# Utility of Topic extraction on Customer Experience data

Kiran Karkera
Bridgei2i Analytics Pvt Ltd
Bangalore, India
kiran.karkera@gmail.com

## ABSTRACT

Unsupervised topic models are capable of deriving important bags of words (which signify topics) from a corpus of text data. We attempt to use unsupervised topic models on customer experience data. We learn that the topics extracted by topic models are hard to interpret if the number of documents in the corpus is small, and if the corpus uses a small number of unique terms. We further explore a novel method of visualizing topic words which may help determine if the data has enough semantic diversity that will enable unsupervised topic models to extract meaningful topics.

**KEYWORDS**: Topic Models, LDA, Word2Vec, T-SNE, Topic coherence

## INTRODUCTION

Our customers have corpora of text data related to customer service, such as employee surveys and customer feedback in the form of reviews, customer surveys and chat agent notes of customer calls. Categorization or segmentation of the text data is often used to enhance their understanding of the data, since it is too large to be read quickly.

Since it is rare to find a training set of categorized data (such as a spam training corpus where the email is categorized into spam or non-spam), unsupervised approaches such as LSA[1] or LDA[2] are used to segment data.

## SHORT INTRODUCTION TO TOPIC MODELS

Before we dive into topic models, let us review the associated terms so that we can follow the discussion

- A “topic” is a bag of words. For example, a document describing a cricket match may find {Bowler, batsman, runs} as one of the topics.
- Within each topic, the words have a probability distribution.
- A document can contain multiple topics. For example, a document about sports administration will find topics about sport and administration.
- Unlike classification or clustering which assigns a single class to a document, topic models will find a probability distribution over all topics for a given document

Given a fixed number of topics ***T***, the LDA model will learn ***T*** bags of words, one for each topic. Typically, after a topic model generates the “bag of words”, an analyst familiar with the domain will choose a human-friendly name for each of those topics.

## INTERPRETING THE EXTRACTED TOPICS

Once the LDA model is trained on a particular corpus, we can observe the resulting topics. Often we discover that the words in the topic are not related, and the topic model isn’t segmenting the data into topics that are meaningful to human consumption. We can employ techniques such as noun phrase extraction, improved stop-word removal, term frequency and inverse document frequency pre-processing as well as changing the number of topics, in an attempt to improve the interpretability of the topics.

In some cases, despite the application of the aforementioned procedures, the topic bag of words are not interpretable. The results could be deficient for the following reasons:

- The same words reappear as top words in multiple topics.
- There is no significant relation or similarity between words in the same topic

Along with human evaluation of the topics, we also attempt to evaluate if the topics segmentation is meaningful by using the following measures of coherence such as perplexity[3], and the Jiang-Conrath distance metric[4]. Chang et al [5] showed that the perplexity, which computes the held-out likelihood, is an error prone way of topic coherence as it often correlates negatively with human annotated results.

An example of topic extracted that is hard to interpret for humans can be seen in Table 1

| Topic 0 | Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|---|
| better | chance | aspect | chance | employee |
| Effort | company | basis | employee | good |
| ijp | distance | environment | ijp | gradation |
| industry | employee | finance | job | ijp |
| manager | environment | growth | level | implement |
| organization | experience | manner | mobility | performer |
| provide | job | performance | people | process |

Table 1: Topic words from a customer service corpus

We opine that such results could be due to the following causes:

- Fewer number of documents in the corpus.
- The number of words in each document is low. This is often observed in customer and employee surveys where the respondent records “good manager” or a similar one or two word response.
- The number of unique words in the corpus is low.

## VISUALIZING THE EXTRACTED TOPICS

Although humans have an instinctive feel for the relatedness or similarity of words in a topic extracted by our model, it turns out that humans annotating the topics have a high amount of disagreement between each other. We therefore decided to explore an approach of visualizing words to see if there exists a natural separation between bags of words.

Word2vec (WTV) [6] is a neural network model that can learn a high dimension representation of words based on their pattern of occurrence. The learned vectors place semantically similar words close to each other in a high dimension vector space. We used a word2vec model trained on 100 billion words on a Google news corpus. The trained model can be downloaded here[7].

