TOWARDS PERSONALIZED HASHTAG RECOMMENDATION SYSTEM: A STUDY ON TWITTER

Project Report

Submitted by

M V SAI TEJA
COE13B016

in partial fulfillment for the award of the degree of

Bachelor of Technology

in

COMPUTER ENGINEERING

Indian Institute of Information Technology
Design and Manufacturing, Kancheepuram, India

November 2016

BONAFIDE CERTIFICATE

This is to certify that the thesis titled "**TOWARDS PERSONALIZED HASHTAG RECOMMENDATION SYSTEM: A STUDY ON TWITTER DATA**" submitted by **Mr. M V SAI TEJA (COE13B016)** to the Indian Institute of Information Technology Design and Manufacturing, Kancheepuram, for the award of **Bachelor of Technology in COMPUTER ENGINEERING,** is a *bona fide* record of the project work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Nargis Pervin
Project Guide

Assistant Professor
Indian Institute of Information Technology
Design and Manufacturing, Kancheepuram
Chennai 600 127
India

Place: Chennai
Date: 01/12/16

ACKNOWLEDGEMENTS

I would like to express my profound sense of gratitude to my guide, Dr. Nargis Pervin, Assistant Professor, Department of Computer Engineering, for her valuable advice and motivation throughout the completeness of this project. Her technical knowledge has been an invaluable asset, but may be more importantly, she has helped to keep me focused and pointed in the right direction.

I am also thankful to all my friends for helping me whenever I encountered hiccups during the work.

ABSTRACT

The thesis is about personalized hashtag recommendations in Twitter. Twitter network is currently overwhelmed by massive number of tweets generated by its users. To effectively organize and search tweets, users have to depend on appropriate hashtags inserted into tweets. Hashtag is a word or phrase preceded by hash sign (#) which is commonly used in social media, e.g., #KABALI, #PVSINDHU etc., The intention behind hashtag was to make it easy for the users (like in Twitter) to search for a content and also for categorizing their posts/tweets. Hashtags also represents the mood of the user like #happy. But, only 8% of the tweets in Twitter contain hashtags.

The aim of this project is to recommend personalized hashtags for the Twitter users. The hashtag recommendations help in effective categorizing of tweets and also helps in fetching better search results.

Contents

List of Figures

List of Tables

Nomenclature

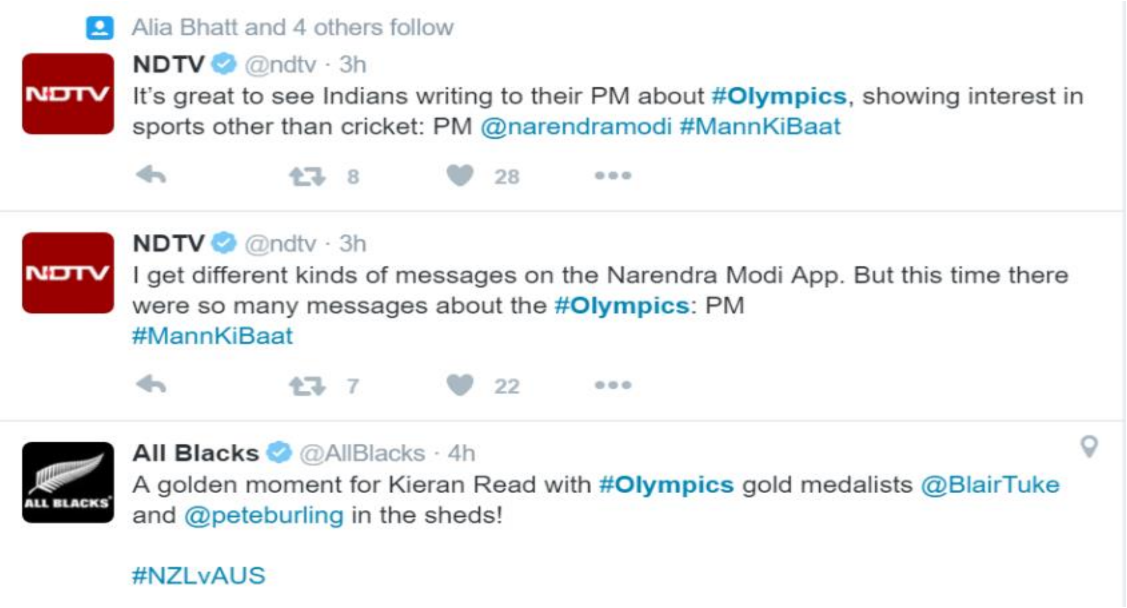
LDA – Latent Dirichlet Allocation

Chapter 1 Introduction

1.1 Motivation

Recommendation systems have been widely used at e-commerce websites to identify from a huge range of products the most appropriate ones for each user. The recommendation systems can be used in social networking sites like Twitter for hashtag recommendations which helps in effective categorizing of tweet and also helps in fetching better search results.

