Text Analysis
Past exam questions on Preprocessing & Vector counts 2012-2015

May 2015: Question 1:

a) Explain the role of **filters** when preparing a document vector for clustering. In your answer, discuss three types of **term filters**. For each filter, include your own opinion of their usefulness when preparing document vectors for clustering.

(*8 marks*)

b) Explain one approach for identifying phrases in unstructured text.

(*6 marks*)

c) In the context of term counts in a document vector, explain **Term Frequency** (TF) and **Inverse Document Frequency** (IDF). What is the purpose of adjusting TF with IDF to give **TF-IDF**? You do not need to include formulas in your answer. Illustrate your answer with reference to Table 1.

(*11 marks*)

Table 1. Document vector counts

	theatre	**report**	**patient**	**hospital**
doc 1: Occurrences	1	2	1	3
doc 1: TF	0.14	0.29	0.14	0.43
Number of documents this term appears in	*3*	*15*	*10*	*2*
doc 1: TF-IDF	0.17	0.13	0.09	0.60

Total: 25 marks

May 2014: Question 1:

a) Are high frequency or low frequency terms likely to be useful when classifying document vectors? Explain your answer.

(*5 marks*)

b) Discuss one approach for identifying useful phrases in unstructured text.

(*6 marks*)

c) Table 1 gives the term counts for four terms that appear in document ‘doc 1’. There are a total of 40 documents in the dataset.

Explain each of the three term counts presented in Table 1, namely occurrences, normalised count, and TF-IDF.

Why does the term ‘student’ get a higher TF-IDF count than the term ‘event’ even though each term appears four times in the document ‘doc 1’?

Based on your own work mining text documents, which of the three counts would you recommend and why?

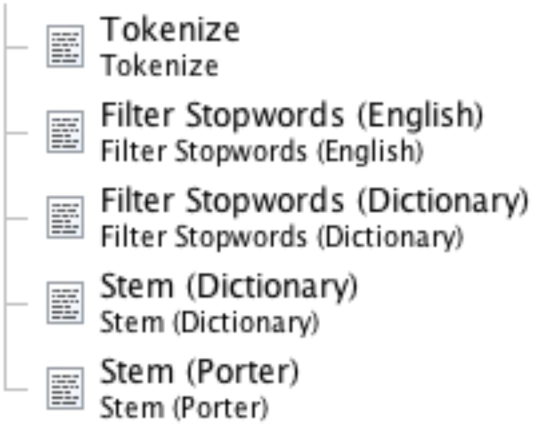
Table 1. Term counts

	course	event	news	student
doc 1 using **occurrences**	3	4	2	4
doc 1 using **normalised count**	0.23	0.31	0.15	0.31
doc 1 using **TF-IDF**	0.29	0.15	0.10	0.45
Number of documents this term appears in	*3*	*15*	*10*	*2*

(*14 marks*)

Total: 25 marks

August 2014: Question 1:



Tokenize
Tokenize

Filter Stopwords (English)
Filter Stopwords (English)

Filter Stopwords (Dictionary)
Filter Stopwords (Dictionary)

Stem (Dictionary)
Stem (Dictionary)

Stem (Porter)
Stem (Porter)

Figure 1. Text preprocessing

a) Figure 1 above lists the pre-processing steps that were used when training a k-NN classification model to label document vectors by topic. For each of the five pre-processing steps in Figure 1, explain if it also needs to be applied to unlabelled documents before classifying those documents using the same k-NN model.

(*12 marks)*

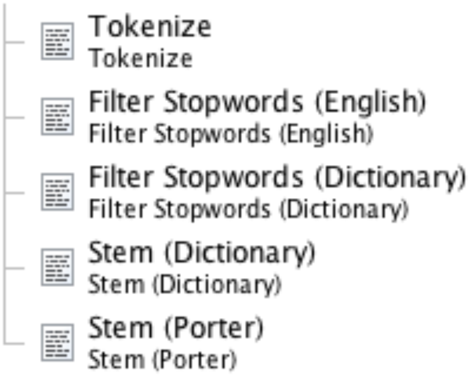
b) Explain the term ‘latent semantic indexing’. Your answer should explain what issue latent semantic indexing aims to address, and how it works.

(8 *marks*)

c) In your experience mining text, does adding an IDF weighting to a term count improve classification model accuracy? Explain your answer.
(*5 marks*)

Total: 25 Marks

May 2013: Question 1.



