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.

<table>
<thead>
<tr>
<th>Website</th>
<th>Domain</th>
<th>Input Text Description</th>
<th>Output Industry/Labeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1-3rv</td>
<td>01Now.com</td>
<td>01Now provides international phone communications at a lower cost than typical calling cards or standard international rates.</td>
<td>Telecom (4)</td>
</tr>
<tr>
<td>1-2-3rv</td>
<td>1-2-3rv.com</td>
<td>1-2-3rv is a multi-channel auction house with a combination of exciting auction action and service-oriented multi-channel hotselling.</td>
<td>Broadcasting, Radio and Television (3)</td>
</tr>
</tbody>
</table>

**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

**ERROR METRIC**

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

$$
\text{Error} = \text{RSS}_{k,X} = \frac{\text{Var}(X_{\text{reconstructed}})}{\text{Var}(X)}
$$

$$
\text{RSS}_{k,X} = \sum_{i=1}^{m} \text{SSR}_{k,i} = \sum_{i=1}^{m} \text{Var}(u_{i}V - \mu_k)
$$

$$
\text{Var}(M \in \mathbb{R}^{m \times n}) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (M_{ij} - \bar{X}_{M})^2
$$

$$
\bar{X}_{M} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}
$$

**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**

**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**