis contains work that has not been submitted previously, in whole or in part, for any other academic award and is solely my original research, except where acknowledged.

This work has been carried out since March 2014, under the supervision of Dr Jenny Zhang, Dr Amanda Kimpton, Dr Daryl D'Souza.

Bin Lu

School of Computer Science and Information Technology

Royal Melbourne Institute of Technology

October 26, 2014

Acknowledgements

TODO:THANKS!

Contents

List of Figures

List of Tables

Abstract

Topic models have been widely used to identify topics in text collections. Many studies tried to fit variety of data with existing model or extension of existing model. Since each type of the data has its own characteristic, it is not possible to find a solution to suite all. Current studies cover consumer reviews, blogs, or even twitter. It left few missing pieces, for example healthcare reviews. We study the data from a website called Patient Opinion, it's unique data structure gives us advantage to improve the topic modelling result without modifying baseline model. To achieve this objective, particular fields from user input from Patient Opinion are retrieved, this collection of terms is used to filter out noise from baseline model. Also each term in topic will get ranked by *tf-idf* score in the scope of result topic. Our evaluation demonstrate the effectiveness of topic coherence improvement by reducing the noise. The effectiveness of ranking is overlooked due to the limitation of evaluation method.

Chapter 1

Introduction

Publicly available opinions and service feedback provide valuable information for decision making for both service providers and consumers. With the help of websites, blogs, forums and social networks, it is never been so easy to express opinions and leave feedback. Analyzing the opinions becomes a challenge, not just because of the quantity of the data, most opinion from general users are free form text. The massive quantity of the data wont be effectively used until there is a systematically approach of analyzing and summarizing, in this project we focus on topic modelling side, aiming to discover a set of terms that can form a topic, hence with the topics the collection of document can be easily categorized or summarized. Many techniques have been proposed to solve this problem. Most previous studies focus on analyzing product reviews. We are interested to discover a model that suite service reviews. More specifically, reviews relate to healthcare. Study shows the effective governance is increasingly recognized as pivotal to improvements in healthcare quality (Bismark and Studdert [2013]), moreover current issue of effectiveness of the authority is affected by insufficient resource and inadequate information received (Bismark et al. [2013]). The object we are going to study is www.patientopinion.org.au, it is a publicly available healthcare forum. It allows user to post their own healthcare related story, the story can be positive or negative or a bit from both side. Although the story body is free form text, user still has to follow a certain template while submit the story. There is a unique feature of the data from Patient Opinion, user could specify the key word while submitting the story, which we could treat as pre-defined terms for topics, and they will be used weight the terms that generate by the topic model algorithm.

MDK-LDA model proposed by Chen (Chen et al. [2013]) , the method extends the Latent Dirichlet Allocation (Blei et al. [2003]), the later one becoming the standard method in topic modelling and been extended in variety ways. The basic idea of LDA is treat each document in a collection as a vector of word count, each document is represented as a probability distribution over a number of topics, while each topic is represented as a probability distribution over a number of words. MDK-LDA introduces a new latent variable s in LDA to model s-sets. Each document is an admixture of latent topics while each topic is a probability distribution over s-sets. Another approach is Aspect-based Summarization (Garcia-Moya.L and Berlanga-Llavori.R [2013]), it is usually composed of three main tasks: aspect identification, sentiment classification, and aspect rating. Generally this model is used to analyzing product review, it is designed to effectively retrieve features and sentiment for products.

Due to the unique characteristic of the data from Patient Opinion, we could improve existing algorithm with the additional information from the data set. LDA has been approved a very effective model, and been used as a based model in many topic modelling studies. We choose LDA as our base model, and incorporate unique feature in Patient Opinion, specifically the section of Whats Good and What could be improved. These two sections are filled in by user while submitting the story, the template is provided by the website. Generally this will be the main topic or features user want to give feedback about in the story. And we assume user labelled story 100

1.1 Research Question

- How to use user specified key words to improve the performance and accuracy in topic modelling.

1.2 Research Contribution

The project has made the following contribution to the field of topic modelling by using the LDA as base framework: By introducing the user specified terms, the number of term that form each topic could be reduced significantly while retain the quality of the topic.

