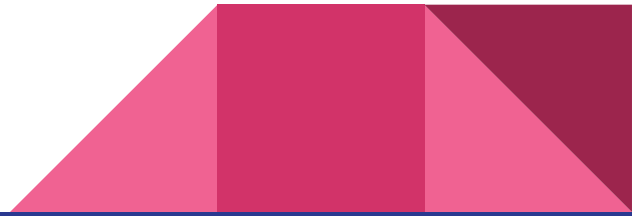
Using NLP to gather information from college reports

Sarah Robinson

The Data

71 Annual Reports submitted by community colleges

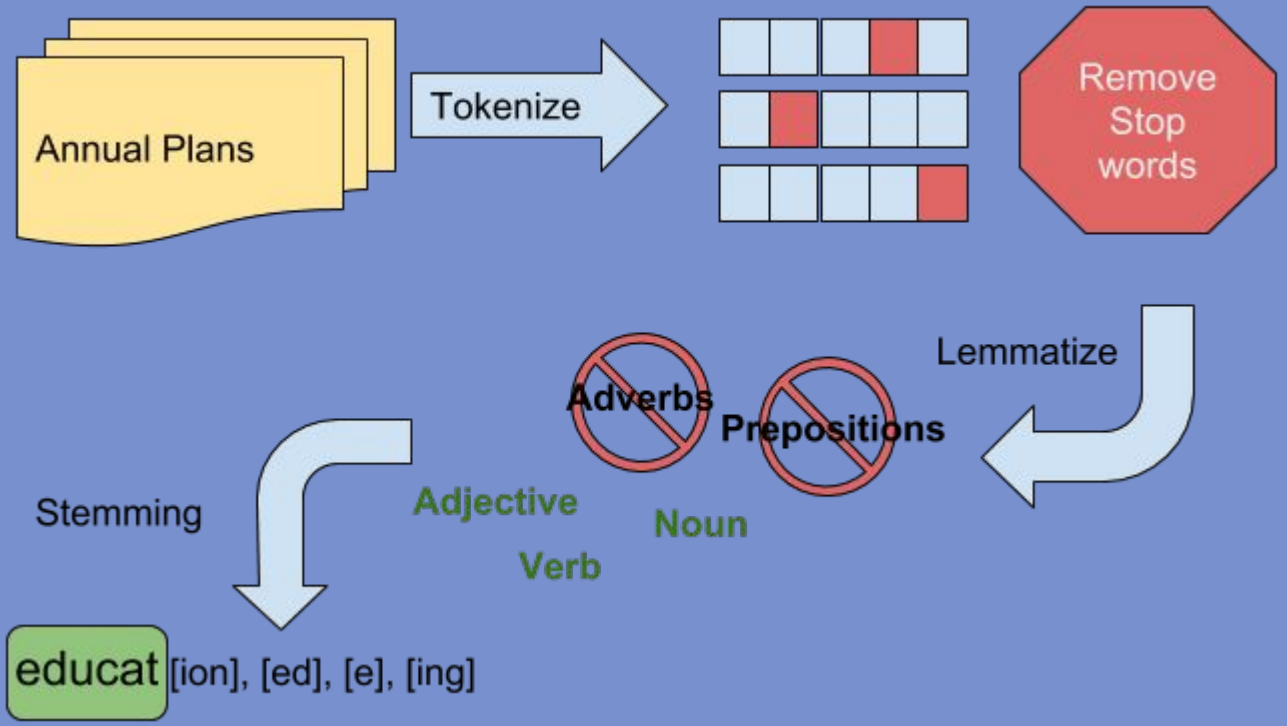
Similar Documents to use for topic modeling



The Goal

To use NLP to identify main themes and trends within the reports and to identify institutions with similar goals.

