

Electronics and Computer Science
Faculty of Physical Sciences and Engineering
University of Southampton

Jamie Davies
16th April 2014

Enhanced Content Analysis and Information
Retrieval Using Twitter Hashtags

A project report submitted for the award of
MEng Computer Science

Supervisor:
Dr. Nick Gibbins

Examiner:
Dr. Klaus-Peter Zauner

Abstract

One of the key characteristics of Twitter and other microblogging platforms is the use of ‘hashtags’ — topical/categorical annotations provided by the authors of the posts (tweets) themselves. This flexible system was designed for the effective organisation and searching of tweets, but with Twitter facing an ever-increasing number of users and tweets it is hard for users to keep track of the vast number of hashtags in popular use. This results in data from the hashtags being fragmented and inaccurate due to the users making poor or uninformed hashtag choices.

If users are presented with a choice of relevant hashtags when writing a tweet, they are more likely to publish tweets with accurate tag data. This project aims to create an intelligent hashtag recommendation tool to improve the quality of the information we can gain from hashtags. However, whilst such a system could improve future tweets, tweets that have already been published will remain untouched by the system. Thus, the system will be extended to also retrofit hashtags to published tweets — allowing for tweets to appear in search results for a particular hashtag even if they don’t actually contain the hashtag in question.

Contents

1	Introduction	5
2	Preliminary Data Analysis	7
3	Background and Literature Review	9
3.1	Recommendation Systems	9
3.1.1	Collaborative Filtering	9
3.1.2	Content-Based Recommendation	9
3.1.3	Relevance Feedback	10
3.2	Hashtag Recommendation Research	10
3.2.1	Current Twitter Hashtag Implementation	10
3.2.2	Comparing Tweets To Other Tweets	11
3.2.3	Creating Personalised Recommendations	11
3.2.4	Overcoming Hashtag Duality	12
3.3	Broader Classification in Twitter	12
3.3.1	Categorising Tweets	12
3.3.2	Categorising Users	13
4	Requirements	14
4.1	Functional Requirements	14
4.2	Non-Functional Requirements	14
5	Design	15
5.1	Architecture	15
5.1.1	Processing Framework	15
5.1.2	Classifier	15
5.2	Interface	15
5.2.1	API Endpoints	15
5.2.2	Web Interface	15
6	Implementation	16
6.1	Stream Processing	16
6.2	Classification System	16
6.2.1	Training	16
6.2.2	Classifying	16
6.2.3	Optimisation	16
6.2.4	RESTful Interface	16
6.3	Web Application	16
6.3.1	Framework	16
6.3.2	Authentication	16
6.3.3	Connecting to Classifier	16
7	Testing	17
7.1	Strategy	17

7.2	Unit Testing	17
7.3	Classification Testing	17
7.4	User Interface Testing	17
7.5	Requirements Validation	17
8	Conclusion	17
8.1	System Evaluation	17
8.2	Project Management	17
8.3	Future Work	17
A	Time Management	18
B	Testing Plans & Results	19
C	REST API Interface	20
D	Original Project Brief	21
	References	22

Acknowledgements

I would like to thank my supervisor, Dr. Nick Gibbins, for providing inspiration and guidance throughout the course of the project.

I would also like to thank Dr. Jonathon Hare and Dr. Sina Samangooei for their continued support and advice.

Statement of Originality

The work contained within this report has not been previously submitted for a degree or diploma at any other higher education institution. To the best of my knowledge and belief, this project contains no material previously published or written by another person except where due references are made.

1 Introduction

Collaborative tagging (sometimes referred to as social tagging) systems are designed to allow users to navigate, organise and manage online resources. The central idea behind these systems is the concept of an annotation: the action of a user applying a personalised tag to a resource. On a large scale, many of these annotations can be taken together to form a complex network of users, resources and tags — commonly referred to as a *folksonomy* (Xu et al., 2008).

One of the most popular applications of a folksonomy is Twitter¹. ‘Hashtags,’ which are simply any combination of letters or digits preceded by a hash character (#), afford users the ability to create free-form tags to describe the content of their posts (tweets). Tweets can be categorized with several hashtags, thereby creating networks of tweets and users, and making it easy to find other related tweets, users and hashtags. This renders hashtags as a powerful tool in aiding the search and navigation of the billions of tweets contained within Twitter.