The WTV model converts a word to a 300-dimension vector representation. We then use the T-SNE [8] manifold learning method to reduce the 300-dimension vector to a 2-dimension space so that it can be visually plotted.

| | **Topic 0** | **Topic 1** | **Topic 2** |
|---|---|---|---|
| **1** | india | genome | sport |
| **2** | russia | virus | cricket |
| **3** | china | dna | football |
| **4** | algeria | cell | player |
| **5** | delhi | biology | coach |
| **6** | washington | disease | field |
| **7** | paris | pathology | olympic |
| **8** | france | rna | games |
| **9** | berlin | disease | nutrition |

Table 2: human-created topic bag of words

We first explore a human-annotated bag of words and observe the plot in Figure 1. Each circle indicates a word and the colour indicates a different topic. It must be clarified that the X and Y axes are relative (not absolute) and are obtained by dimension reduction using T-SNE.

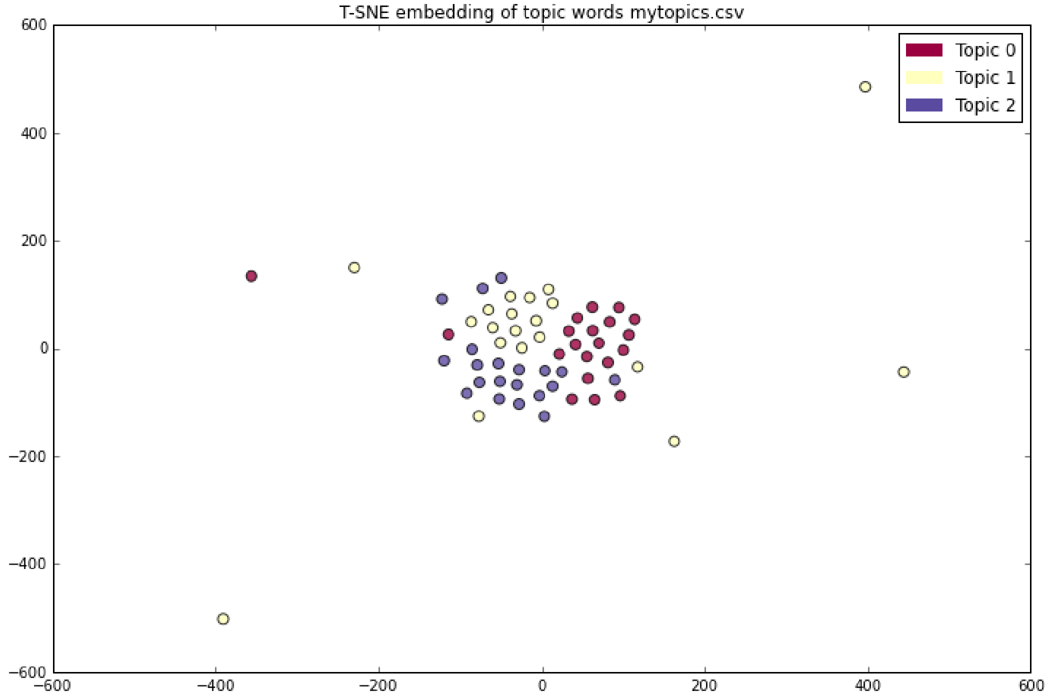

Figure 1: plot bag of words displayed in Table 2.

It is fairly clear from the plot that the WTV vectors (in their reduced dimensionality) are capable of differentiating the words from different topics.

We now use the topics generated from a dataset of articles published in the journal Science from 1980-2002 in Figure 2. 50 topics were generated and the data points are the top words from a few topics. Although the points are not linearly separable, it may be observed that some, if not all points of the same colour (topic) are clustered together.

