Unsupervised Trait-Clustering for Adobe Audience Manager

Adobe Audience Manager allows marketers to create audience segments based on fine-grained user trait data.

Our goals:
- Develop a trait-clustering algorithm
- Make a trait recommendation system
- Build an interactive tool to visualize the results

Adobe provided us with a fully anonymized dataset that has over 137 million rows and 5,196 columns. Each row represents a unique uuid, and each column represents a particular trait (binary input). The sparsity rate of data is 0.0018, indicating that the data is extremely sparse.

Our dataset has properties which are similar to text datasets:
- Appealing properties of NLP models:
  - Remarkable success in practice
  - Heavy ongoing research field
  - Lend themselves to top-N recommendations

Term Frequency-Inverse Document Frequency
- Traits as words, users as documents

\[ tf-idf(d, t) = tf(t) \times idf(d, t) \]

\[ idf(d, t) = \log \frac{n}{df(d, t)} \]

Item2Vec

\[
\text{minimize } J = - \log P(w_c | w_{c-1}, \ldots, w_{c-8}, w_{c+1}, \ldots, w_{c+m}) \\
= - \log P(u_c | \theta) \\
= - \log \frac{\exp(u_c^T \phi)}{\sum_{j=1}^{|V|} \exp(u_j^T \phi)} \\
= -u_c^T \phi + \log \sum_{j=1}^{|V|} \exp(u_j^T \phi)
\]

Where \( u_c \) is the embedding of the output vector \( w_c \), and \( \theta \) is the average of the embeddings of the 8ws context words

Item2Vec Algorithm
1. Build dataset containing 5 random shuffles of each row.
2. Run word2Vec on this dataset
   - vector size: 100
   - min count: 1
   - window: 8
   - negative sampling: 16
   - iterations: 20

Conclusion & Future Work
Both the models described allow the traits to be treated as a set, in NLP terms a bag of words. Hence, The results of these models, particularly the Item2Vec model, suggests that NLP can be applied with success to non-textual datasets. There has been some work on parallelization of word2vec models[2], in future we will look more closely into such models and architectures.

Acknowledgments
We would like to acknowledge our faculty mentor Eleni Drinea and industry mentors Charles Menguy and Handong Zhao for their support, patience, and guidance.

Reference(s)