Preprocessing



Methods

1. **TFIDF and KMeans Clustering:**
   Grouping based on relative phrase frequency
2. **Doc2Vec and KMeans Cluster:**
   Grouping based on how similar words seem using a pre-trained model
3. **LDA Topic Modeling:**
   Identifying common topics based on how similar words seem using a pre-trained model

1.TFIDF and KMeans Clustering

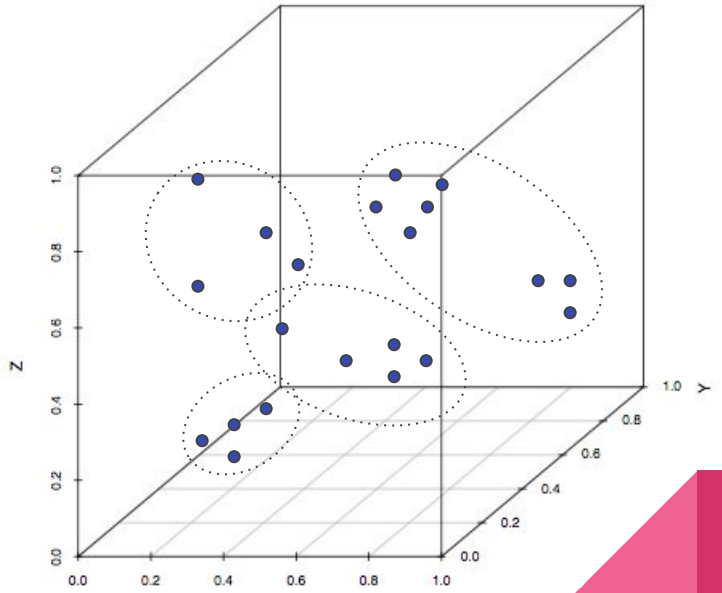
Term **F**requency **I**nverse **D**ocument **F**requency

$$\text{TFIDF} = \text{TF}\left(\frac{\text{\# of times term appears in document}}{\text{Total \# of terms in document}}\right) * \text{IDF}\left(log\_e\left(\frac{\text{Total \# of documents}}{\text{\# of documents containing term}}\right)\right)$$

1. TFIDF and KMeans Clustering

KMeans Clustering

- Documents given a coordinate in multi-dimensional space
- Grouped to minimize distance from cluster center and maximize distance from one center to another



Top words from each cluster

Cluster 0 words:
'classes', 'sites', **'adults'**, 'team', 'well', 'instruction', **'building'**, 'activities', 'last', *'disabilities',* 'help', 'exploration', **'noncredit'**, 'progress', 'market', 'project', 'original', 'basic', 'groups', 'consist',

Cluster 1 words:
'aep', 'three-year', 'fiscally', 'board', 'ensure', **'engage',** 'program', **'county',** 'across', 'approved', **'academic'**, 'gaps', 'managed', 'comprehensive', **'noncredit'**, **'adults'**, 'make', 'approaches', 'ongoing',

Cluster 2 words:
'county', 'training', 'classes', 'skills', '3-year', 'learners', 'high', 'certificates', 'well', 'local', 'reviewed', 'within', *'diploma',* 'adults', 'agencies', 'instruction', 'serves', 'market', 'current', **'learning'**

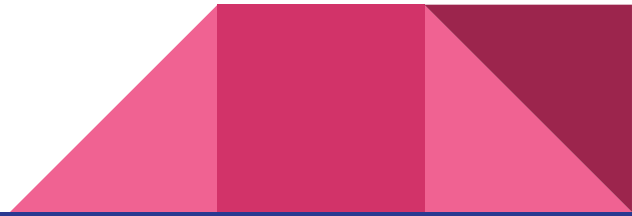
Cluster 3 words:
'saec', 'objectives', 'committee', **'academic',** 'access', 'strategic', **'building',** 'retreat', 'program', 'staff', 'counseling', 'support', 'seamless', 'current', 'within', 'jo', 'first', 'board', 'county', 'assist',

Cluster 4 words:
'3-year', 'created', 'initiated', 'three-year', 'progress', 'made', 'leveraging', 'staff', 'stakeholders', 'outcome', 'address', **'local',** **'engage',** 'effectiveness', 'reviewed', 'activities', 'ensure', 'san', 'approved', **'learning'**

2. Doc2Vec and KMeans

Gensim

- Pre-trained model
- already has words mapped in space
- Can identify words similar in meaning
- Doc2Vec maps documents in space (vectors) using this pretrained network



