Creating Labeled Clusters of Startups Based on Customer-Value Proposition

Finance & Commerce

Meeran Ismail (meeran@stanford.edu), Daniel Semeniuta (dsemeniu@stanford.edu)

Motivation

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. To do this, they must learn an incredible amount about the startups they want to invest in. One piece of information that is especially valuable to investors is the industry that a startup is in and the industry competition it faces. As such, classifying startups by industry function is an important tool in investing, as well as understanding the industry role of these different categories; however, doing this for the many thousands of startups that are formed every day is impossible by hand. 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.

Methods

In general, this project will make use of unsupervised learning techniques, some of which are well-established while others will be more a creation of our own. As a baseline, we'll use words as features to create an *n*-dimensional space (where *n* is the number of distinct words in the corpus) on which we'll apply K-means. However, we don't expect this approach to work that well because in extremely high-dimensional states, all points become close to equidistant from one another. This means that our other approaches will be reducing the dimensionality of this space we are in. Minimizing objective functions for UV factorization, topic modeling, and LDA (latent dirichet allocation) are all approaches we will consider and test out. The goal will ultimately be to reduce the dimensionality enough to the point where points start to separate from another into distinct regions.

After each startup is grouped into regions, we will then assign labels and keywords to these regions to make them more readable to the human reader. Examples of the input/output behavior follow:

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
011now.com	011Now provides international phone communications at a lower cost than typical calling cards or standard international rates.	Telecom

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

Our system would take in these text descriptions, put them into intermediary groupings, and then label these intermediary groupings during the learning phase. The user side application would be inputting a text description for a potential idea, and outputting a labeling alongside similar companies existing in our dataset.

Evaluation

These methods will be evaluated through normal means of evaluating clustering algorithms. First, to evaluate the dimensionality reduction, we'll use mean squared error (MSE) to see how well the decomposition of our original corpus worked out. Then, to evaluate the success of our clustering algorithm, we'll use standard reconstruction loss (with respect to the dimensions that we've reduced our corpus down to).

Both team members are enrolled in both CS221 and CS229. Our projects for both classes will be utilizing the same base idea and dataset, with different applications.