Twitter is one of the largest social networking and microblogging site that enables users to write and read messages called "tweets". In Twitter following someone means the user who has followed a person can view his/her tweets in his/her timeline. A tweet that a user shares publicly with his/her followers is known as retweet. Hashtag is mainly used as a means to share the mood of the user or to emphasize on a specific topic. Chris Messina, a social technology expert was the first person to use hashtag on Twitter. The popularity of Twitter has gained popularity in other social networks as well, e.g., Facebook and Instagram.



Alia Bhatt and 4 others follow

NDTV @ndtv · 3h
It's great to see Indians writing to their PM about #Olympics, showing interest in sports other than cricket: PM @narendramodi #MannKiBaat

8 28

NDTV @ndtv · 3h
I get different kinds of messages on the Narendra Modi App. But this time there were so many messages about the #Olympics: PM #MannKiBaat

7 22

All Blacks @AllBlacks · 4h
A golden moment for Kieran Read with #Olympics gold medalists @BlairTuke and @peteburling in the sheds!

#NZLvAUS

Figure 1.1: Usage of hashtags in tweets.

In this project my aim is to develop a personalized hashtag recommendation system for Twitter users which motivates the people to use hashtags in their tweets. This recommendation also helps people to maintain a common hashtag for a particular event like #RioOlympics2016 instead of different hashtags like #Rio, #Olympics2k16 or #Rio2016 which represents same event of Rio Olympics 2016. The usage of different hashtags for the same event hinders the search results of a particular event.

1.2 Problem Statement

The objective of the project is to recommend personalized hashtags for Twitter users based on the context in the tweets and also based on the user's previous tweets.

1.3 Overview of the Project

In the project, for hashtag recommendations, topic modelling of tweets is applied. Firstly, Twitter data was collected and topic modelling using LDA was applied on the collected Twitter data and the results were observed.

Chapter 2 Literature Review

Lei Yang et al. [1] states the role of hashtag in Twitter. In [1] authors stated that the hashtag has dual role. On one hand, hashtag serves as bookmark of the content which links tweets with similar topics, on the other hand it serves as symbol of community membership. Content related hashtag recommendations depend on relevance and preference. In [1] authors compared the recommendations of hashtags with the adaption of bookmark in Del.icio.us (an e-commerce website). The relevance between hashtags and content is measured by using relevance function. Preference relates to how closely hashtag is related to a person which is also known as personalized preference. It was measured with aggregate function of similarity between hashtag and all other hashtags used by the user. Measure related to joining community depends on prestige and influence. Prestige of a hashtag community is the aggregate measure of prestige of individual users of the community. Influence is like if my friend uses a hashtag or joins a community, so do I. In [1] authors have used regression analysis to check whether the above stated factors have the power of hashtag recommendations with hashtag adaption behavior as dependent variable. They tested on various data sets and the results suggested the dual role of hashtags.

Frederic Godin et al. [2] used unsupervised topic models for hashtag recommendations. Topic modelling is a type of statistical modelling which is used for discovering abstract topics that occur in collection of data. In[2] author used Latent Dirichlet Allocation(LDA) for hashtag recommendation. LDA is a way of automatically discovering topics that a document contains. They have collected Twitter data using Twitter Streaming API. They filtered data by removing URLs, HTML entities, digits, punctuations and hash characters and by taking the tweets only from English language. This method of approach helps to recommend new recommendations but it is not personalized.

Jieying She et al. [3] proposed a supervised model called TOMOHA (TOpic MOdel based HAshtag recommendation). In [3] author stated that using LDA, an unsupervised approach as in [2] failed for short texts and only works for long texts. They have collected Twitter data using Twitter REST API. They also proposed TOHOMA-follow model where users may follow the topics of their followees. As they used supervised model, they have used parallel training for speeding up the training of the system. After their experiment, their solution of TOHOMA and TOHOMA-follow model [3] showed more advantages in efficiency than LDA model [2].