Tokenize
Tokenize
Filter Stopwords (English)
Filter Stopwords (English)
Filter Stopwords (Dictionary)
Filter Stopwords (Dictionary)
Stem (Dictionary)
Stem (Dictionary)
Stem (Porter)
Stem (Porter)

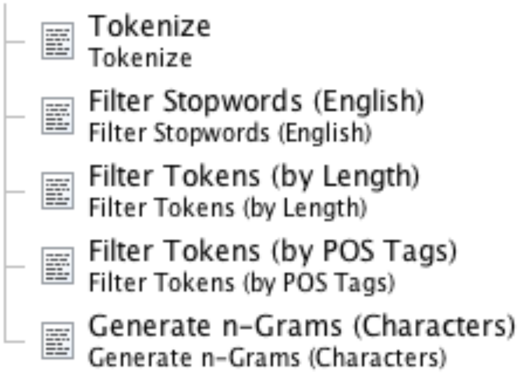
Figure 2. Text preprocessing

a) Figure 1 above lists a number of text pre-processing steps that could be used on a collection of documents. Explain the role of each step, and whether or not, in your experience, it is of benefit when creating document vectors for subsequent data mining.
(*15 marks*)

b) The five steps in Figure 1 were used as pre-processing steps before training a Support Vector Machine (SVM) model. You now need to apply that SVM model to unlabelled documents. Which of the pre-processing steps illustrated in Figure 1 must also be used on the unlabelled text documents? For each of the five steps, justify your answer.
(*10 marks*)

[Total: 25 marks]

Repeat 2013: Question 1.



Tokenize
Tokenize
Filter Stopwords (English)
Filter Stopwords (English)
Filter Tokens (by Length)
Filter Tokens (by Length)
Filter Tokens (by POS Tags)
Filter Tokens (by POS Tags)
Generate n-Grams (Characters)
Generate n-Grams (Characters)

a) Figure 1 above lists a number of text pre-processing steps that could be used on a collection of documents. Explain the role of each step, and whether or not, in your experience, it is of benefit when creating document vectors for subsequent data mining.

(*15 marks*)

b) When deciding on terms to include in a document vector, are high frequency or low frequency words more useful? Explain your answer.

(*6 marks*)

c) Explain the purpose of adding an IDF weighting to a term count.

(*4 marks*)

[Total: 25 marks]

May 2012: Question 3

a) Explain what is meant by a TF-IDF count. Your answer should cover both TF and IDF. Based on your experience, does TF-IDF give more accurate results that an occurrences count?
Note: You do not need to include formulas in your answer.

(11 marks)

b) The table below holds an extract from a Hidden Markov Model.

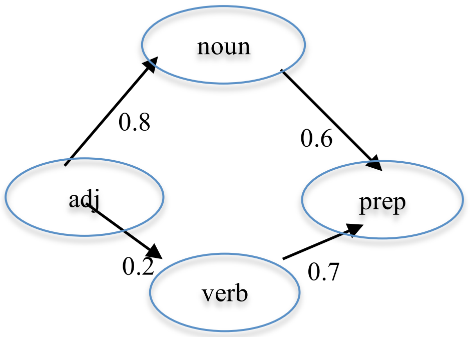
i) Explain the components of a Hidden Markov Model and their relevance to text mining.

(8 marks)

ii) Using the sample sentence and Hidden Markov Model given below, illustrate how to determine if 'walk' is a noun or a verb. You do not need to calculate the final answer, just indicate how this calculation would be done

(6 marks)

Sample sentence	**Model**	

He went for a long walk on the beech.
 	
Relevant Parts of Speech:
long = adj
walk = noun or verb
on = preposition (prep
		Prob(noun|walk) = 0.4
		Prob(verb|walk) = 0.6

[Total: 25 marks]

Text Analysis

Past exam questions on Preprocessing & Vector counts 2012-2015

Repeat 2012: Question 2.

a) There are a number of considerations when selecting terms for a document vector including canonical forms, identifying phrases and handling synonyms.

i. Explain the role of stemmers in improving on a simple bag of words. In your experience, do stemmers improve the accuracy of classifiers trained on a document vector?

(6 marks)

ii. Give an example of why a 2- or 3- word phrase can be more predictive than individual words. How can POS tagging be used in combination with an n-gram tokeniser to identify phrases in a text?

(7 marks)

iii. Why is it important to identify and link synonyms when creating a document vector? Explain how this can be done.

(6 marks)

b) The table below illustrates two processes: the first process creates a dataset of document vectors which is then used to train a classifier such as a decision tree; the second process also converts text samples into a document vector, and then applies the decision tree model generated in the first process.

Identify **three** problems with the steps taken in the second process.

Process 1: **Train a model**
1. Process documents using a binary count
2. Filter stop words
3. Apply a dictionary stemmer
4. Apply a porters stemmer
5. Train a decision tree model

Process 2: **Apply the model to unseen data**
1. Process documents using TF-IDF
2. Apply a Lovins stemmer
3. Apply the decision tree model to the data

(6 marks)

[Total: 25 marks]