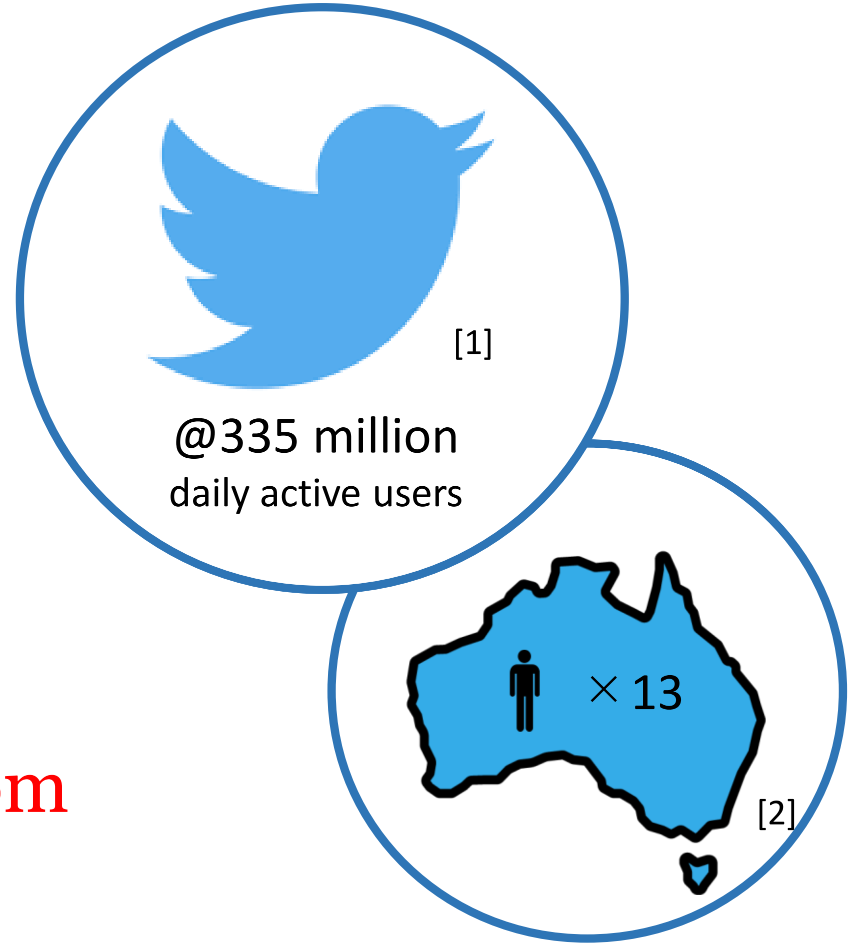
Tweet filtering based on user interests using doc2vec & unsupervised clustering

IFN712 Research in IT Practice

n10423010 Xinruo Wang

Project Background

- 500 million tweets posted on Twitter daily
- People want to finding specific tweets which interest them
- Difficult to find relevant tweets
  - A large volume of tweets
  - Unannotated data
- How to automatically detect user's interests from their tweets and retrieve related posts for them



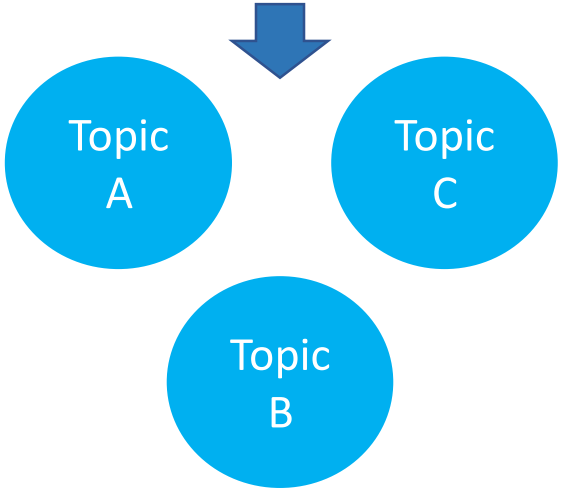
Project Overview (1/2)

- Group tweets based on topics (user interests) using doc2vec, tf-idf and unsupervised clustering techniques

- Traditional tf-idf methods fail in tweet topic categorisation
  - tweets are too short & noisy
  - tf-idf ignores semantics of words
  - tf-idf loses ordering of words

- doc2vec can be a better feature representation technique


[3]



Project Overview (2/2)

- The research will answer two questions:
  - whether doc2vec can capture tweet topics better than tf-idf
  - what is the most effective clustering method to group tweets together

Related Works (1/2)

- To apply tweet clustering, the first step is to use a feature representation method to convert the texts into vectors
- Tf-idf is a common method applied in many research

Godfrey et al. (2014) [4]

Feature representation: tf-idf
Clustering: k-means and NMF
Evaluation: Non-Negative Matrix Factorization (NMF)
Data: labelled 30,000 tweets about World Cup
Results: both clustering had similar performance

