Clustering phrases extracted from raw text in such a way that semantically similar phrases are grouped together.

**Word2Vec Skip Gram with custom weight Initialization**

- Create a copy of the data with phrases in each sentences replaced by a single entity
  - ‘I love New York’ → ‘I love New-York’
- Equate corresponding sentences in each dataset
  - Average embedding(“I love New York”) = Average embedding(“I love New-York”)
  - Extract phrase embedding by using word2vec embeddings for all words and treating the phrase as an unknown
- Train a skip-gram model on the hyphenated dataset, with word embeddings initialized with Google Skip Gram embeddings and phrase embeddings initialized with the extracted values

**Sample Resultant Clusters**

- **Occasions**
  - Nice Wedding
  - Rememberance Day
  - Birthday Party
  - Mothers Day
  - Great Christmas
- **Locations**
  - Medical District
  - Drum Tower
  - Bomb Shelter
  - National Forest
  - Pall Mall
- **Food**
  - Dried Bread
  - Strawberry jam
  - Honey Sauce
  - Lemon Juice
  - Ground Beef
- **Transportation**
  - Black Taxis
  - Land Cruiser
  - Limo Ride
  - Bus Coach
  - Renting Bikes

**Future Work**

- Translate the idea of equating sentences into a deep learning model
- Prepare an evaluation criterion for phrase clusters