WEB SCIENCE COURSE WORK 4 Matriculation Number - 2576183S

1. **First you should use topic modelling on tweet's text data. You should analyse the results and identify issues with short text topic modelling. You can use any topic modelling library**

Issues with short text topic modelling:

Because short texts are so common on the Internet, mining their underlying subject patterns has become a key and basic challenge for many applications. Short text material is sparse, and there is no "word co-occurrence information." Traditional topic models that infer cohesive subjects from short texts are hampered by this. The issue of extracting latent ideas from short texts remains difficult. The problem of sparsity exacerbates performance concerns. Because of the restricted contexts, topic models have a harder time determining the meanings of ambiguous terms in short papers. Document-level co-occurrence patterns are challenging to capture in short texts.

2. **Group tweets based on some criteria. The idea here is to group similar tweets content wise and/or from the same users. Develop topic models on them and analyse the performance.**

   Code in ipynb file

3. **Compare the performance differences and discuss the reasons. Comparingperformances you should use more than metric we discussed in slide.**

	Perplexity	Coherence
Model with 5 topics	-8.620935265730441	0.20457114492991196
Model with 20 topics	-8.841893195320267	0.32670822598172916
Model with 50+ topics	-9.234749827641524	0.3994241455622251

1. Perplexity is a popular metric for assessing a topic model's generalisation performance The value of perplexity is equal to the inverse of the geometric mean per-word likelihood When the perplexity is lower, the model's generalisation ability is better; when it is higher, the model's generalisation ability is poorer. The greater the coherence score, the more probable it is that the topic's top T related words will appear in the same document. Topic coherence assesses the degree of semantic similarity between high-scoring terms in a topic's score. These metrics aid in the distinction between semantically interpretable issues and statistical inference artefacts. If a set of statements or facts supports each other, it is said to be coherent. As a result, a coherent set of facts can be interpreted in a context that encompasses all or nearly all the facts. Coming to our model. When the number of topics were low the perplexity and coherence scores were also low. When we increase the number of topic to 20 still we were not able to see much difference. When we increase the topic to around 68 we were able to see good. perplexity and coherence score Hence more topic with better parameters will give you good score and which means our model is better