Chapter 2

Related Works

2.1 Topic Modelling

Also known as Latent Dirichlet Allocation or discrete PCA is a Bayesian graphical model for text document collections represented by bags-of-words (Newman et al. [2009], Blei et al. [2003], Griffiths and Steyvers [2004], Buntine and Jakulin [2004]). The model allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. Generally, only a small number of words have high likelihood in each topic and each document only presents certain number of topics. Following is the equation of collapsed Gibbs sampling:

$$p(z_{id} = t \mid x_{id} = w, Z^{\neg id})\alpha \tag{2.1}$$

$$\frac{N_{wt}^{\neg id} + \beta}{\sum_w N_{wt}^{\neg id} + W\beta} \frac{N_{wt}^{\neg id} + \alpha}{\sum_w N_{wt}^{\neg id} + T\alpha} \tag{2.2}$$

where $z_{id} = t$ assigns topic t with i^{th} word in document d, and word w currently observed indicated by $x_{id} = w$. $Z^{\neg id}$ is the vector of all topic assignments not including the current word. N_{wt} represent integer count arrays, and α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the per-topic word distribution.

2.2 Multi-Domain Prior Knowledge in Topic Modelling

As mentioned before, LDA is a powerful topic modelling framework, however recent studies found that these unsupervised models may not produce topics that conform to the user's existing knowledge(Chen et al. [2013]). Chen et al (Chen et al. [2013]) proposed a novel knowledge-based model, called MDK-LDA, which is capable of using prior knowledge from multiple domains to help topic modelling in the new domain. A new latent variable 's' is added to model the s-set, each document represent admixture of latent topics while each topic is a probability distribution over s-set. MDK-LDA uses s-set to distinguish topics in multiple senses. For example the word light can be represented by two s-set: S1 {light, heavy, weight} and S2 {light, bright, luminance}, if light co-occurs with bright or luminance it will be assigned to S2. In conclusion, MDK-LDA outperforms base LDA model, in other word, the extra information improved the quality of result. ccurs with bright or luminance it will be assigned to s-set light, bright, luminance.

Chapter 3

Improve topic coherence by using user input

Although LDA provides a powerful framework for extracting latent topics in text document, but sometimes learned topics are lists of words that do not convey much useful information (Newman et al. [2009]). Some extrinsic evaluation has been used to demonstrate the effectiveness of the learned topic in the application domain, but standardly, no attempt has been made to perform intrinsic evaluation of the topics themselves, either qualitatively or quantitatively (Newman et al. [2010]). To solve the problem, base LDA model had been extended either by incorporating human judgement in to the model-learning framework or creating a computational proxy that simulates human judgements (Chang et al. [2009]), for example the MDK-LDA model (Chen et al. [2013]) we introduced in section 2. Due to the unique characteristic of the data of Patient Opinion, we use user input to simulate human judgement, hence to produce a better quality topic modelling result.

3.1 Description of Data

Data from Patient Opinion contains many informations, however we only interested in few parts of them in our project Figure 1: 1) The title of the story, its the summary of story by the user. 2) The author role and time of the post, the role could be the patient, patients relative, carer or doctor. 3) The more about section is from website moderator, it inserts relevant tags to the story. 4) & 5) are the most important fields to our project, these field are inserted by the user, the fields indicate what user thinks the story is about, and we use these fields to simulate user judgement in topic modelling.

