COSC 2671 Social Media and Network Analytics

Tute Week 2

Introduction to Topic Analysis & Machine Learning Revision

Learning outcomes:

- Reinforce topic analysis concepts from lectures
- Revise using machine learning to solve problems

Introduction

1. Consider the following text. Construct the unigram Bag of Words representation for the words in bold (there are other unbolded words that are of interest, but to avoid using too much time for this, just focus on the bolded ones). Assume we also converted all words to lower case before constructing this bag of words representation. You might consider using a table to answer this question. Can you see potential issues with using unigrams, if we changed to bigrams, would that fix these issues?

   Data science is a **multi disciplinary** field that uses scientific methods, processes, **algorithms** and **systems** to **extract knowledge** and insights from structured and unstructured **data**. **Data science** is the same concept as **data mining** and **big data**: "use the **most powerful hardware**, the **most powerful programming systems**, and the **most efficient algorithms** to solve problems".

   (Wikipedia article on Data Science, retrieved 28/07/2019)

2. Construct the TF-IDF weightings for the following set of documents and their unigram frequencies. Discuss the difference in weightings compared to raw frequencies.

	football	**basketball**	**cricket**	**Badminton**
Doc1	1	1	0	1
Doc2	2	0	2	1
Doc3	2	0	2	1

3. In lectures we covered several (NLP) based approaches that are useful for pre-processing, including tokenisation, stemming/lemmatisation and part of speech tagging. Outline the general process when given some text and want to do some analysis such as sentiment analysis and what are some techniques that can be used for each step of the process.

4. For supervised learning, explain what the differences between Decision trees and K-nearest neighbour are. In your answer you might want to first explain what they are doing first.

5. How does K-means and hierarchical/agglomerative clustering work and how are they different?

6. Given we are analysing social media and networks, discuss what are some ethnical and privacy concerns and what are potential solutions to these? Note there is no right or wrong answer for this question.