| | |
|---|---|
| **Started on** | Friday, 11 May 2018, 1:04 PM |
| **State** | Finished |
| **Completed on** | Friday, 11 May 2018, 1:34 PM |
| **Time taken** | 29 mins 59 secs |
| **Grade** | **9.0** out of 10.0 (**90**%) |
| **Feedback** | Well done! |

**Question 1**

Correct

Mark 1.0 out of 1.0

Text mining is founded in Information Retrieval research, which also developed Web Search. How does Information Retrieval relate to Database Systems?

IR systems work well for some things that Database systems do not, such as → unstructured and semi-structured documents, approximate queries, best answers a ✓

Database systems are designed for some things that IR systems do not offer, including → transaction management, structured objects, specialist data types like spatial and te ✓

Your answer is correct.

The correct answer is: IR systems work well for some things that Database systems do not, such as → unstructured and semi-structured documents, approximate queries, best answers amongst many alternatives, Database systems are designed for some things that IR systems do not offer, including → transaction management, structured objects, specialist data types like spatial and temporal, updates

**Correct**

Marks for this submission: 1.0/1.0.

**Question 2**

Correct

Mark 1.0 out of 1.0

NEGATIVE MARKS ARE AWARDED FOR WRONG ANSWERS

It is usual for any text retrieval or text mining application to pre-process collection data as follows.

## Select one or more:

- [x] a. *Normalise* to translate each token to a form that will be considered equivalent to the original form ✓
- [x] b. Remove *stop words* ✓
- [x] c. *Tokenise* to split ext into words and remove punctuation ✓
- [ ] d. Build a term document matrix
- [ ] e. *Map* terms to words
- [x] f. *Stop words* could be customised to the domain of application so that overly frequent, non-selective words in the domain are removed ✓
- [ ] g. Generate a *word cloud*

Your answer is correct.

The correct answers are: *Tokenise* to split ext into words and remove punctuation, *Normalise* to translate each token to a form that will be considered equivalent to the original form, Remove *stop words*, *Stop words* could be customised to the domain of application so that overly frequent, non-selective words in the domain are removed

**Correct**

Marks for this submission: 1.0/1.0.

**Question 3**

Correct

Mark 1.0 out of 1.0

Which of the following terms has been stemmed?

## Select one:

- ( ) a. run very fast
- ( ) b. ran
- ( ) c. running
- ( ) d. Usain Bolt
- (•) e. run ✓

Your answer is correct.

The correct answer is: run

**Correct**

Marks for this submission: 1.0/1.0.

**Question 4**

Correct

Mark 1.0 out of 1.0

NEGATIVE MARKS ARE AWARDED FOR WRONG ANSWERS

| Term | Total count | Doc1 | Doc2 | Doc3 |
|---|---|---|---|---|
| animal | 6 | 0 | 1 | 5 |
| crazy | 3 | 0 | 0 | 3 |
| diet | 2 | 2 | 0 | 0 |
| eat | 5 | 3 | 0 | 2 |
| fast | 1 | 0 | 0 | 1 |
| feed | 2 | 2 | 0 | 0 |
| food | 8 | 5 | 1 | 2 |
| slow | 5 | 0 | 1 | 4 |
| melbourne | 4 | 4 | 0 | 0 |
| zoo | 1 | 0 | 1 | 0 |

The table here shows a term-document matrix (TDM). Which of the following are true statements about the TDM?

## Select one or more:

- [x] a. If we update the TDM for Doc 4 which says "Elephants roam freely in Melbourne." then the Total count for term "melbourrne" will increase by 1 to 5 ✓
- [ ] b. Allowing for stop words and normalisation, it is plausible that Doc2 is the sentence "Zoo animals in Melbourne eat their food slowly."
- [x] c. Allowing for stop words and normalisation, it is plausible that Doc2 is the sentence "Zoo animals eat their food very slowly." ✓
- [ ] d. The feature vector (2,3,2,5,4 ) represents Doc1
- [ ] e. "animal" is the most frequent word in the corpus

Your answer is correct.

The correct answers are: Allowing for stop words and normalisation, it is plausible that Doc2 is the sentence "Zoo animals eat their food very slowly.", If we update the TDM for Doc 4 which says "Elephants roam freely in Melbourne." then the Total count for term "melbourrne" will increase by 1 to 5

**Correct**

Marks for this submission: 1.0/1.0.

**Question 5**

Correct

Mark 1.0 out of 1.0

Consider the  query "Who is crazy enough to eat food in Melbourne"?

Referring to the DTM above, what is the feature vector for the query?

## Select one:

- a. (0,1,0,1,1,1)
- b. (0,0,1,0,0,1,1,0,1)
- c. (0,0,1,0,0,1,1,0,1,0)
- d. (0,0,2,3,0,2,5,0,4,0)
- e. (0,1,0,1,0,0,1,0,1,0) ✓

Your answer is correct.