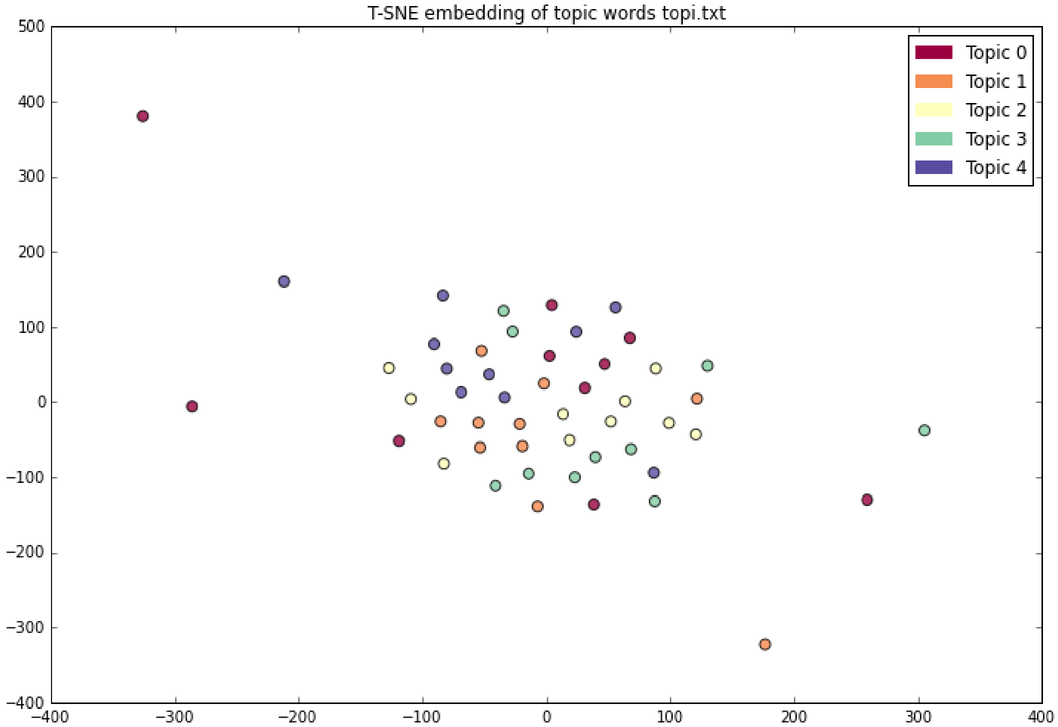

Figure 2: Topics extracted from the Journal Science (1980-2002)

| | **Topic 0** | **topic 1** | **topic 2** | **topic 3** | **topic 4** |
|---|---|---|---|---|---|
| **1** | computer | chemistry | cortex | orbit | infection |
| **2** | methods | synthesis | stimulus | dust | immune |
| **3** | number | oxidation | fig | jupiter | aids |
| **4** | two | reaction | vision | line | infected |
| **5** | principle | product | neuron | system | viral |
| **6** | design | organic | recordings | solar | cells |
| **7** | access | conditions | visual | gas | vaccine |
| **8** | processing | cluster | stimuli | atmospheric | antibodies |
| **9** | advantage | molecule | recorded | mars | hiv |
| **10** | important | studies | motor | field | parasite |

Table 3: Topic bag of words extracted from the journal Science(1980-2002)

## TOPICS FROM CUSTOMER EXPERIENCE DATA

We now look at topic extraction on customer experience data. This corpus consists of call centre agent notes as well as customer survey results from a fitness brand. The following plot consists of topics extracted from text written by a call centre agent for inbound calls.