+ Use DBSCAN to remove noisy tweets as pre-processing

- Problem of tweet length remains

- Ordering and semantics of words are ignored

- Produced matrices are high-dimensional and sparse

Related Works (2/2)

- To solve the problems of tf-idf, do2vec algorithm was developed

Le & Mikolov (2014)) [5]

Feature representation: doc2vec, tf-idf
Data: triplets of paragraphs obtained from search queries
Evaluation: Use distance measures to find which two paragraphs are from the same query
Results: doc2vec outperformed tf-idf by 30%

+ Can be applied to different length of documents

Curiskis et al. (2020) [6]

Feature representation: doc2vec, tf-idf etc.
Clustering: k-means, hierarchical etc.
Data: two labelled Twitter datasets etc.
Evaluation: NMI matrix etc.
Results: doc2vec performed better than tf-idf

- Require labelled datasets, which does not suit real scenario

Methods (1/3)

1.Data collection
- 14million tweet data set
- use #music
- remove retweets

2.Data pre-processing
- remove URL, emoj, HTML entities, @, hashtags
- lowercase
- remove digits
- remove punctuation
- lemmatisation
- remove stopwords
- tokenisation
- remove tokens with < 3 char
- Remove short tweets with < 4 tokens

3.Feature representation
- TF-IDF
- doc2vec

4.Clustering
- k-means
- Hierarchical agglomerative

5.Evaluation
- elbow method
- silhouette coefficient

6.Tweet filtering

7.Reporting

Methods (2/3)

- After pre-processing
  - # of tweets: **3974**
  - # of users: **1799**
  - # of tokens in each tweet
    - Average: **7.11**
    - Median value: **7**
    - Standard deviation: **2.54**

Methods (3/3)

- Silhouette score

The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters

For each data point i, define:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

represent the average distance of the point i to all the other points that belongs to the same cluster Ci.

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

which represent the average distance of the point i to all the other points in the next nearest cluster

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

- $-1 \leqq s(i) \leqq 1$
- s(i) = 1 is a good indicator of good clusters

Results & Discussion (1/4)

- In general, doc2vec outperformed tf-idf
- doc2vec combined with k-means gave the best performance in all cases

Table 1. Optimal number of clusters / Comparison of silhouette scores

HAC: Hierarchical Agglomerative Clustering

	Optimal number of clusters (k)	
Feature Rep.	**K-means**	**HAC**
TF-IDF	3	5
Doc2vec	8	3

K = 3	**Silhouette Score**	
Feature Rep.	**K-means**	**HAC**
TF-IDF	0.0487	0.0394
Doc2vec	0.4692	0.4663

K = 5	**Silhouette Score**	
Feature Rep.	**K-means**	**HAC**
TF-IDF	0.0517	0.0522
Doc2vec	0.4133	0.3300

K = 8	**Silhouette Score**	
Feature Rep.	**K-means**	**HAC**
TF-IDF	0.0671	0.0421
Doc2vec	0.3403	0.3048

Results & Discussion (2/4)

- Visualisation of best model using doc2vec & k-means (optimal k=8)
  - Each cluster has similar distribution of silhouette score
  - Clusters are close to each other but not scattered