Despite its numerous benefits, the hashtag system presents new challenges to overcome before it can become truly useful. Due to the open nature of folksonomies, it is important that users have the freedom to create and use exactly the tags they wish to use. However, this unsupervised tagging can result in vast numbers of hashtags for users to choose from — often including redundant or ambiguous hashtags. When posting a tweet, there is nothing stopping a user from creating an entirely new hashtag to describe something with exactly the same meaning as a collection of other hashtags. This tag redundancy can confuse users and fragment the true meaning behind the synonymous hashtags.

There is also the converse issue of tag ambiguity to consider, in which a single hashtag could be associated with many meanings and connotations. The use of these hashtags can incorrectly imply that tweets are similar and related to one another, when in fact they are from distinct unrelated categories.

Because these problems are complicating the already vast hashtag information space, an intelligent recommendation system is needed to aid users in their choice of hashtags. For example, when a user writes a tweet such as: “*The golden gate bridge looks so nice today! I LOVE SF in this weather!*”, the system should recommend hashtags such as #SanFrancisco, #California or even #sohappy, but should not recommend hashtags such as #ScienceFiction, #amreading or #fantasy.

Such a recommendation system would also need to aid users in the navigational and exploratory aspects of folksonomies. For example, when a user performs a search on Twitter for the hashtag #BGT, the system should dis-

¹www.twitter.com

play not only those tweets that contain #BGT, but also tweets that contain other relevant hashtags such as #BritainsGotTalent, #VoteAttraction or #VoteJackCarroll.

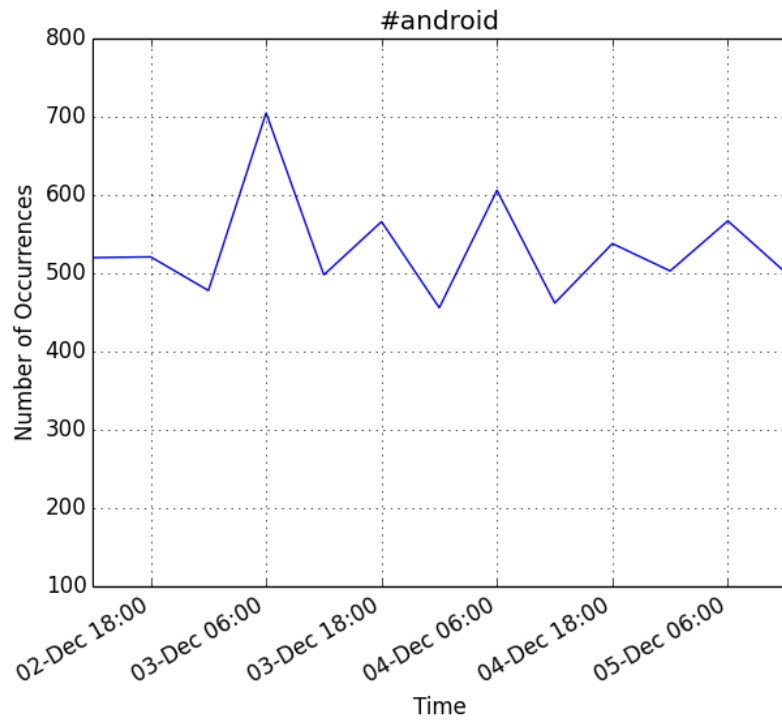
2 Preliminary Data Analysis

To investigate the usage of hashtags and the problems associated with them further, a dataset of 500,000 tweets was collected from the sample Twitter stream over a time period of approximately 4.5 days. The tweets collected were filtered to ensure that they were in English, and contained at least one hashtag. Through the use of some trivial Python scripts, it is easy to find evidence of poor hashtag choices. In particular, the most overwhelming observation is the quantity of redundant hashtags in use throughout Twitter. Table 1 shows several hashtags taken from the dataset and their number of occurrences. Despite the tags all having exactly the same semantic meaning, the use of the tags is fragmented and spread across several tags.