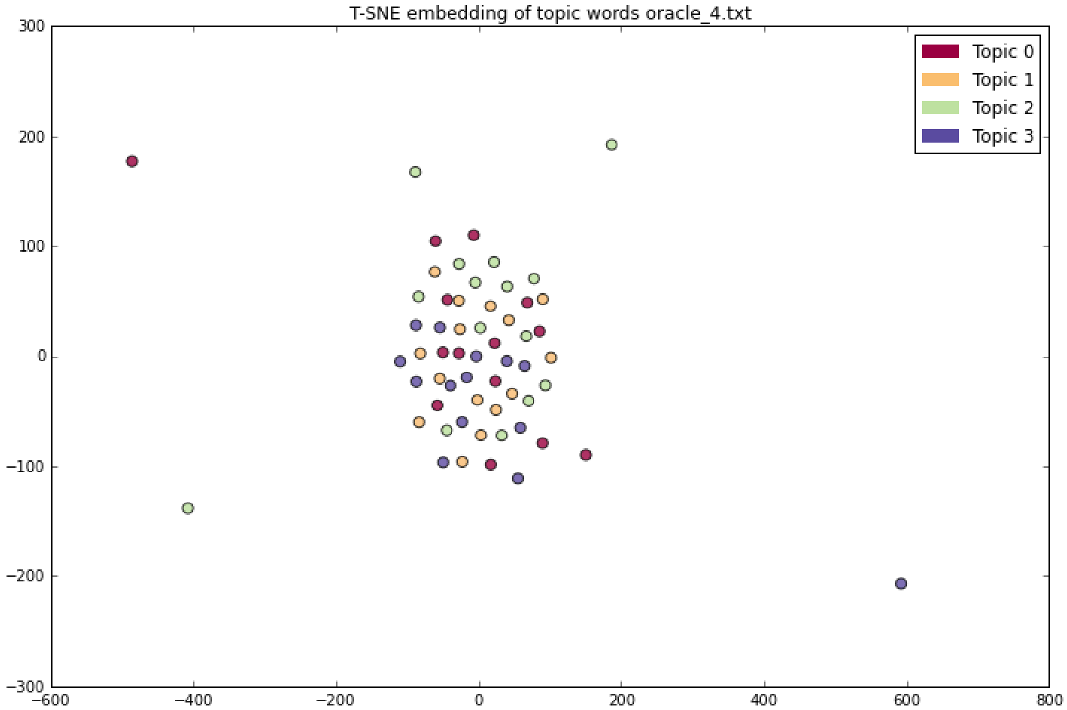

Figure 3: Plot of topics extracted from call centre agent notes

| | **Topic 1** | **Topic 2** | **Topic 3** | **Topic 4** |
|---|---|---|---|---|
| **1** | vitamin | usable | confirmation | program |
| **2** | cold | multivitamin | shipment | stop |
| **3** | note | unmanned | dvd | shipment |

|  |  |  |  |  |
|---|---|---|---|---|
| **4** | continuity | reason | app | calling |
| **5** | cancelled | installment | product | installment |
| **6** | email | sell | continuity | beach |
| **7** | challenge | continuity | process | body |
| **8** | customer | instruction | instruction | cancelled |
| **9** | refund | club | status | instruction |
| **10** | membership | return | package | credit |

Table 4: Topic words extracted from call centre agent notes

We can observe that there is little coherence among the words in each topic, and the word2vec representation of the words share the same confusion.

We plot another dataset consisting of (topics extracted from) customer survey results. It can be observed that there is insufficient semantic separation between bags of words in each topic.

