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)

Answer:

Unigram:

Word	**Frequency**	**Word**	**Frequency**
data	5	mining	1
science	2	big	1
multi	1	algorithms	2
disciplinary	1	most	3
systems	2	powerful	2
extract	1	hardware	1
knowledge	1	programming	1
efficient	1		

Words such as data science has been broken up when consider them as unigrams, but as bigrams they remain as a unit of analysis.

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

Answer:

	football	**basketball**	**cricket**	**Badminton**
Doc1	0	log (3)	0	0
Doc2	0	0	2 log(3/2)	0
Doc3	0	0	2 log(3/2)	0

Note the TF-IDF values for football and badminton columns. Because they occur over all documents, then they get reweighted to 0, as they are not discriminative words if we want to distinguish between the different documents.

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.

Answer:
General process is to first do pre-processing, then apply NLP then do the analysis.
For pre-processing, this would include things like reading it in, removing boilerplate, and typically use a combination of tools, heuristics and scripting to do so.
Then we apply NLP based approaches to enrich the text. This could include tokenisation, to break up the raw text into useful tokens. These tokens are then stemmed or lemmatised to their stems/lemmas, to reduce the large set of tokens to a smaller set of stems/lemmas, which convey the same meanings. We might also use part of speech tagging to enrich the stems/lemmas with POS tags, which are useful to understand contexts and semantics and helpful for subsequent analysis such as sentiment analysis.

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.

Answer:
Examine your Practical Data Science, Machine Learning or Data Mining notes.

Very briefly, to avoid repeating what was discussed in the readings, a decision tree looks at the data, finds which attribute and threshold, and data, such that (for information gain) the resulting nodes are as pure as possible, or the instances for that node are of one class if possible. In contrast, a k-NN classifier predicts a class of an instance based on the majority vote of its k nearest neighbours. Both make very different assumptions, decision tree assume that we can divide the data feature by feature, which isn’t always true. Conversely, k-NN doesn’t make that assumption. However, K-NN assumes all features are relevant, and is highly dependent on the similarity measure selected, if that is selected wrong or the dimensions are very high, k-NN may not work at all.

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

Answer:
Examine your Practical Data Science, Machine Learning or Data Mining notes.
K-means assumes all clusters can be described by a centroid and all instances belong to the cluster with the nearest centroid. Hierarchical clustering (agglomerative) assumes all instances are singleton clusters initially, and merges the two closest ones, based on a cluster similarity measure, e.g., single, average or complete merging.

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.

Answer:
No right or wrong answer here ☺ Hope the discussion was useful.