Content-based Hashtag Recommendation Methods for Twitter

Dipesh Poudel '22 (dpoudel18@earlham.edu)

Department of Computer Science, Earlham College

1. Abstract

This project designs and evaluates different content-based methods (Hashtag Frequency, Naïve Bayes model, and KNN-based cosine similarity) for Twitter hashtag recommendation. The results shows that Naive Bayes model outperforms other two models on almost all metrics with an average accuracy score of 0.83.

2. Introduction

- A hashtag recommendation system is an important tool to organize similar content together for topic categorization.
- A very little research has been done on evaluating the performance of different existing models using the same dataset and the same evaluation metrics.
- This project first designs three content-based methods (Tweet similarity using hashtag frequency, Naïve Bayes model, and KNN-based cosine similarity) for hashtag recommendation.
- It, then, evaluates the performance of these content-based methods (Tweet similarity using hashtag frequency, Naïve Bayes model, and KNN-based cosine similarity) using 5 evaluation metrics namely Precision, Recall, F1 Score, Hit Rate and Hit Ratio.
- The final product is a web application which lets the user input their tweet, and recommend hashtags for their tweet using these three models.

References

- Dipesh Poudel. Content-based Hashtag Recommendation for Twitter. https://code.cs.earlham.edu/dpoudel18/senior-capstone-hashtag-recommendation

