**CS5560 Knowledge Discovery Management**

**In Class Assignment #5**

**20th February 2020**

**Feedback Link :** [**ICP\_FEED\_BACK\_LINK**](https://docs.google.com/forms/d/1dEA7tI8CaV-MZvdDNvvZbc_wZbyyhcNJe5JEGGQVidA/edit?ts=5e2144a4)

1. **Generate the output (changes or transformations in the data) manually when the following tasks are applied on the input text. Show your output in details.**

***Input***:

*5 Abstracts saved in separate files*

***Tasks:***

* 1. ***Find out the top TF-IDF words for the above input.***
  2. ***Find out the top TF-IDF words for the lemmatized input***
  3. ***Find out the top TF-IDF words for the n-gram based input.***

1. **Write a simple spark program to read a dataset and find the W2V Synonyms for the Top TF-IDF Words**

* 1. Try without NLP
  2. Try with Lemmatization
  3. Try with NGrams

Compare the results from (a) , (b) and (c)