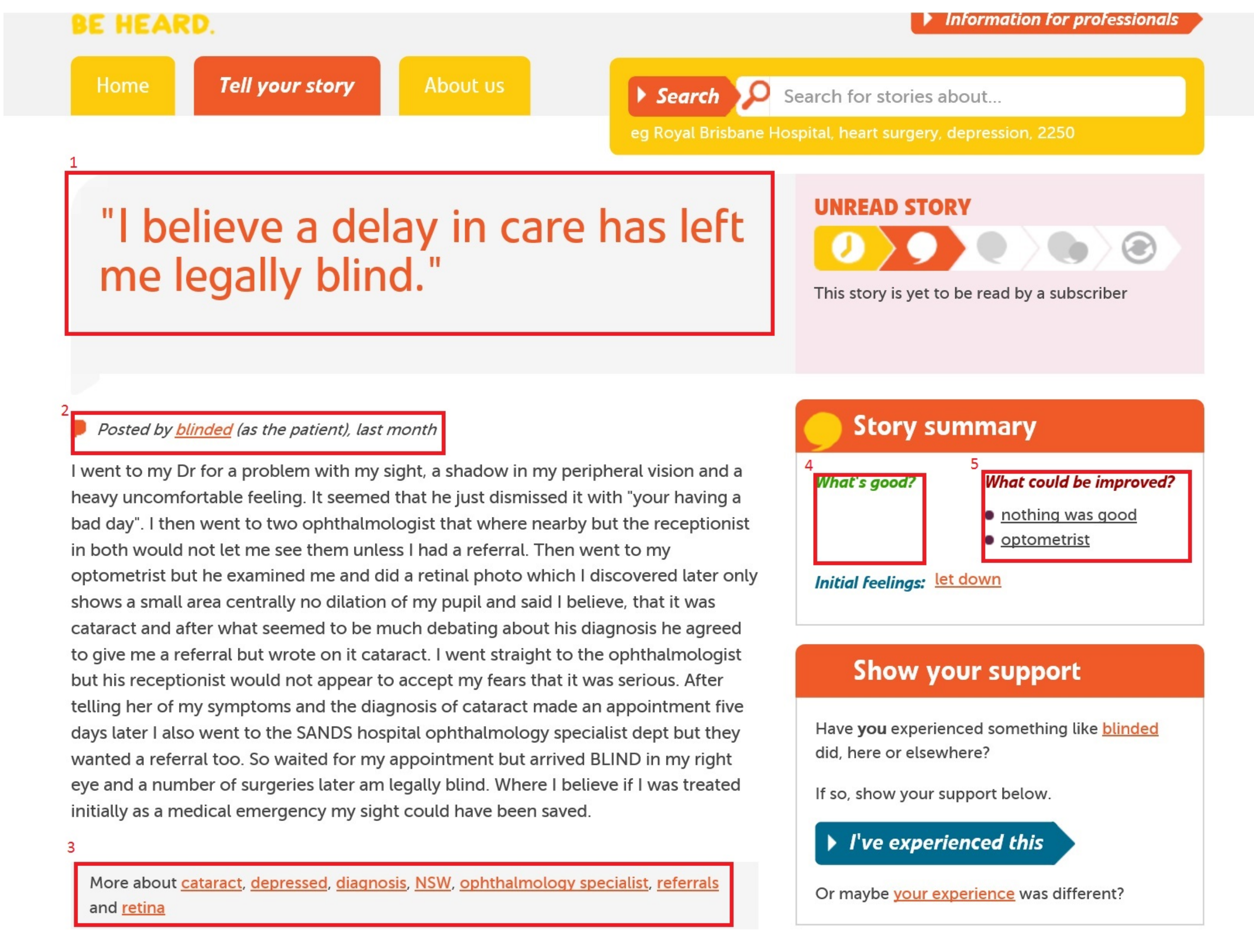


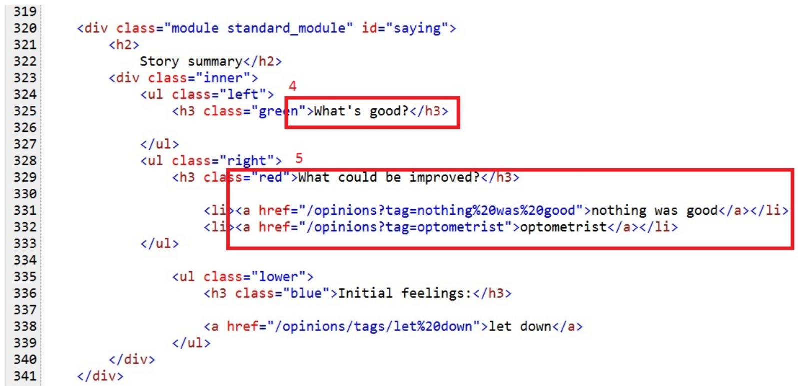
Figure 3.1: Patient Opinion Story Sample

```html
    <article id="story" data-po-opinionid="59518" itemscope itemtype="http://data-vocabulary.org/Review">

<h1>
    <span class="top_dec"></span>
    <blockquote>
        &quot;<span id="opinion_title" itemprop="summary" class="">I believe a delay in care has left me legally blind.</span>&quot;
    </blockquote>
    <span class="btm_dec"></span>
</h1>

<p class="info">

        Posted by
            <span itemprop="reviewer"><a href="/opinions?author=blinded" title="Other opinions from blinded">blinded</a></span>

        (as <span id="opinion_author_role" class="">the patient</span>),
    <time itemprop="dtreviewed" datetime="2014-07-22T04:35:56Z" title="Submitted on 22/07/2014 at 04:35 and published by Patient Opinion on 04/08/2014 at 05:04">last month</time>
</p>

    <div class="story_copy">
        <blockquote id="opinion_body" itemprop="description" class="text ">
            <p>I went to my Dr for a problem with my sight, a shadow in my peripheral vision and a heavy uncomfortable feeling.  It seemed that he just dismissed it with "your having a bad day".  I then went to two ophthalmologist that where nearby but the receptionist in both would not let me see them unless I had a referral.  Then went to my optometrist but he examined me and did a retinal photo which I discovered later only shows a small area centrally no dilation of my pupil and said I believe, that it was cataract and after what seemed to be much debating about his diagnosis he agreed to give me a referral but wrote on it cataract.  I went straight to the ophthalmologist but his receptionist would not appear to accept my fears that it was serious.  After telling her of my symptoms and the diagnosis of cataract made an appointment five days later I also went to the SANDS hospital ophthalmology specialist dept but they wanted a referral too. So waited for my appointment but arrived BLIND in my right eye and a number of surgeries later am legally blind.  Where I believe if I was treated initially as a medical emergency my sight could have been saved.</p>
        </blockquote>

    </div>

    <div class="related clearfix">
        <p>