The encoding of the query is (0,1,0,1,0,0,1,0,1,0)

The correct answer is: (0,1,0,1,0,0,1,0,1,0)

**Correct**

Marks for this submission: 1.0/1.0.

**Question 6**

Correct

Mark 1.0 out of 1.0

Consider the  query "Who is crazy enough to eat food in Melbourne"?

Referring to the DTM above, what is the cosine similarity of the query and Doc1?

Give your answer to two decimal places.

Answer: 0.79

Who is crazy enough to eat food in melbourne?

Use the formula 12/sqrt(232)

$$Sim(D_i, D_j) = \frac{\sum_{t=i}^{N} w_{it} * w_{jt}}{\sqrt{\sum_{t=1}^{N} (w_{it})^2 * \sum_{t=1}^{N} (w_{jt})^2}}$$

Sim(q, doc1) = 12/15.23 = 0.79

The correct answer is: 0.79

**Correct**

Marks for this submission: 1.0/1.0.

**Question 7**

Correct

Mark 1.0 out of 1.0

NEGATIVE MARKS ARE AWARDED FOR WRONG ANSWERS

Refining the term frequency vector, we can use l*ogarithmic term frequency* or *inverse document frequency* instead of raw term frequency as above because

## Select one or more:

- a. commonly, document vectors in the corpus will be weighted by term frequency times inverse document frequency, but query terms by logarithmic term frequency, and the cosine similarity of those determines the relevance of the document to the query
- b. inverse document frequency measures the frequency of a term in just one document, not the whole corpus
- c. compared to raw term frequency, logarithmic term frequency recognises that while term importance to a document increases with each repetition, the significance of each additional repetition drops off when frequencies are higher. ✓
- d. logarithmic term frequency (or even raw term frequency) and inverse document frequency may be multiplied to give *TF-IDF* which gives more weight to frequent terms (from the TF part) and to discriminative, selective terms (the IDF part). ✓
- e. logarithmic term frequency assigns a higher weight to terms that are rare in the corpus compared to terms that are frequent in the corpus as such terms discriminate amongst alternative documents
- f. compared to raw term frequency, logarithmic term frequency reduces the importance of terms seeming very frequent due to the document being long. ✓

Your answer is correct.

The correct answers are: compared to raw term frequency, logarithmic term frequency reduces the importance of terms seeming very frequent due to the document being long., compared to raw term frequency, logarithmic term frequency recognises that while term importance to a document increases with each repetition, the significance of each additional repetition drops off when frequencies are higher., logarithmic term frequency (or even raw term frequency) and inverse document frequency may be multiplied to give *TF-IDF* which gives more weight to frequent terms (from the TF part) and to discriminative, selective terms (the IDF part).

**Correct**

Marks for this submission: 1.0/1.0.

**Question 8**

Correct

Mark 1.0 out of 1.0

Match the mining method to the text mining problem it can solve.

| | |
|---|---|
| Hierarchical clsutering can be used to ✓ | group related documents and to derive a taxonomy of terms to describe the gro |
| Classification algorithms, such as SVM, Naive Bayes, and Decision trees can be used to ✓ | classify documents modelled as weighted term vectors into topic classes |
| By viewing the set of terms in a document as a transaction, associating mining can be used to ✓ | discover multi-word named entities because they occur frequently together |

Your answer is correct.

The correct answer is: Hierarchical clsutering can be used to → group related documents and to derive a taxonomy of terms to describe the groups, Classification algorithms, such as SVM, Naive Bayes, and Decision trees can be used to → classify documents modelled as weighted term vectors into topic classes, By viewing the set of terms in a document as a transaction, associating mining can be used to → discover multi-word named entities because they occur frequently together

**Correct**

Marks for this submission: 1.0/1.0.

**Question 9**

Correct

Mark 1.0 out of 1.0

Web mining techniques include

## Select one or more:

- [x] a. mining site access logs for trend analysis, improved site design, and drivers to purchase ✓
- [x] b. mining the structure of XML or HTML to relate text to images for image labelling or to recreate the relationship of textual fragments to each other ✓
- [x] c. mining the content (text) in nodes together with the link structure, for example for finding authoritative pages on some topic ✓
- [ ] d. web site design
- [ ] e. captioning images by crowdsourcing
- [ ] f. digging deep for a gold standard corpus

Your answer is correct.

The correct answers are: mining the structure of XML or HTML to relate text to images for image labelling or to recreate the relationship of textual fragments to each other, mining the content (text) in nodes together with the link structure, for example for finding authoritative pages on some topic, mining site access logs for trend analysis, improved site design, and drivers to purchase

**Correct**

Marks for this submission: 1.0/1.0.

**Question 10**

Incorrect

Mark 0.0 out of 1.0

In Tom Mitchells' ACM Webinar 15 June 2017, he asks how we can develop **a theory of neural representations** of concepts, rather than just discovering a list of particular concepts for which we know how they are represented. What is his answer to that question?

Answer: Study the brain activities while giving people fMRI. ✗

The correct answer is: predic model

**Incorrect**

Marks for this submission: 0.0/1.0.