Assignment2

Name:Arushi Dogra
Roll No : 201302084

We had to implement k-means clustering and make the co-occurence matrix. There were 3 cases to be implemented:

1. Feature words not containing stop words.
2. Feature words containing stop words.
3. Feature words not contaning top-50 unigrams.

Procedure:
• The data was first tokenized.
• Then the co-occurence matrix was built using top-250 words as feature words.
• Words were clustered using k-means algorithm and 50 centroids.
• 25 words from each group are taken and displayed.

Observations:
Most of the similar words occured in the same group. Like tokens representing quantity appeared in same cluster.

Difficulties:
• Iterations took too much time as the data was large.
• Some groups didn't contain any words.
• Tokenizer was ours so it wasn't perfec