            More about <a href="/opinions/tags/cataract">cataract</a>, <a href="/opinions/tags/depressed">depressed</a>, <a href="/opinions/tags/diagnosis">diagnosis</a>, <a href="/opinions/tags/nsw">NSW</a>, <a href="/opinions/tags/ophthalmology%20specialist">ophthalmology specialist</a>, <a href="/opinions/tags/referrals">referrals</a> and <a href="/opinions/tags/retina">retina</a>
        </p>
    </div>
```

Figure 3.2: Patient Opinion Story Sample Source 1

3.2 Preprocessing of Data

Everything been converted to lower-case, collect all unique words in user specified field. This collection is used to filter out the words in each topic that generated by LDA. A list of related document ID to each word also collected, (see Figure3.1) and indexed for fast look up for document frequency.



```
<div class="module standard_module" id="saying">
    <h2>
        Story summary</h2>
    <div class="inner">
        <ul class="left">
            <h3 class="green">What's good?</h3>

        </ul>
        <ul class="right">
            <h3 class="red">What could be improved?</h3>

                <li><a href="/opinions?tag=nothing%20was%20good">nothing was good</a></li>
                <li><a href="/opinions?tag=optometrist">optometrist</a></li>
        </ul>

            <ul class="lower">
                <h3 class="blue">Initial feelings:</h3>

                <a href="/opinions/tags/let%20down">let down</a>
            </ul>
    </div>
</div>
```

Figure 3.3: Patient Opinion Story Sample Source 2

3.3 Using user input to improve topic modelling result

Topics learned from LDA sometimes dont convey much useful information, sometime it is caused by overfeeding the result set, for example it will include top 20 words for each topic (based on the settings, the total number in each topic can be configured), some words may not make any sense in current topic but statistically significant to the topic. Our goal is try to use user input to reduce the noise while retaining as much information as possible to describe or label the topic. The generative process is given as follows:

1. Collect unique words from user specified field as S set.
2. Generate a set of topics T with LDA model.
3. Calculate result set R as: for each topic $t \in \{1, ..., T\}, r_n = t_n \cap S$
4. Treat the result T as collection of document, calculate idf for each term with $idf(t, D) = log\frac{N}{|\{d \in D : t \in D\}|}$, where N = 100 represent the 100 topics generated with LDA. The original tf-idf method is not used, for each term the tf for current document always equals to 1.

