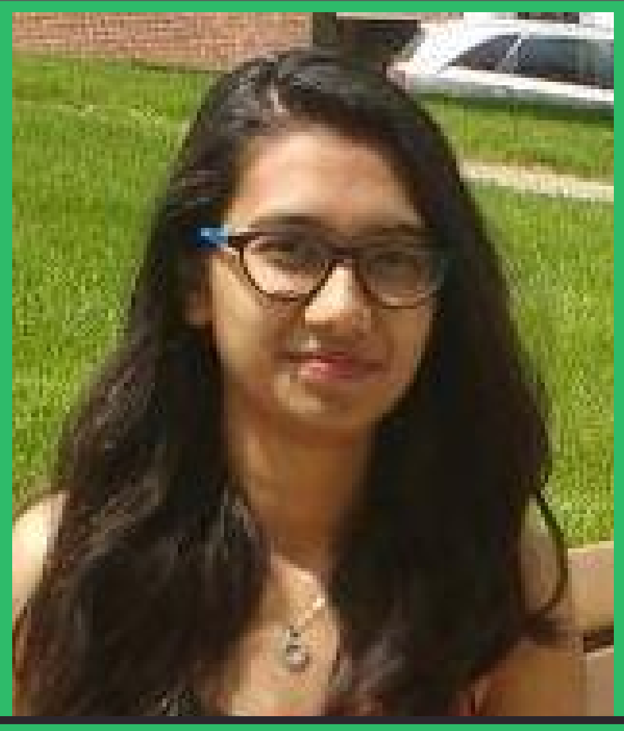


Tanvi Sahay

MS, CS, UMass Amherst
June '17 RnD Intern, Lexlaytics

Machine Learning
Natural Language Processing
Deep Learning

Graduating May 2018
Seeking Full Time Opportunities

Ankita Mehta

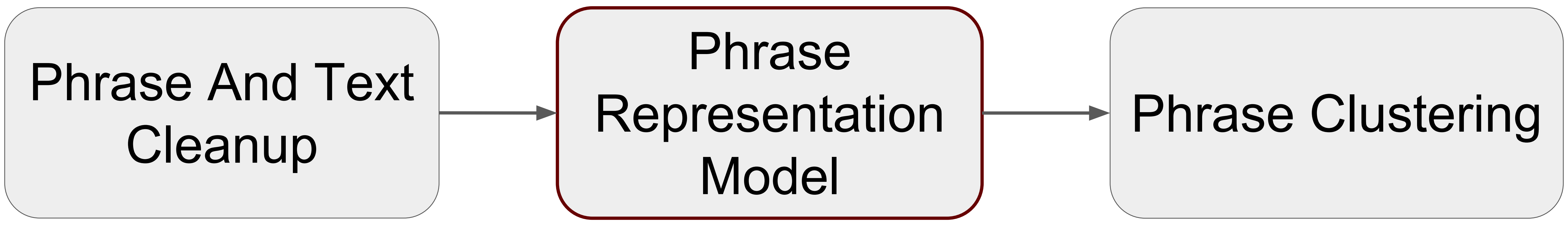
MS, CS, UMass Amherst
June '17 RnD Intern, Lexlaytics

Machine Learning
Natural Language Processing
Deep Learning

Graduating May 2018
Seeking Full Time Opportunities

Concept/ Theme Rollup

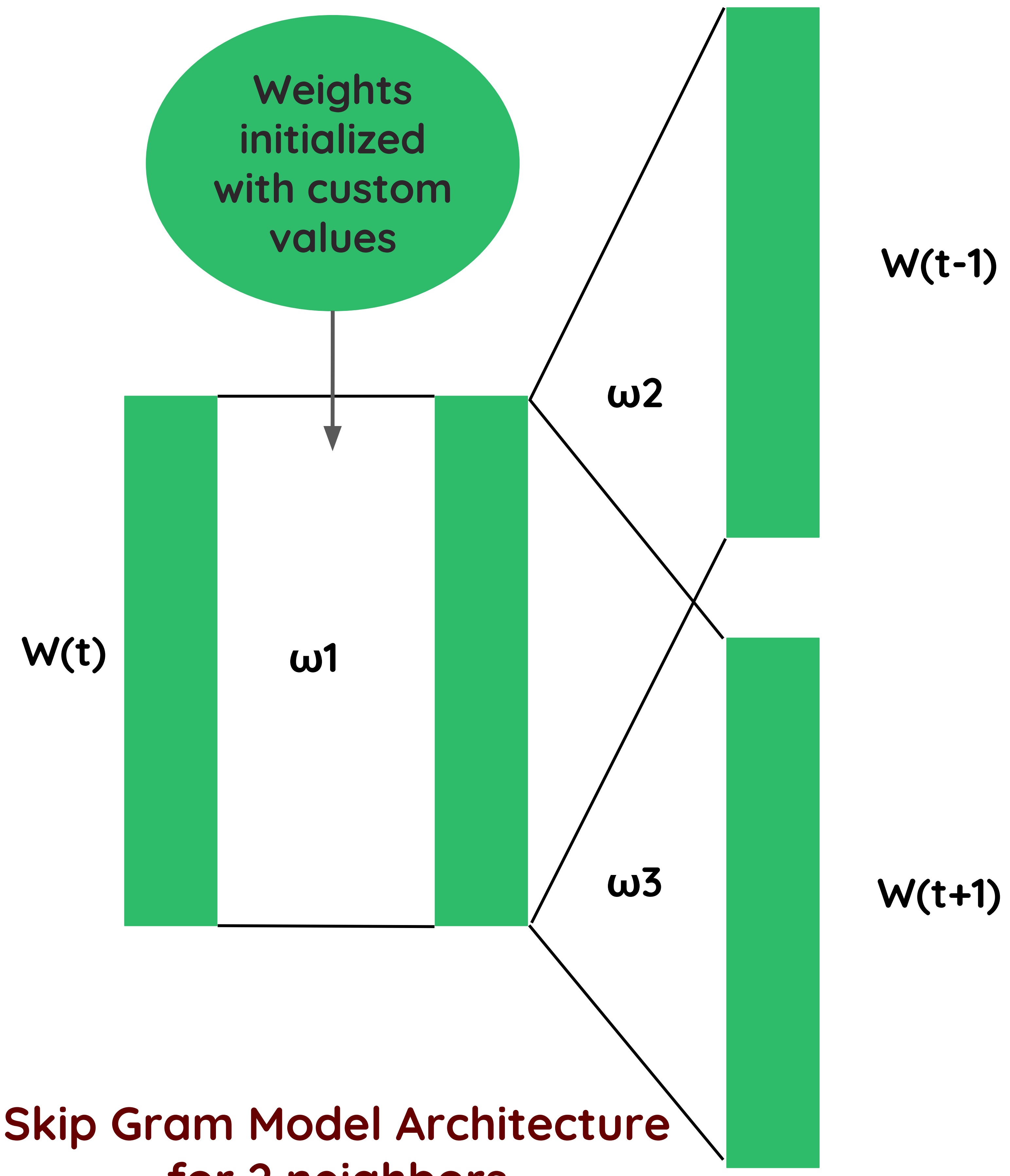
Clustering phrases extracted from raw text in such a way that semantically similar phrases are grouped together



Phrase And Text Cleanup → Phrase Representation Model → Phrase Clustering

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



Skip Gram Model Architecture for 2 neighbors

Sample Resultant Clusters

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

Future Work

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