David M. Blei et al. [4] described latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora.

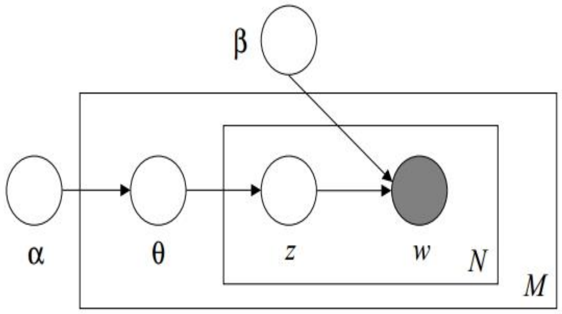


Figure 2.1: Graphical model representation of LDA

The Fig 2.1 represents the graphical model representation of LDA. The terms represent the following

M – Number of documents

N – Number of words in a document

Z – Topic assigned for each word

α - Per document topic distribution

θ - Topic distribution

β -Per topic word distribution

LDA is used for finding the topics for given documents automatically. It assigns topics for each document with probability that the given document belongs to that particular topic.

Chapter 3 Implementation of topic modelling

The work was done by implementing LDA topic modelling for the Twitterdata. The following steps were followed for implementation

3.1 Data Collection

The first objective was to collect data from Twitter. Twitter's REST API was used for collecting Twitter data. Next step is authorizing with the Twitter API. We have to create a Twitter

application and it will generate API keys and access tokens which are required for authentication and collection of Twitter data.

The next step was to collect Twitter data from the topics which were trending in Twitter. Some of the tweets include hashtags-

#INDIANARMY

#LINKINPARK

#RAJINIKANTH

#ISIS

#PRAYFORPARIS

A total of 10,000 tweets were collected with 1000 tweets from each topic.

3.2 Data pre-processing

The pre-processing of the data was done like-

1.Removing all the non-ASCII characters

2.Converting all the text to lower case.

3.Removing numbers, punctuations

4.Stemming words.

5.Removing words whose length is not in bertween 3 and 30.

6.Parts of speech tagging

Stemming is a process of reducing the inflected (or sometimes derived) words to their base or root form. For example, walking and walked are stemmed to their root word walk. The text pre-processing was done through "tm" package in Twitter. A Corpus was created by combining all the tweets with each tweet as a document. Next, unimportant words from the Corpus were removed by using Tf-idf (term frequency inverse document frequency). It is a principle approach to filter out unimportant words from the text. The Tf-idf values for each word are computed and the words whose Tf-idf values are less than median of Tf-idf are removed. After that all the tweets that do not contain any words were removed from the Corpus.

Next step is parts of speech tagging where openNLP package was used in Rstudio. The result will be the tags for each word in the string.

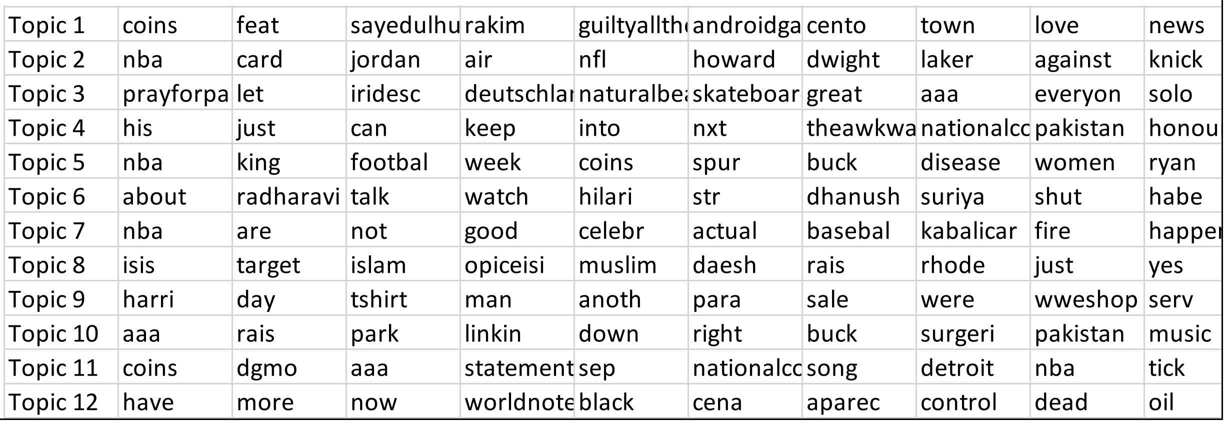
3.3 Performing LDA topics modelling

Next, topic modelling was applied on the processed Twitter data by using LDA. The number of topics required for the topic modelling were given randomly and the output of the topic modelling was observed for each value of k (number of topics). The output of LDA topic modelling is topics and words corresponding to the particular topic arranged in decreasing order of probability of the word belonging to that particular topic.