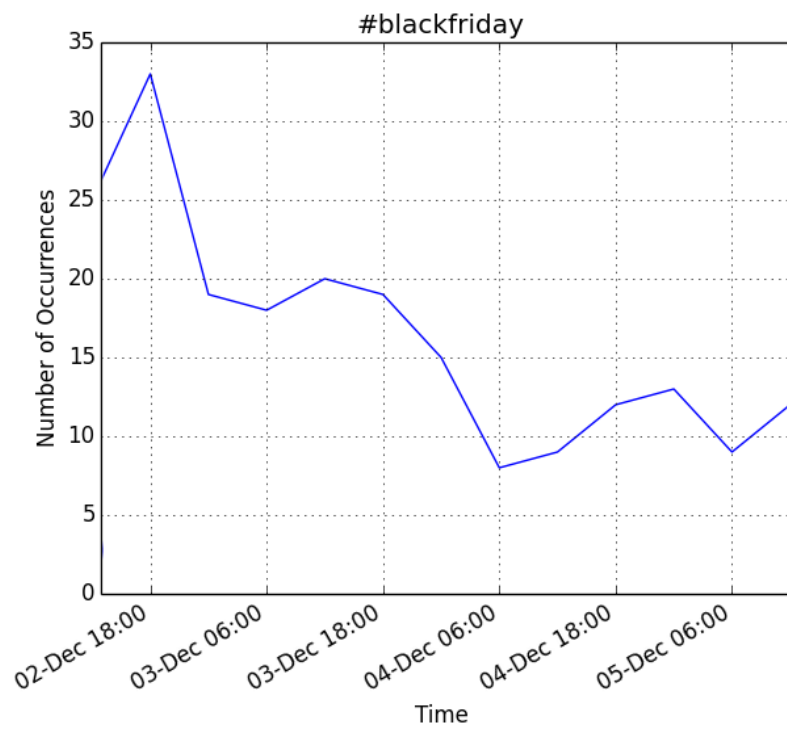
Hashtag	Number of Occurrences
#peopleschoice	94849
#peopieschoice	2043
#peoplechoice	439
#peoplesch	287
#peopleschoi	269
#peoplesc	230
#peoplescho	219
#peolpeschoice	164
#peopleschioce	137
#pca	94

Table 1: Occurrences of hashtags referring to the People’s Choice Awards

Another interesting observation is that of the times that different hashtags are used, which varies greatly from hashtag to hashtag. Some tags are used in tweets at a fairly uniform rate (1(a)), whilst others feature large spikes of usage over a short period of time (1(b)). This demonstrates the communities behind the hashtags — rather than a hashtag being used at a particular time for a particular event, some hashtags are used consistently by users to join in on a large-scale conversation.



(a) #android



(b) #blackfriday

Figure 1: Usage of hashtags over the period that the dataset was collected.

3 Background and Literature Review

The main design goals behind hashtags are to categorise tweets and allow them to show up more easily in searches². Whilst the task that this project is aiming to complete is novel and fairly unexplored, it is well connected with other experiments, systems and projects within the research community.

3.1 Recommendation Systems

Traditional recommendation systems are in place all over the web today. From music discovery services (such as Last.fm³) to suggested purchases on retail sites (like that in place at Amazon⁴, these systems are all personalised recommendation engines that take an individual user’s preferences and use them to provide suggestions tailored to that user.

3.1.1 Collaborative Filtering

Most personalised recommendation systems employ a set of techniques known as collaborative filtering. These techniques were first coined by Goldberg et al. (1992), where a system named *Tapestry* was created that allowed people to attach annotations to documents, and then use that information to filter the documents for other users.

One common implementation of collaborative filtering is the so-called “user-to-user” approach. “User-to-user” collaborative filtering works by taking the preferences of a user A , and finding a small subset of other users in the system that have similar preferences. For each user B in the subset any items that B has adopted that A hasn’t are added to a ranked list of suggestions. A is now more likely to adopt items in the list than the items of another random person (Schafer et al., 2001).

