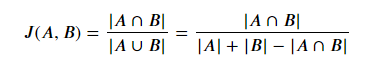
**Social Media Analytics Lab**

**Instructor: Dr. Janani Selvaraj**

**Lab 10. Clustering the job titles of LinkedIn Connections using Greedy Heuristic Algorithm**

1. Retrieve the Connections data of your LinkedIn Profile from the SMA Lab 8
2. Transform the titles by normalising it to known forms
3. Understand the concept of Jaccard similarity by calculating Jaccard distance between titles

The Jaccard distance is based on the Jaccard Index, which is used for comparing two sets of strings:



1. Find Jaccard distance between two titles

Sample code:

from nltk.util import bigrams

ceo\_bigrams = list(bigrams("Chief Executive Officer".split(), pad\_left=True, pad\_right=True))

cto\_bigrams = list(bigrams("Chief Technology Officer".split(), pad\_left=True, pad\_right=True))

print(ceo\_bigrams)

print(cto\_bigrams)

print(len(set(ceo\_bigrams).intersection(set(cto\_bigrams))))

1. Calculate the distance and similarity between 2 titles

from nltk.metrics.distance import jaccard\_distance # pip install nltk

job\_title\_1 = 'Chief Executive Officer'.split()

job\_title\_2 = 'Chief Technology Officer'.split()

print(job\_title\_1)

print(job\_title\_2)

print()

print('Intersection:')

intersection = set(job\_title\_1).intersection(set(job\_title\_2))

print(intersection)

print()

print('Union:')

union = set(job\_title\_1).union(set(job\_title\_2))

print(union)

print()

print('Similarity:', len(intersection) / len(union))

print('Distance:', jaccard\_distance(set(job\_title\_1), set(job\_title\_2)))

1. Apply the above distance and similarity measures to the job titles and find clusters