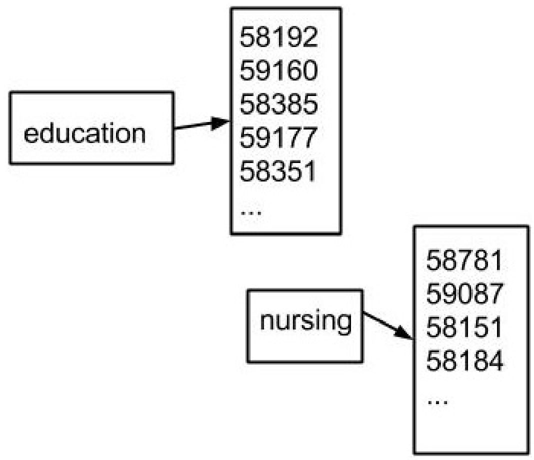


Figure 3.4: Patient Opinion Story Sample

5. Re-arrange terms in set R with idf score, so the more significant word appear in front of each topic.

Chapter 4

Experiments and Result

We collected all 624 stories from Patient Opinion by August 2014. The count of unique user specified term is 659. 100 topics are generated using Mallet [1] with setting of optimize interval equals to 20. The number of unique terms in the result set R is 527, compare to 1440 in the original T set. Table 4.1 Shows the topic composition, which is the per topic probability distribution over documents. Rows represent documents with document ID in first column and the remaining columns represent topic probability distributions over current document.

The total composition score for each topic is calculate with the help from term to document ID index we build before. Clearly, the total composition of T is expected to greater than it's subset R. The average difference of composition between T and R over 100 topics is 1.0829, there are 58 topics has the difference below the average, we consider the probability of distribution of the topic over the documents still significant in that 58 topics. A sample of the topics are selected from remaining topics, see Table4.2.

[1] http://mallet.cs.umass.edu/

Table 4.1: Example of Composition

Doc ID	Topic Index	Composition	Topic Index	Composition	...	Topic Index	Composition	Topic Index	Composition
58954	61	0.1171875	91	0.0390625	...	78	0.0234375	72	0.0234375
58832	83	0.076388889	78	0.048611111	...	10	0.048611111	71	0.034722222
58953	83	0.077380952	65	0.053571429	...	29	0.041666667	60	0.029761905
58956	78	0.025	76	0.025	...	65	0.025	62	0.025
58834	42	0.065517241	11	0.065517241	...	58	0.037931034	44	0.037931034
58710	12	0.108208955	18	0.063432836	...	90	0.041044776	71	0.041044776
58952	91	0.044642857	73	0.026785714	...	59	0.026785714	36	0.026785714
58830	94	0.108490566	80	0.051886792	...	99	0.04245283	71	0.04245283
58951	93	0.0625	81	0.052884615	...	0	0.052884615	79	0.043269231
58719	51	0.27238806	27	0.063432836	...	42	0.026119403	22	0.026119403
58716	77	0.041666667	90	0.025	...	73	0.025	62	0.025
58718	83	0.242307692	47	0.05	...	71	0.042307692	11	0.034615385
58839	25	0.070512821	63	0.032051282	...	59	0.032051282	58	0.032051282
58717	43	0.050660793	57	0.046255507	...	92	0.04185022	72	0.04185022
58723	91	0.028301887	68	0.028301887	...	35	0.028301887	99	0.009433962
58965	83	0.097014925	1	0.037313433	...	77	0.02238806	72	0.02238806