Fig 1. Silhouette plot

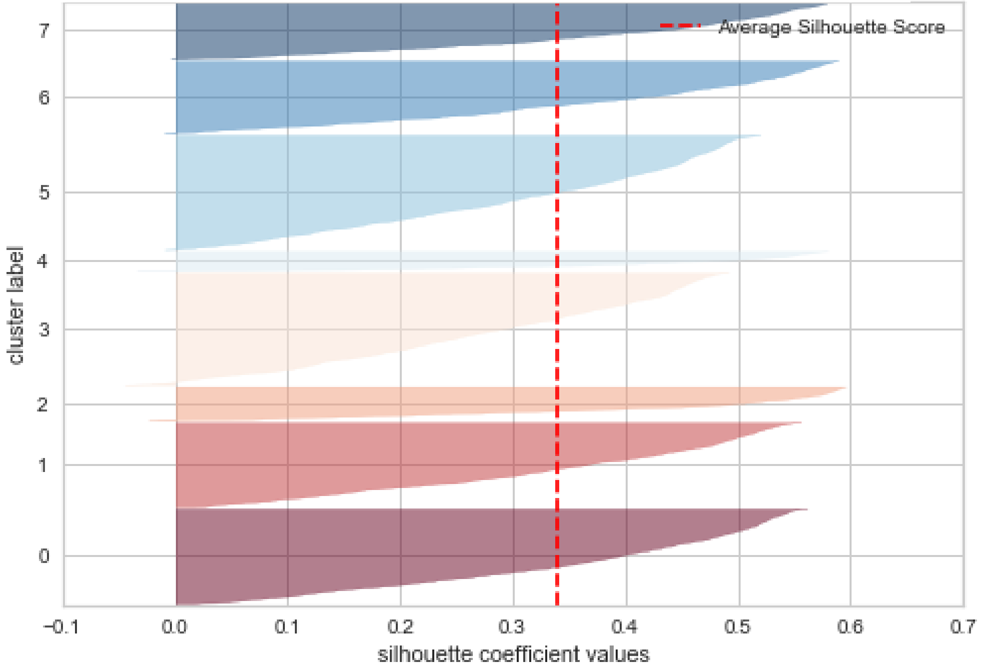
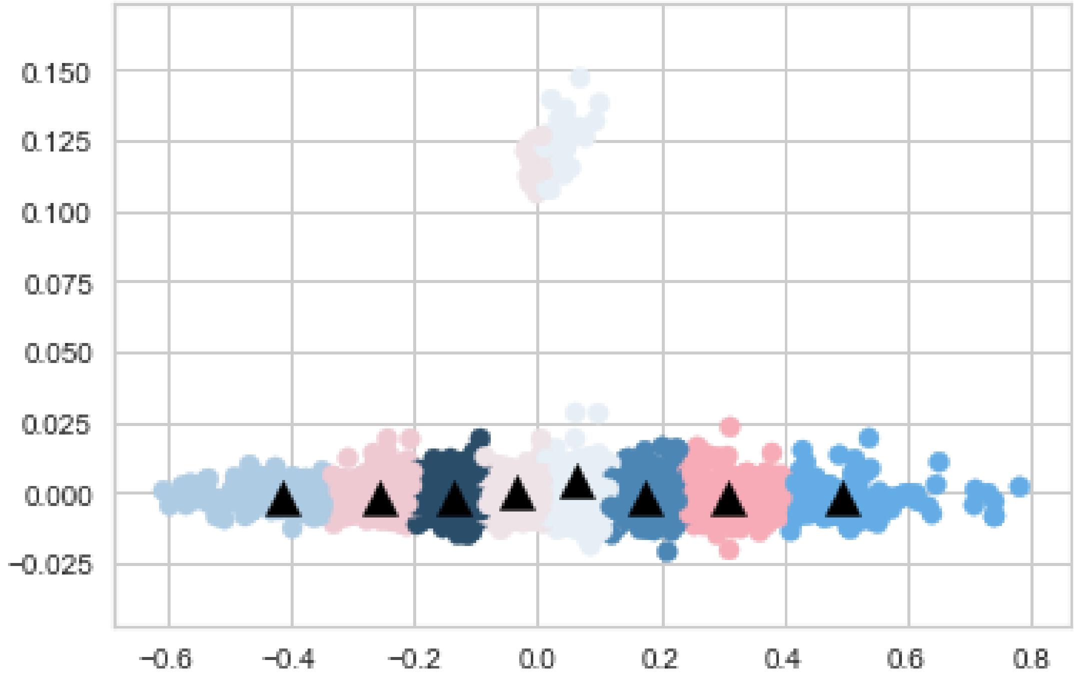


Fig 2. Principal Component Analysis (2d)



Results & Discussion (3/4)

- Determine cluster topics
  - It is difficult to interpret what the true topics are from manual inspection of top keywords in each cluster

Cluster 2

- love
- music
- new
- show
- today
- get
- one
- time
- like
- night

Cluster 3

- new
- song
- post
- music
- follower
- want
- retweet
- love
- video
- check

Results & Discussion (4/4)

- Summary of limitations
  - Do2vec only resulted in the maximum silhouette score of 0.4692
  - Cluster topic interpretability is very low: top-keywords are not meaningful

- Improvement suggestions
  - Add a pre-processing step to detect synonyms (u & you, 2night & tonight) to reduce dimensions of corpus
  - Use N-gram tokeniser to keep phrases to improve topic interpretability

- Future study
  - Investigate how to interpret cluster topics when using doc2vec

Conclusion

- Take-away messages
  - Doc2vec is better at capturing tweet topics than tf-idf
  - Doc2vec combined with k-means gave the best performance

- Novelty of the research
  - Require no labelled dataset

- Significance
  - Use to retrieve relevant tweets for users
  - Apply to analyse follower interests for marketing purposes (business perspective)

References

[1] Image source: https://www.stickpng.com/img/icons-logos-emojis/tech-companies/twitter-logo

[2] Image source: https://www.pinclipart.com/pindetail/xTomTi_accelerated-reader-bookfinder-logo-australia-map-landscape-hd/

[3] Image source: https://www.thebalanceeveryday.com/twitter-terms-for-beginners-896935

[4] Godfrey, D., Johns, C., Meyer, C., Race, S., & Sadek, C. (2014). A case study in text mining: Interpreting twitter data from world cup tweets. arXiv preprint arXiv:1408.5427.

[5] Le, Q. V., & Mikolov, T. (2014). *Distributed representations of sentences and documents. Proceedings of the 31th international conference on machine learning, ICML 2014, Beijing, China, 21–26 June 2014*1188–1196.

[6] Curiskis, S., Drake, B., Osborn, T., & Kennedy, P. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing and Management*, *57*(2). https://doi.org/10.1016/j.ipm.2019.04.002