Assigning Startups to Clusters Based on Customer-Value Proposition

Meeran Ismail and Daniel Semeniuta, [meeran, dsemeniu]@stanford.edu

PREDICTING

Countless new startups are born every single day, and venture capitalists are always on the lookout to find which one will be the next big thing. One piece of information that is especially valuable to investors is the industry that a startup is in and the industry competition it faces. We thus want to use machine learning to cluster companies by customer value proposition, given nothing more than short one to two lines describing what the company does.

DATA

Our data comes from a CSV containing company descriptions from Pitchbook and Crunchbase

Website Domain	Input Text Description	Output Industry/Labeling
0-in.com	Operator of an assertion-based verification company. The company develops and supports electronic design automation tools and functional verification products that help clients to verify multi-million gate application-specific integrated circuit and system-on-chip designs. Its system also automates the engineered methodologies.	Automation/Workflow Software (1)
011now.com	011Now provides international phone communications at a lower cost than typical calling cards or standard international rates.	Telecom (4)
1-2-3.tv	1-2-3.tv is a multichannel auction house with a combination of exciting auction action and service-oriented multi-channel homeshopping.	Broadcasting, Radio and Television (3)

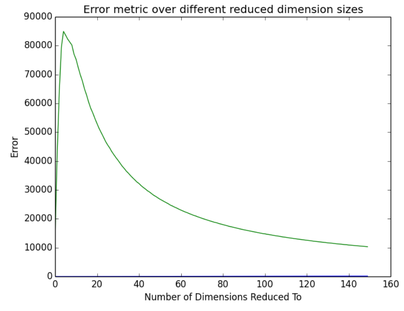
FEATURES

- Raw input results in a feature set of 90,000 words
- After dimensionality reduction through SVD, we bring it down to 20-30 features

MODELS

- K-means clustering for grouping descriptions
  - Multinomial vs Indicator
  - Euclidian vs Manhattan (other metric??)
- Singular-value decomposition for dimensionality reduction
- Other Clustering Models
  - Gaussian Mixture Models? DBSCAN?

RESULTS



ERROR METRIC

Let $X \in \mathcal{R}^{m \times n}$ be our document-term matrix (m documents, n terms), U, Σ, V be our SVD decomposition, K be the number of clusters we have and let $\mu_k \forall k \in [1, K]$ be the centroid location of cluster k. Then we define the following error metric:

$$Error = RSS_{X,K} \times \frac{Var(X_{transormed})}{Var(X)}$$

$$RSS_{X,K} = \sum_{i=0}^{m} \min_{k \in [1,K]} (x_{i*}V - \mu_k)$$

$$Var(M \in \mathcal{R}^{m \times n}) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{i=1}^{n} (x_{ij} - \bar{x}_M)^2$$

$$\bar{x}_{M \in \mathcal{R}^{m \times n}} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{i=1}^{n} X_{ij}$$

DISCUSSION

- Clustering text using features to gauge sentiment is hard
- Euclidean distance is not the best measure for vectors of word features
- Distillation of text features to the most essential words is best
- K-means may not be the best clustering algorithm for text. Look to GMM, EM, etc.

FUTURE

- Better clustering through more meaningful feature reduction
- Better weighting of words (e.g. tf-idf)
- Scrape web for some existing categorization of startups to provide a ground-truth

REFERENCES

1. A. Huang, “Similarity measures for text document clustering,” in *Proceedings of the New Zealand Computer Science Research Student Conference*, Apr 2008, pp. 49–56.