3.1.2 Content-Based Recommendation

Another approach to provide relevant recommendations to a user is the use of content-based recommendation systems. This is a type of system that recommends items relevant to other items by comparing the details and descriptions of the items themselves. This can be extended to suggest items for a user by comparing a content-based description of the user’s preferences with the descriptions of the items (Pazzani and Billsus, 2007).

²<https://support.twitter.com/articles/49309-using-hashtags-on-twitter>

³www.last.fm

⁴www.amazon.co.uk

A key issue with content-based filtering is that the recommendations can only be as accurate as the algorithm used to derive a user’s profile. There are a number of algorithms available to build user profiles, depending upon the context, but essentially a content-based profile is created using a weighted vector of item features. The weights mark the importance of each feature to the user, and can be computed from individually rated content vectors.

Cantador et al. (2010) studied and evaluated a number of content-based recommendation models based upon the premise of user and item profiles being described in terms of weighted lists and tags. Through their experiments they found that models that focused on user profiles outperformed the models oriented towards item profiles in nearly every test. They go on to suggest that a better way of profiling users would be through the use of tag clustering.

3.1.3 Relevance Feedback

Relevance feedback is a process that was originally designed for information retrieval, and works on the assumption that a user can not always correctly encapsulate into a query what it is they are searching for. It works by allowing a user to create an initial query to which an initial set of results is returned. Out of these initial results, the user can then mark certain results as relevant or irrelevant, and this information is then submitted and used to refine the original query and return more relevant results to the user (Salton and Buckley, 1990).

Instead of limiting recommendation systems to the accuracy of their classifiers, a common approach is incorporate relevance feedback techniques. Utiyama and Yamamoto (2006) showed that it is possible to combine collaborative filtering, content-based filtering and relevance feedback techniques into one system to provide better recommendations.

3.2 Hashtag Recommendation Research

Even though providing hashtag recommendations and suggestions is still a new and largely unexplored field, there have been several efforts to improve the hashtag experience for Twitter users.

3.2.1 Current Twitter Hashtag Implementation

The current hashtag system on Twitter (Figure 2) uses a non-personalised auto-complete tool to provide suggestions to the user. Whenever a hash symbol (#) is typed in the tweet composer, the system simply suggests hashtags starting with the letters that the user has typed so far. These suggestions are

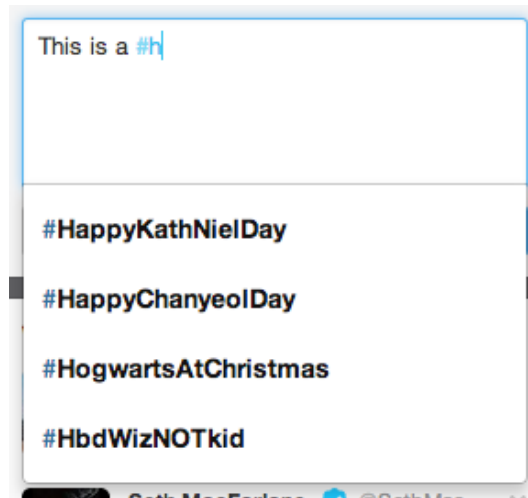


Figure 2: Twitter’s current hashtag suggestion system.

chosen from a tiny subset of hashtags, taken from a mixture of those currently trending⁵ and from the user’s history. Whilst better than not having suggestions at all, this system is only truly useful in one of two specific use cases: when the user knows the starting letters of a trending hashtag they want to use, or are trying to recall a hashtag that they have previously used. This system does not help the user choose the correct hashtag for the content of their tweet.

3.2.2 Comparing Tweets To Other Tweets

By assuming that the primary purpose of hashtags is to categorise tweets and improve searching (as Twitter envisioned), Zangerle et al. (2011) created a system that recommends hashtags for a tweet by taking tweets from other tweets that are textually similar to the query. The similarity between tweets is calculated with the TF-IDF (term frequency – inverse document frequency) model. The hashtags are then extracted from the similar tweets, ranked according to how similar the tweets were to the original query, and returned as a list of suggestions to the user. A number of different ranking algorithms were tested, but this was found to be the most successful.

