The goal of this project is to predict the impact of a research paper/project using keywords related to the paper’s title, which could help guide researchers a feel for the level of interest in a proposed project. An effective measure of research impact is the number of citations received over the years, so our proposed algorithms attempt to predict the number of citations per year a paper will get in the area of nanotechnology, based on its title keywords.

Dataset / Data Processing
Top papers on Google Scholar from 3 journals:
- Nanoletters, ACS Nano, Nature Nanotechnology
- 523 papers total containing titles, publication years, and citation counts

Input raw data:
Title: (Fracture of Silicon Nanoparticles During Lithiation) -> x
Citations: # of citations / Years passed since it appeared -> y

- Remove non-alphanumeric characters, replace with spaces.
- Make all characters lowercase, split on whitespaces to form a vector of keywords.
- ’fracture’ of ‘silicon’ nanoparticles ‘during’ ‘lithiation’

- Remove prepositions/small unimportant words
  ’fracture’ ‘silicon’ nanoparticles ‘during’ ‘lithiation’

- Find root substring of each word by comparing to the vocabulary.
  If it is a new word, it is added to the vocabulary.
  ’fracture’ ‘silicon’ particle ‘during’ ‘lithiation’

Theory / Algorithms

### Gaussian Discriminant Analysis
Divide the dataset into $b$ bins based on number of citations/year. Calculate $p(x = b|y)$ and $p(y)$ from the dataset. Then make prediction using max probability given by:

$$p(y = b|x) = \frac{p(x|y = b)p(y = b)}{\sum_{j=1}^{y} p(x|y = j)p(y = j)}$$

### Softmax Regression
Here we do a linear regression to fit the entire space of vocabulary words to predict a given title.

$$h_{\theta_i}(x^{(i)}) = \phi_1$$
$$p(y = i|x^{(i)}; \theta) = \frac{e^{\theta_i x^{(i)}}}{\sum_{k=1}^{y} e^{\theta_k x^{(i)}}} = \phi_i$$
$$\bar{\theta}_i := \bar{\theta}_i + \alpha\gamma \left(1 - h_{\theta_i}(x^{(i)})\right)x^{(i)}$$

### K-means Clustering
Cluster the titles based on similarity, i.e. number of common words.

Algorithm:
1. Sort titles into clusters based on similarity to the cluster keywords
2. Redefine the each cluster’s keywords to be the most common words in the cluster.

Results

### Classification Accuracy
![Classification accuracy graph]

#### Mean Error
Error metric:
$$E = \frac{\|C_{true} = c_{y=b}\|}{m}$$
$c_{y=b}$ is the center of the predicted bin.

### K-means Clusters
- High Performance
  - Solar cells
  - Transistor technology
  - Battery technology
  - Drugs/Materials
- Medical
  - Plasma
  - Medical devices
  - Therapeutics
  - Nanomaterials

### Data Distribution
- We observe GDA to have lower test error than softmax, but each offer only a modest improvement over random guessing.

- Titles words alone are insufficient information for a meaningful prediction. In future should consider abstract, authors, references, and even body text.

- But clustering can successfully group titles into current relevant research categories.