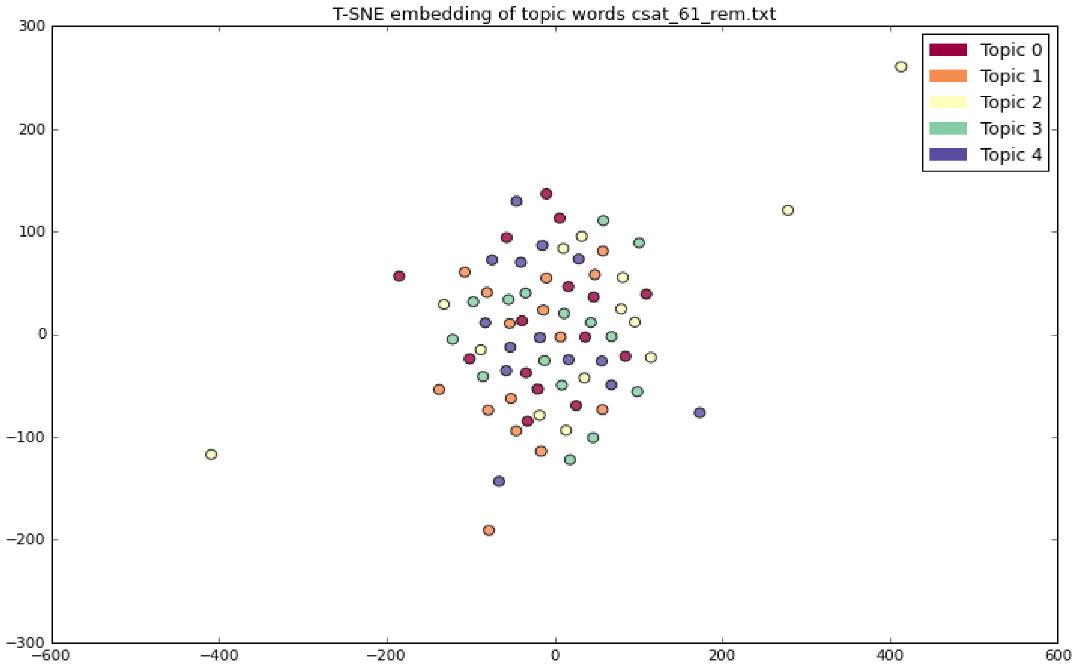

Figure 4: Topic words from customer survey results

|  | Topic 1 | topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|---|
| 1 | people | reply | phone | supervisor | challenge |
| 2 | information | sent | shipping | spoke | account |
| 3 | resolution | resolution | date | understanding | information |
| 4 | point | chat | home | card | program |
| 5 | contact | cancellation | credit | faq | business |
| 6 | line | form | account | request | experience |
| 7 | chat | account | flavor | line | website |
| 8 | faq | information | form | experience | shipping |
| 9 | online | survey | card | accent | people |
| 10 | answer | confirmation | bag | online | company |

Table 5: topics extracted from customer survey results.

## SUMMARY STATISTICS OF WORDS IN DATASET

| Dataset | instances | number of unique words |
|---|---|---|
| Fitness co-customer survey | 13832 | 26115 |
| fitness-call center agent note | 60869 | 57066 |
| employee survey | 164 | 3192 |
| Science (1990-2002) | 21000 | 16M |

Table 6: Summary statistics of text data from datasets.

The data in the Table 6 informs us about the summary statistics of the text datasets. We can observe that for low number of unique words, the interpretability of the extracted topics is quite low. Further, when the number of words increase, techniques such a Term Frequency/Inverse document frequency can be used in pre-processing to increase the importance of relevant words in the corpus, and that results in coherent topics, even if the final number of words input to the topic model is low (10k terms for the Science dataset).

## OBSERVING SEPARATION BETWEEN TOPIC CENTROIDS

The averaged Word2Vec vector for each topic's bag of words can be assumed as the centroid for that particular topic. It is reasonable to assume that topic words that are semantically similar would be sufficiently separated from the centroids of other topics.

We calculated the spatial distance of each centroid from all other centroids, and then took the mean of the distances.