3.2.3 Creating Personalised Recommendations

After studying the advantages of providing personalised recommendations in retail situations on a per-user basis, Kywe et al. (2012) realised that a similar approach towards hashtags could prove fruitful. Hashtag use varies from user

⁵Trending hashtags are those with the highest rise in usage within a given time period.

to user, with some users using the latest trending hashtags, other users only using a specific set or type of tag, and with some users barely using them at all. They proposed a personalised hashtag recommendation system that considers both user preferences and the query tweet content: the system creates a ranked list of hashtags from both the most similar users and most similar tweets. This gave promising results, although it was noted that this may not be the best recommendation system for all types of tweets and hashtags.

Shepitsen et al. (2008) used a hierarchical agglomerative clustering algorithm to profile users and provide personalised recommendations in collaborative tagging systems. They found that clusters of tags can be effectively used to ascertain a user's interests, which could then be used in a traditional content-based recommendation approach. This technique worked well, particularly for dense folksonomies such as Last.fm.

3.2.4 Overcoming Hashtag Duality

Observers of social media have realised that hashtags play a dual role within the online microblogging communities, such as Twitter. On one hand, hashtags fulfil the design goals that Twitter created them to accomplish (bookmarking and improving search); on the other hand, however, they serve as a badge of community membership: connecting users together. Yang et al. (2012) took this duality into account when attempting to create a hashtag recommendation system by training a SVM (support vector machine) classifier with a variety of features taken from the tweet metadata to overcome the duality and suggest relevant hashtags.

3.3 Broader Classification in Twitter

Twitter is a thriving metropolis of users expressing themselves on a daily (and often more frequent) basis, and has grown exponentially⁶ in size since its inception in 2006. Due to this, the data that it contains has caught the attention of researchers throughout computer science and even other disciplines. Whilst the concept of recommending hashtags is relatively unexplored, there have been many other classification experiments run with Twitter data.

3.3.1 Categorising Tweets

Sriram et al. (2010) proposed an approach to classify tweets into 5 general categories: *news*, *opinions*, *deals*, *events* and *private messages*. This was achieved

⁶www.bloomberg.com/news/2013-10-15/twitter-revenue-more-than-doubles-in-third-quarter.html

by training a classifier using a small set of specific features from each tweet, instead of using the traditional “Bag-Of-Words” (BOW) text classification method. The BOW approach is centred around counting occurrences of words in the text, but in the case of Twitter and its 140 character limit, it is very rare that words are actually repeated in a tweet.

3.3.2 Categorising Users

Another approach to deciphering the vast quantity of data on Twitter is to classify the users themselves. Twitter has become a powerful platform for people posting content about events, and as such it would be useful to automatically establish *who* is participating in these events. By taking a number of features from each user account and passing them through a K-Nearest Neighbours (KNN) algorithm, De Choudhury et al. (2012) developed a system that would classify a user’s behaviour into one of three categories: *organisations, journalists/media bloggers* and *ordinary individuals*.

4 Requirements

There are two main requirements the system in this project is aiming to fulfil. It must:

- Allow users to compose and publish tweets whilst suggesting hashtags relevant to the content of their tweets.
- Allow users to search for a hashtag and view related tweets, including those that don't contain that hashtag.

4.1 Functional Requirements

1. The system must allow the user to log in and publish tweets to their Twitter account.
2. The system must provide hashtag recommendations as the user is creating a tweet.
3. The system must perform a hashtag search through a large dataset of tweets and return all relevant tweets, including those that do not contain the search query.
4. The system must use information from a large dataset of tweets to generate a model representing each hashtag.
5. The system must be able to compare tweets against its representational hashtag models.
6. *Optional:* The system must be able to update its classification models using information from the live Twitter stream.
7. *Optional:* The system must provide probabilities for how likely a hashtag is to be related to a tweet.

4.2 Non-Functional Requirements

1. The system must be accessible via a web interface.
2. The system must be responsive and easy to use.
3. The system must be able to perform searches quickly.
4. The system must be able to make hashtag recommendations quickly.
5. The system must be able to produce visualisations to provide an easy way to interpret the hashtag recommendations/assignments.
6. *Optional:* The system must be accessible via mobile web browsers.

