Lab9-Natural Language Processing – Finding Text Similarities

Total Marks: **8 Marks + 2 marks (individual assessment)= 10 Marks**

This assignment is about finding the similarities in given sentences.

Id Text

1. In the past John liked only sport but now he likes sport and politics
2. Sam only liked politics but now he is fan of both music and politics
3. Sara likes both books and politics but in the past she only read books
4. Robert loved both books and nature but now he only reads books
5. Linda liked books and sport but she only likes sport now
6. Alison used to loved nature but currently she likes both nature and sport

Using Python language, perform the followings NLP tasks to find the similarities between the given sentences:

1. Using NLTK word_tokenize function, tokenize the given sentences
2. Using NLTK PorterStemmer, perform the stemming for the tokens of the sentences
3. Using NLTK WordNetLemmatizer, perform the lemmatization for the stemmed tokens
4. Using sklearn K-means clustering technique, cluster the given sentences. Find the feature vectors for the input of a K-means algorithm using the below techniques. Also, find an appropriate K-value using a KneeLocator method from the python kneed library.
    a. TF-IDF
    b. TF
    c. BOW
    d. Word2Vec
5. Visualize the clusters using the word clouds.

Tips:

- You will need to adjust the vector size in order to find the appropriate k-value