3. Software Architecture

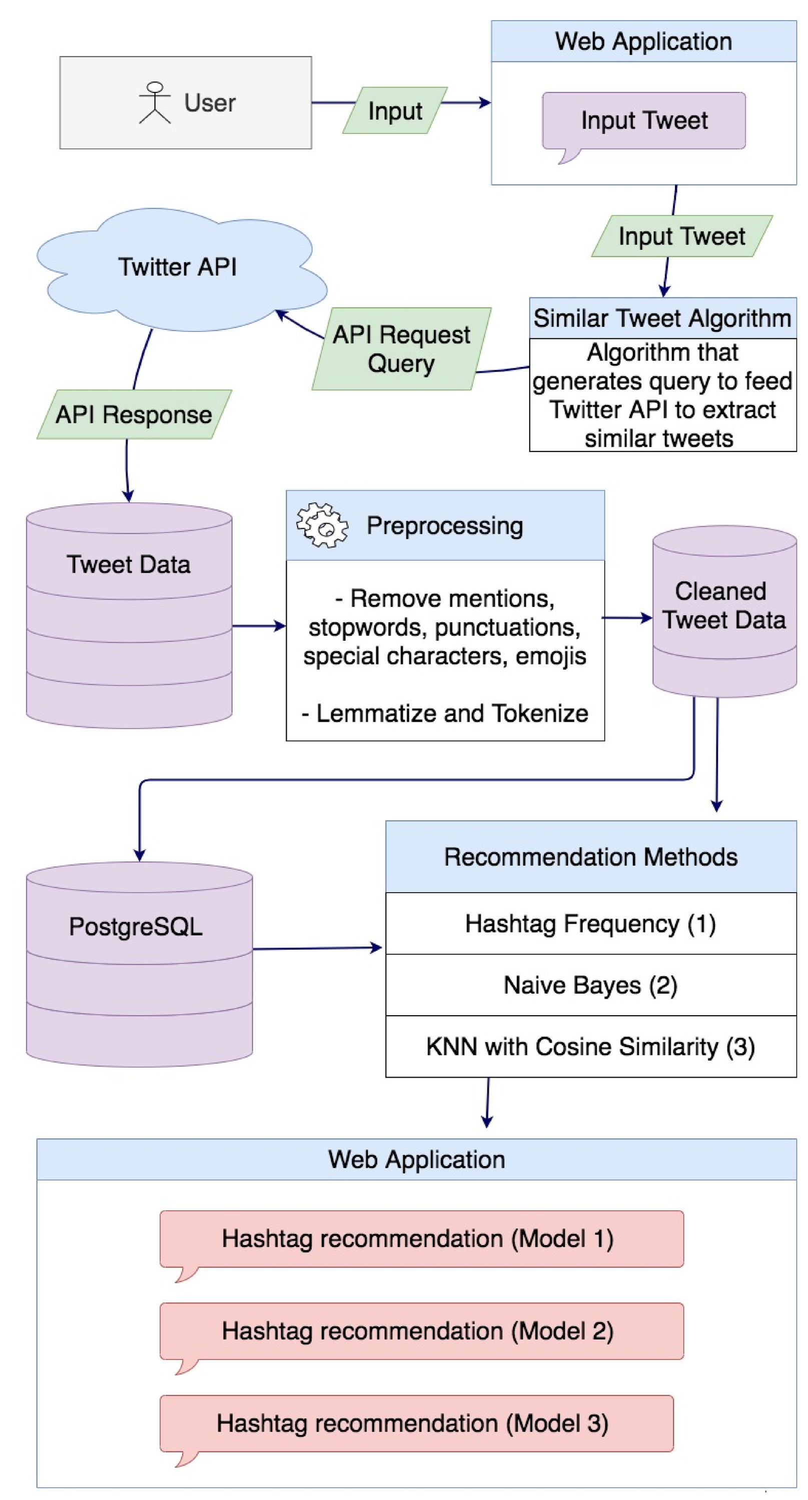

Figure 1:Software Architecture Diagram

4. Evaluation Framework

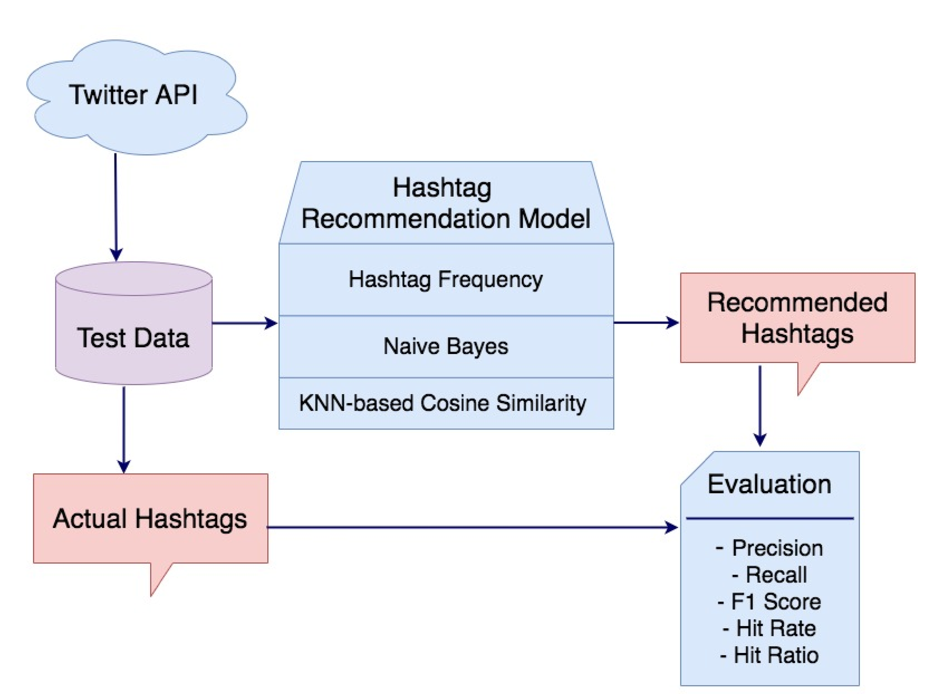

Figure 2:Evaluation framework

5. Data Pre-processing

- This project make use of *Regex* and *NLTK* library to clean the raw twitter data as shown below

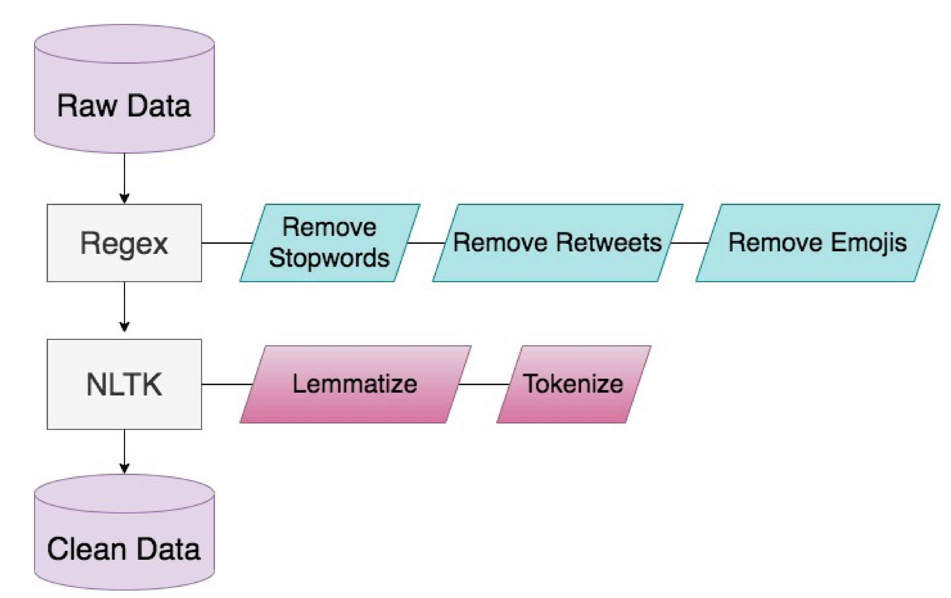

Figure 3:Pre-processing framework

6. Recommendation Methods

Hashtag Frequency

- The hashtags in the retrieved similar tweets are selected to be candidate hashtags. which are, then, ranked based on which hashtags are present more and top-k hashtags are recommended.

Naive Bayes

- Hashtags are recommended using Bayes' Theorem as illustrated in the given formula.

$$P(Hashtag \;/\; Tweets) = \frac{P(Hashtag) \cdot P(Tweets \;/\; Hashtags)}{P(Tweets)}$$

- The top-k hashtags with the higher probabilities are recommended to the user.

KNN-based Cosine Similarity

- First, the cosine similarity between the test tweet and the train tweet is calculated.

$$CosineSimilarity(T_{train}, T_{test}) = \frac{T_{train} \cdot T_{test}}{\|T_{train}\| \cdot \|T_{test}\|}$$

.

- Then,K-nearest neighbor of the test tweet ranks hashtags.

7. Evaluation Results

- Model was evaluated and verified, by the described evaluation model, using 100 high quality tweets from verified Twitter personality with at least 3 hashtags on each tweets ensuring the quality of the tweet.
- The results of the evaluation can be seen below.
- Naive Bayes model outperforms other models in almost all metrics.
- Cosine Similarity model performs the worst

Model	Precision	Recall	F1-Score	Hit-Rate	Hit-Ratio
HF	0.69	0.87	0.76	0.91	0.88
NB	0.72	0.86	0.78	0.93	0.9
CS	0.65	0.81	0.72	0.84	0.80

Table 1:Evaluation Results

8. Conclusion / Future Work

- It is observed that even simple content-based recommended methods like Naive-Bayes produces good average accuracy score of 0.83.
- This project was conducted with a fairly small datasets due to Twitter API restrictions
- There is potential to improve the performance of the models if an robust date pipeline is created with PostgreSQL database to store and query real-time tweets.

9. Acknowledgement

- I would like to acknowledge Dr. Charlie Peck and Dr. David Barbella for the immense help throughout the project.