5 Design

5.1 Architecture

5.1.1 Processing Framework

5.1.2 Classifier

5.2 Interface

5.2.1 API Endpoints

5.2.2 Web Interface

6 Implementation

6.1 Stream Processing

6.2 Classification System

6.2.1 Training

6.2.2 Classifying

6.2.3 Optimisation

6.2.4 RESTful Interface

6.3 Web Application

6.3.1 Framework

6.3.2 Authentication

6.3.3 Connecting to Classifier

7 Testing

7.1 Strategy

7.2 Unit Testing

7.3 Classification Testing

7.4 User Interface Testing

7.5 Requirements Validation

8 Conclusion

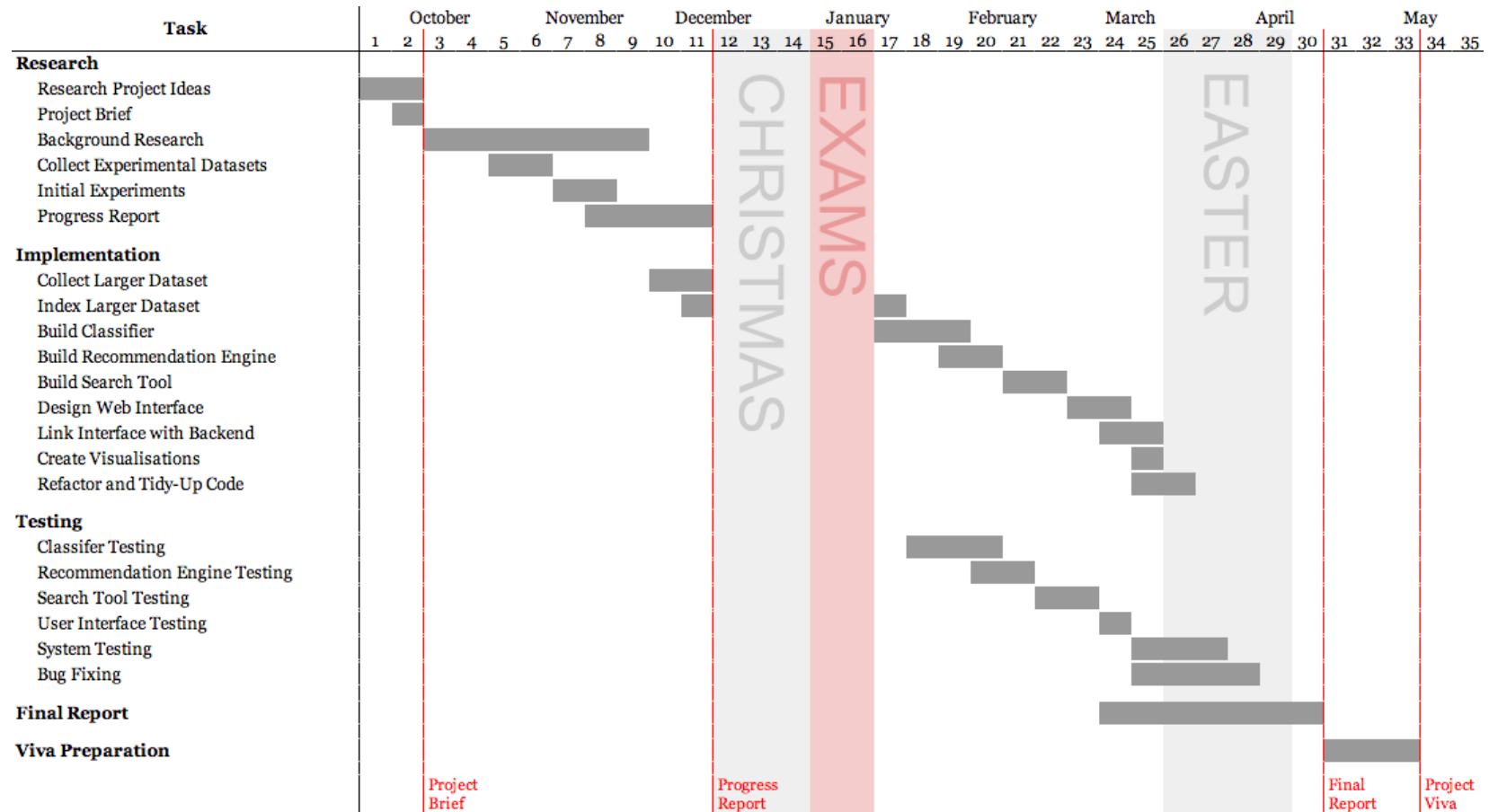
8.1 System Evaluation

8.2 Project Management

8.3 Future Work

A Time Management

This is a Gantt chart showing the scheduled progression through the different aspects of the project.



B Testing Plans & Results

C REST API Interface

D Original Project Brief

References

- Cantador, I., Bellogín, A. and Vallet, D. (2010). ‘Content-based Recommendation in Social Tagging Systems’. *Proceedings of the Fourth ACM Conference on Recommender Systems*. RecSys ’10. Barcelona, Spain: ACM, pp. 237–240.
- De Choudhury, M., Diakopoulos, N. and Naaman, M. (2012). ‘Unfolding the Event Landscape on Twitter: Classification and Exploration of User Categories’. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. CSCW ’12. Seattle, Washington, USA: ACM, pp. 241–244.
- Goldberg, D., Nichols, D., Oki, B. M. and Terry, D. (1992). ‘Using Collaborative Filtering to Weave an Information Tapestry’. *Communications of the ACM* 35.12, pp. 61–70.
- Kywe, S. M., Hoang, T.-A., Lim, E.-P. and Zhu, F. (2012). ‘On Recommending Hashtags in Twitter Networks’. *Proceedings of the 4th International Conference on Social Informatics*. SocInfo ’12. Lausanne, Switzerland: Springer-Verlag, pp. 337–350.