Chapter 4 Results

After applying the LDA topic modelling, the results are shown by the following table

Table 1: Table for topics and corresponding words.



Topic	Terms									
Topic 1	coins	feat	sayedulhu	rakim	guiltyallth	androidga	cento	town	love	news
Topic 2	nba	card	jordan	air	nfl	howard	dwight	laker	against	knick
Topic 3	prayforpa	let	iridesc	deutschla	naturalbe	skateboar	great	aaa	everyon	solo
Topic 4	his	just	can	keep	into	nxt	theawkwa	nationalc	pakistan	honou
Topic 5	nba	king	footbal	week	coins	spur	buck	disease	women	ryan
Topic 6	about	radharavi	talk	watch	hilari	str	dhanush	suriya	shut	habe
Topic 7	nba	are	not	good	celebr	actual	basebal	kabalicar	fire	happe
Topic 8	isis	target	islam	opiceisi	muslim	daesh	rais	rhode	just	yes
Topic 9	harri	day	tshirt	man	anoth	para	sale	were	wweshop	serv
Topic 10	aaa	rais	park	linkin	down	right	buck	surgeri	pakistan	music
Topic 11	coins	dgmo	aaa	statement	sep	nationalc	song	detroit	nba	tick
Topic 12	have	more	now	worldnote	black	cena	aparec	control	dead	oil

The output of the topic modelling was stored in a csv file in local disk which contains top 100 words for each topic. Now, we look at words of one topic in detail. Let us take Topic 8, the top 10 terms in Topic 8 are

Isis, Target, islam, opiceisi, muslim, daesh, rais, rhode, just, yes “isis” was one of the topic which was collected. So, in Topic 8, words like “islam”, “muslim”, “target” were related to isis in different contexts. Isis was the top word in Topic 8. Of the tweets collected, we can also find the topic to which each tweet was assigned.

Chapter 5 Conclusion and Future Work

In this Project, the hashtag recommendations using LDA topic modelling is performed and the results were observed. Though many topics include words of a similar domain, some of the topics even include outliers. The number of topics were experimented manually and results were observed. Though this experimental analysis can be useful for small datasets, when dealing with large datasets, the experimental approach may not be recommended. Methods to automatically find number of topics for a given document should be implemented. Data pre-processing should still be improved to have better results because the output will be determined based on input words.

We can perform sentiment analysis to the Twitter data and we can recommend hashtags based on the score we obtain from sentiment analysis. (For example, a score of +7 indicate a very positive tweet and we can recommend words like happy, delighted).

Next, we need to come up with a personalized recommended system where the recommendations not only rely on the context of the text in the tweet but also be based on the previous activities of the user in Twitter. After that we need to test our algorithm with existing baseline algorithm to measure the accuracy.

References

[1]. Lei Yang1,Tao Sun2 , Ming Zhang, Qiaozhu Mei. We Know What @You #Tag:Does the Dual Role Affect Hashtag Adoption?. Proceedings of the 21st international conference on World Wide Web (2012).

[2]. Frederic Godin, Viktor Slavkovikj, Wesley De Neve. Using Topic Models for Twitter Hashtag Recommendation. Proceedings of the 22nd International Conference on World Wide Web (2013).

[3]. Jieying She, Lie Chan. TOMOHA: TOpic MOdel-based HAshtag Recommendation on Twitter. Proceedings of the 23rd International Conference on World Wide Web (2014).

[4]. David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation
. Journal of Machine Learning Research 3 (2003)

[5]. Timothy Graham, Robert Ackland. Topic Modeling of Tweets in R: A Tutorial and Methodology.http://www.academia.edu/19255535/Topic_Modeling_of_Tweets_in_R_A_Tutorial_and_ Methodology

[5]. David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation
. Journal of Machine Learning Research 3 (2003)

[6]. David Alfred Ostrowski. Using Latent Dirichlet Allocation for Topic Modelling in Twitter. Proceedings of 9th International Conference on Semantic Computing(2015).