Lab assignment 2
CIS 660/EEC 525 Data Mining
Name: Adil Hashmi Syed
I'd:2824849

Part 1:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.
Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.
Data preprocessing is a proven method of resolving such issues.
The document vector that is the result of the process in the step is a structured table consisting of row one for every blog entry in the training set and attributes or columns each token within an article that meets the filtering and stemming criteria defined by operators inside Process Documents.

Part 2:
Data transformation is the mapping and conversion of data from one format to another. For example, XML data can be transformed from XML data valid to one XML Schema to another XML document valid to a different XML Schema. Other examples include the data transformation from non-XML data to XML data.

Part 3:
1)Cosine similarity is a metric used to determine how similar two entities are irrespective of their size.It measures the cosine of the angle between two vectors projected in a multi-dimensional space. If 'a' and 'b' are two vectors, cosine equation gives the angle between the two.

2)The result from computing the similarity of Item A to Item B is the same as computing the similarity of Item B to Item A.

3)In terms of 7 given topics 2 and 3 are most similar