- Pazzani, M. J. and Billsus, D. (2007). ‘The Adaptive Web’. Ed. by P. Brusilovsky, A. Kobsa and W. Nejdl. Vol. 4321. Berlin, Heidelberg: Springer-Verlag. Chap. Content-based Recommendation Systems, pp. 325–341. ISBN: 978-3-540-72078-2.
- Salton, G. and Buckley, C. (1990). ‘Improving Retrieval Performance by Relevance Feedback’. *Journal of the American Society for Information Science* 41.4, pp. 288–297.
- Schafer, J. B., Konstan, J. A. and Riedl, J. (2001). ‘E-Commerce Recommendation Applications’. *Data Mining and Knowledge Discovery* 5.1-2, pp. 115–153.
- Shepitsen, A., Gemmell, J., Mobasher, B. and Burke, R. (2008). ‘Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering’. *Proceedings of the 2008 ACM Conference on Recommender Systems*. RecSys ’08. Lausanne, Switzerland: ACM, pp. 259–266.
- Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M. (2010). ‘Short Text Classification in Twitter to Improve Information Filtering’. *Proceedings of the 33rd International ACM SIGIR Conference on*

Research and Development in Information Retrieval. SIGIR '10. Geneva, Switzerland: ACM, pp. 841–842.

Utiyama, M. and Yamamoto, M. (2006). ‘Relevance Feedback Models for Recommendation’. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. EMNLP '06. Sydney, Australia: Association for Computational Linguistics, pp. 449–456.

Xu, S., Bao, S., Fei, B., Su, Z. and Yu, Y. (2008). ‘Exploring Folksonomy for Personalized Search’. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore, Singapore: ACM, pp. 155–162.

Yang, L., Sun, T., Zhang, M. and Mei, Q. (2012). ‘We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption?’ *Proceedings of the 21st International Conference on World Wide Web*. WWW '12. Lyon, France: ACM, pp. 261–270.

Zangerle, E., Gassler, W. and Specht, G. (2011). ‘Recommending #-Tags in Twitter’. *Proceedings of the Workshop on Semantic Adaptive Social Web*. SASWeb '11. Girona, Spain: CEUR Workshop Proceedings, pp. 67–78.