1116 out of original 1440 terms appears once in T set. So more than half of the term has the highest idf value 2 in the data set. 236 terms appear twice with idf=1.7 and 55 terms appear 3 times with idf = 1.5. There are significant variance in the result topics. Some topics may remain the existing order or only have one word shifted, some may look very differently. For example, {program, kate, lifestle, meet, included, learnt, learning, relationship}, only $idf_{kate} = 1.7$, the idf for the rest equals to 2. So the new topic becomes {program, kate, meet, included, learnt, learning, relationship}. Another example {bed, family, care, staff, time, attending, unit, palliative}, $idf_{care} = 1.3, idf_{staff} = 1.2 and idf_{palliative} = 2$, the rest has idf equals to 1.7, then the new topic should look like {palliative, bed, family, time, attending, unit, care, staff}.

Table 4.2: Sum of Composition

Topic Index	Original Composition	Filtered Composition
1	3.920362	3.382284
2	3.742091	3.275150
3	4.396231	3.930039
4	4.275353	3.603821
5	4.569670	4.444751
6	3.153133	2.775077
7	3.368214	2.622281
8	4.541646	4.018145
9	5.399709	5.194911
10	4.498450	3.711763
11	4.255079	4.101976
12	6.755018	6.489172

4.1 Topic Coherence Evaluation

Apart from quantitative and qualitative evaluation as above, evaluating topic coherence is a component of the larger question of what are good topics, what characteristics of a document collection make it more amenable to topic modelling, and how can the potential of topic modelling be harnessed for human consumption (Newman et al. [2010]). The topic coherence is measured with Pointwise Mutual Information (Newman et al. [2011]) (PMI) score:

$$PMI - Score(w) = (\frac{N^2 - N}{2})^{-1} \sum PMI(w_i, w_j), ij \in \{1...N\} \tag{4.1}$$

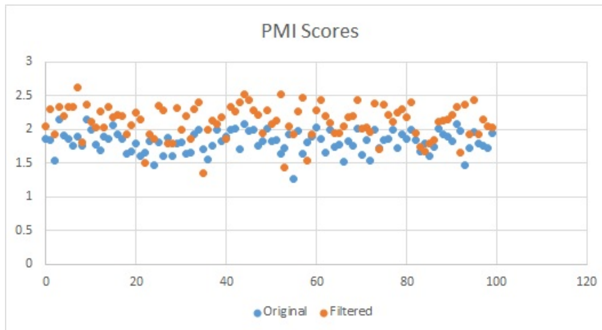


Figure 4.1: PMI-Score

$$wherePMI(w_i, w_j) = log\frac{P(w_i, wj)}{P(w_i)P(w_j)}, \tag{4.2}$$

Since the number of term that form each topic isn't normalized, we calculate the average of the topic, where N is the number of terms in that topic. $(\frac{N^2-N}{2})^{-1}$ gives the number of distinct pairs in N. The measure is symmetric $P(w_i, wj) = P(w_j, wi)$, which means we only measure the difference of topic coherence between original topic and filtered topic.

Table4.? Shows some samples of the PMI scores, plus the difference between two PMI in last column. As we can see from the result, the majority of the topics result in improvement of coherence, and the overall average is 0.297. A paired-t test shows the $P = 3.76e^{-22}$, which means the filtered topic is significant from the original one.