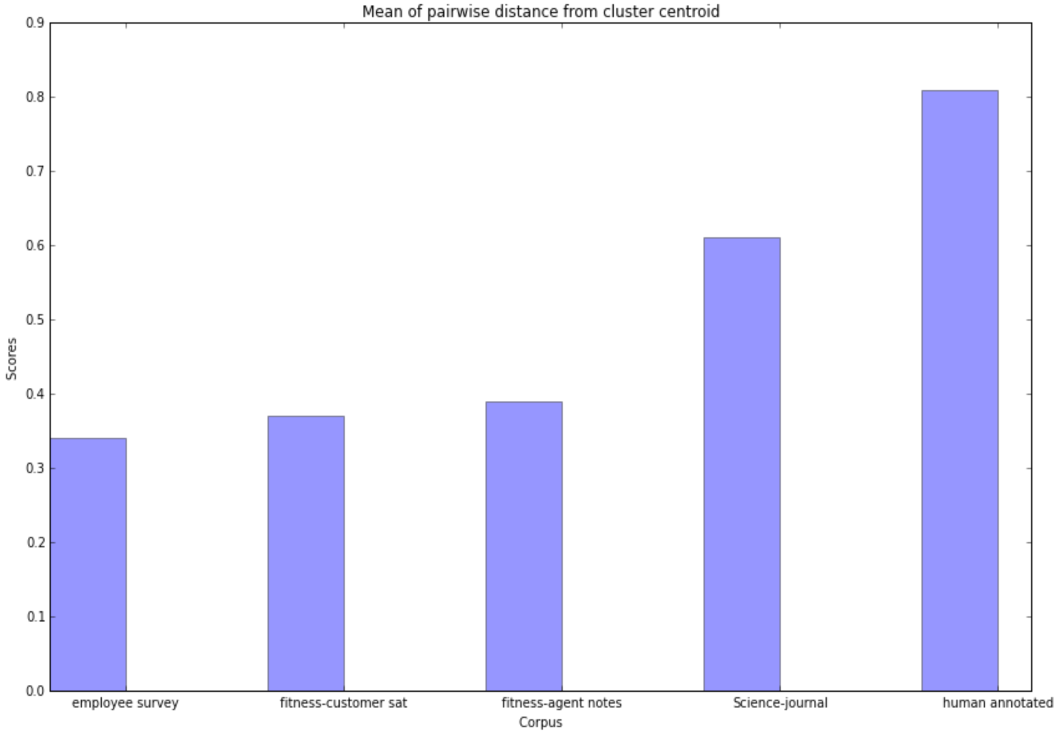

Figure 5: Mean of pairwise distance between topic centroids from multiple datasets

Note that the scores on the y-axis are relative scores. It can be observed that the human annotated topics as well as those extracted from Science show that they have a large spatial distance between topic centroids, which is consistent with the fact that the extracted topic scored high on interpretability.

The scores for the employee survey, Fitness co customer satisfaction survey and Fitness co inbound agent call notes are low. This figure did not change even when the number extracted was varied from 3 to 10, reinforcing the idea that the corpus had little semantic diversity in the first place.

## FURTHER WORK

We plan to investigate the effect of training Word2vec models on a corpus that consists of domain specific data in addition to public corpus such as Wikipedia. This will be useful as customer experience data has terms that refer to products or services which are not present in Wikipedia and therefore will not be available in the word2vec trained model.

## SUMMARY

We compared the output of running unsupervised topic model algorithms on large public datasets and smaller datasets related to the customer experience domain. We observed that the topic words extracted are hard to interpret when the corpus is small, or if the number of unique words are low.

We then generated high dimension representations of the topic words using a Word2vec model trained on a large Google new corpus, and visualized a low dimension representation of it using T-SNE. It was observed that topic models that score highly on human interpretability, show up as distinct clusters. In the absence of clear separation between multiple topics, it is advised that unsupervised topic models are not satisfactory for human interpretation, and advanced techniques such as must link-cannot link[9], interactive topic models[10], or correlated topic models [11] should be used to improve the interpretability of the topics.

## ACKNOWLEDGEMENTS

I would like to thank my team members Mano Bharati J, Prasanna J and Durga Ravichandran for their invaluable assistance in research and development.

## REFERENCES

1. Papadimitriou, Christos; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, Santosh (1998). "Latent Semantic Indexing: A probabilistic analysis" (Postscript). *Proceedings of ACM PODS*
2. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John (January 2003). "Latent Dirichlet allocation". *Journal of Machine Learning Research*
3. Perplexity as means to evaluate topic models (http://qpleple.com/perplexity-to-evaluate-topic-models/)
4. Jiang J; Conrath W (1997) "Semantic Similarity based on Corpus statistics and Lexical Taxonomy", Proceedings of International Conference Research on Computational Linguistics (ROCLING X).
5. Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *NIPS*.
6. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013
7. GoogleNews-vectors-negative300.bin.gz (https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing)
8. van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9:2579-2605, 2008
9. Andrzejewski David, Xiaojin Zhu, Craven Mark. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors, in Proceedings of the 26h International Conference on Machine Learning, Montreal (2009)
10. Hu, Yuening, et al. "Interactive topic modeling." *Machine learning* 95.3 (2014): 423-469.
11. Blei, David, and John Lafferty. "Correlated topic models." *Advances in neural information processing systems* 18 (2006): 147.