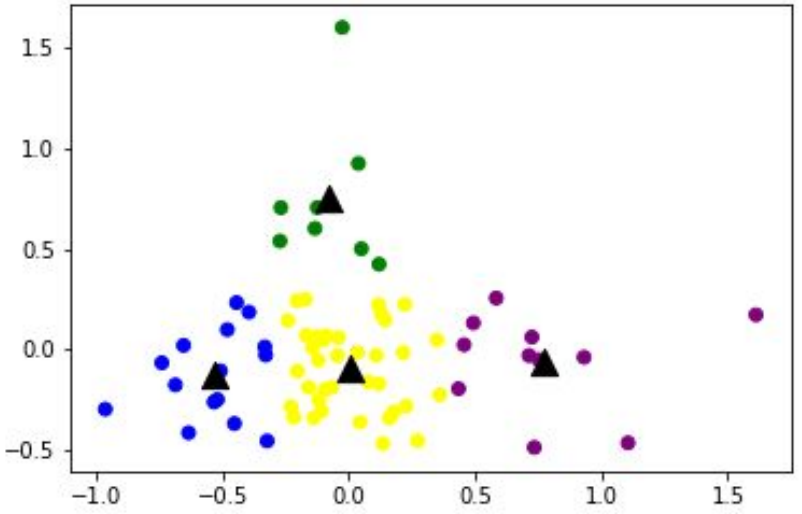
2. Doc2Vec and KMeans

Cluster 0 words: 'fiscally', 'focus', **'gaps'**, **'address'**, **'curriculum'**, **'allocate'**, **'goals'**, 'high', **'partnerships'**, 'esl', 'program', 'learners', **'basic'**, 'data', 'workforce', 'education', 'initiated', 'skills', 'community'

Cluster 1 words: 'fiscally', **'goals'**, **'address'**, **'gaps'**, **'allocate'**, **'curriculum'**, **'partnerships'**, **'basic'**, 'esl', 'focus', 'workforce', 'high', 'program', 'data', 'community', 'approved', 'education', 'leveraging', 'skills'

Cluster 2 words: 'fiscally', **'goals'**, **'allocate'**, **'address'**, **'gaps'**, **'basic'**, **'curriculum'**, **'partnerships'**, 'workforce', 'approved', 'community', 'esl', 'data', **'leveraging'**, 'board', 'college', 'education', 'program', 'areas'

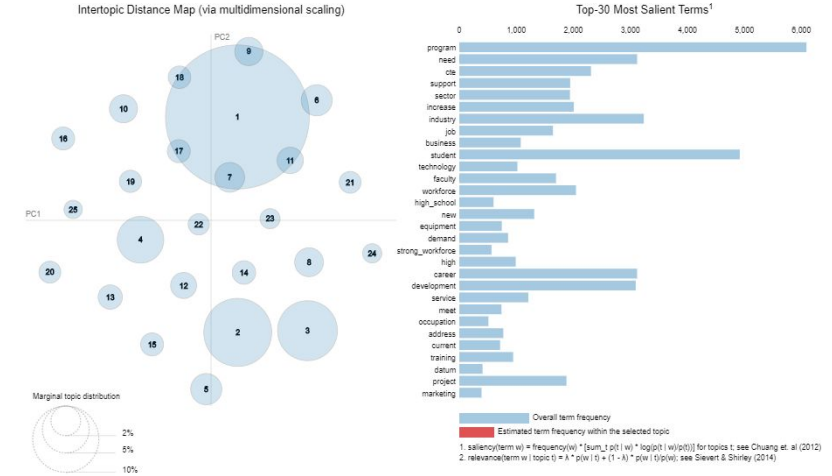
Cluster 3 words: 'fiscally', b"'s", **'gaps'**, **'address'**, **'goals'**, **'allocate'**, **'curriculum'**, **'partnerships'**, **'basic'**, 'focus', *'esl'*'esl', 'workforce', 'high', 'data', 'program', 'education', 'community', 'approved', 'learners', 'areas'



3. LDA Topic Modelling

- Similarly, uses gensim group topics, instead of documents

Click to see LDA Visualization



Final Thoughts

- Gained a high level overview of plan topics
- Clustering attempts identified some of the same top words for each cluster
  - Try removing these words as stop words
  - Set parameter to only include less common words
  - Try other clustering methods