Table 4.3: Sample topics

Original	Filtered
program healthy kate lifestyle sessions eat meet programme included organised held healthier encouraging learnt foods beneficial learning relationship handle	program kate lifestyle meet included learnt learning relationship
gp local recently records government copy prescription paper multiple tasmania gps referring avail beginning calls surprised cairns super shared	gp local records government prescription paper gps cairns
physio gp mri injury follow shoulder xray week asked hospital discussed full physiotherapist neck stand complaining neurologist princess forte	physio gp mri xray hospital full physiotherapist neck
call waiting phone told back called list unit rang ring explain apparently assumed clerk calling noticed mcewin lyell requested	call waiting phone back list unit clerk calling
father bed family care appears staff time attending difficult dad speak comfort unit incident law visitor awake gosford palliative	bed family care staff time attending unit palliative
time advised team contact causing consultant tumour professional manner independent safe stressful arrival note closed usual considerate empathetic seizures	time team contact consultant professional manner independent empathetic
looked experience er bad partner full approach worry give free skills male chronic terrible running provider building drive welcomed	looked experience er partner full approach skills male building
night stay thing major hospital admission support suggested accommodation fully sydney unable relatives sleep developed tuesday staff added environment	night stay hospital admission accommodation relatives sleep staff environment
waiting wait hours room hour area waited reception number temperature hurt remember panadol minutes geelong sunday impressed time er	waiting wait hours room area reception number time er
child issues jean aboriginal helped understanding knowledge school minds behaviour hay mighty woods anger louise clinician strategies interaction love	child jean aboriginal understanding knowledge school minds behaviour woods strategies
public brisbane system live mater hospital pa run booking advise toowoomba meds qld health expect west lift weekend weak	public brisbane system hospital run booking meds health west lift
health community sarah local support primary medicare rural group libby people murrumbidgee provide part clients topics art groups guest	health community sarah local primary medicare rural group people murrumbidgee clients guest

Table 4.4: PMI Scores

Topic Index	Original Topic PMI Scores	Filtered Topic PMI Scores	Difference
1	1.864771	2.040466	0.175695
2	1.844573	2.305665	0.461092
3	1.543335	1.928804	0.385469
4	2.145729	2.333656	0.187927
5	1.912059	2.195632	0.283573
6	1.854318	2.341428	0.48711
7	1.748795	2.33346	0.584665
8	1.899982	2.624272	0.72429
9	1.752184	1.80262	0.050436
10	2.152952	2.371954	0.219002
11	1.986177	2.120463	0.134286
12	1.776924	2.030256	0.253332

Chapter 5

Conclusion and Future Work

This study shows the possibility of using user input to improve topic coherence in topic modelling of healthcare related blogs. To the best of our knowledge, this has not been done before. We proposed a method that use user input words as a filter for the result from baseline LDA model. We successfully reduced the number terms in each topic while still keep the topic meaningful, the terms that been omitted can be considered as noise, which means the documents they associate to do not significantly contribute to the possibility distribution over the topic, this is evaluated by the total composition score. Hence the overall topic coherence is improved as less noise in the topic. Furthermore, we experiment using tf-idf to re-rank the terms in each topic. Unfortunately, PMI-score evaluation is symmetric, which means the order of each term in topic isn't taken in count. We couldn't find an existing statistical model to fit in the evaluation. Our method of ranking the terms by idf isn't ideal, since the term frequency for each term for in a topic always equals to 1, this approach is reasonable for step 1. With more resources in the future work, this approach could be expended to count term frequency in the scope of whole collection of documents, and idf still from the topic list. If the scope of the factor expended to the whole collection, it is also reasonable to rank the topics not only the terms in each topic, hence the model could make suggestions of which topic is more likely an important one.

Bibliography

M. M. Bismark and D. M. Studdert. Governance of quality of care: a qualitative study of health service boards in victoria, australia. *BMJ quality & safety*, pages bmjqs–2013, 2013.

M. M. Bismark, M. J. Spittal, L. C. Gurrin, M. Ward, and D. M. Studdert. Identification of doctors at risk of recurrent complaints: a national study of healthcare complaints in australia. *BMJ quality & safety*, 22(7):532–540, 2013.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

W. Buntine and A. Jakulin. Applying discrete pca in data analysis. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 59–66. AUAI Press, 2004.

J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Leveraging multi-domain prior knowledge in topic models. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2071–2077. AAAI Press, 2013.

A.-S. Garcia-Moya.L and Berlanga-Llavori.R. Retrieving product features and opinions from customer reviews. *Intelligent Systems*, 28(3):19–27, 2013.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

D. Newman, S. Karimi, L. Cavedon, J. Kay, P. Thomas, and A. Trotman. External evaluation of topic models. In *Australasian Document Computing Symposium (ADCS)*, pages 1–8. School of Information Technologies, University of Sydney, 2009.

D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

D. Newman, E. V. Bonilla, and W. Buntine. Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems*, pages 496–504, 2011.