For a paper of interest, looking for more papers with similar topic would be time consuming and effort-taking. We propose a new method of exploring papers through Machine Learning methods that can save people’s effort.

**Motivation**

Existing Paper → Multiple Possible Topics → Multiple Papers for each topic → Result Papers

Paper selection by human – Time and effort consuming

Existing Paper → Machine Learning → Machine Extracted Topics → Paper & Topics in Data → Result Papers

Paper selection by machine – Higher efficiency & accuracy

**Data & Preprocessing**

Data source: 2011 ~ 2015 CS229 course project reports

<table>
<thead>
<tr>
<th>Number of documents</th>
<th>Total words</th>
<th>Unique words</th>
<th>Dictionary size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,298</td>
<td>2.4 million</td>
<td>6,522</td>
<td>39,588</td>
</tr>
</tbody>
</table>

For each document:
- Convert format from PDF to txt
- Remove non-English words (i.e. numbers, symbols, signs etc.)
- All words to lowercase
- Word suffix removal (word stemming) (obtained dictionary)
- Trivial words removal (i.e. the, and, was, we)

Vectorize the selected feature into w columns and the documents into d rows. And thus we have the **text feature matrix X**.

\[ X_{i,j} = \sum_{w \epsilon d_i} \{ \text{word} = w_j \} \]

\( X_i \) is the time of appearance of a specific word \( w_j \) in document \( i \). Remove non-common words like names or rare jargons, we filter out words with times of **time of appearance < 3** [obtained unique words].

An example of a paragraph before and after processing will be:

Welcome to CS 229! This course provides a broad introduction to machine learning and statistical pattern recognition. Topics include: supervised learning (generative / discriminative learning...)

welcome course provide broad introduce machine learn stat pattern recogni topic include supervis learn generat discriminit learn

**Methodology**

Training

- Setting cluster number to 20
- **K-means:**
  - Obtain topic assignment on each document
- **Latent Dirichlet Allocation (LDA):**
  - Obtain topic distribution on each word

**Analyzing**

- **Matrix Y:** Labeling Matrix for documents
- **Matrix Z:** Labeling Matrix for words
- Convert Matrix X / Z to matrix of topic distribution over documents in k-means / LDA method.

**Testing**

- Add a row vector representing the word composition over the test document and plug into k-means / LDA method.

**Results**

Clusters for k-means & unique words

**Comparison**

- Documents recommended by k-means method have a very similar distribution with the reading list papers compound distribution
- Distribution of documents recommended by LDA deviate more. This may indicate more variance error.

**Methodology**

- Setting cluster number to 20
- **K-means:**
  - Obtain topic assignment on each document
- **Latent Dirichlet Allocation (LDA):**
  - Obtain topic distribution on each word

**Analyzing**

- **Matrix Y:** Labeling Matrix for documents
- **Matrix Z:** Labeling Matrix for words
- Convert Matrix X / Z to matrix of topic distribution over documents in k-means / LDA method.

**Testing**

- Add a row vector representing the word composition over the test document and plug into k-means / LDA method.

**Results**

Clusters for k-means & unique words

**Comparison**

- Documents recommended by k-means method have a very similar distribution with the reading list papers compound distribution
- Distribution of documents recommended by LDA deviate more. This may indicate more variance error.

**Methodology**

- Setting cluster number to 20
- **K-means:**
  - Obtain topic assignment on each document
- **Latent Dirichlet Allocation (LDA):**
  - Obtain topic distribution on each word

**Analyzing**

- **Matrix Y:** Labeling Matrix for documents
- **Matrix Z:** Labeling Matrix for words
- Convert Matrix X / Z to matrix of topic distribution over documents in k-means / LDA method.

**Testing**

- Add a row vector representing the word composition over the test document and plug into k-means / LDA method.

**Results**

Clusters for k-means & unique words

**Comparison**

- Documents recommended by k-means method have a very similar distribution with the reading list papers compound distribution
- Distribution of documents recommended by LDA deviate more. This may indicate more variance error.