University of Glasgow

Tuesday 3 May 2022
09:30-11:00 BST
Duration: 1 hour 30 minutes
Additional time: 30 minutes
Timed exam – fixed start time

DEGREES of MSci, MEng, BEng, BSc, MA and MA (Social Sciences)

TEXT AS DATA H
COMPSCI 4074

(Answer all 3 questions)

This examination paper is an open book, online assessment and is worth a total of 43 marks.

1. Question on Distributional Semantics and Word Embedding. (Total marks: 14)

Consider the problem of finding an outlier word from among a list of other similar words, e.g., out of the following set of words -

```
linux,  windows,  solaris,  android,  java,
```

the word 'java' is an outlier (because the other words are names of operating systems).

Given a list of such words your task is to automatically infer the outlier word. With respect to this task, answer the following

(a) An approach to solve the word intrusion problem is to represent words as vectors and then make use of the relative distances/similarities between the vectors for finding the outlier.

Assuming you know (by the output of some process) the vector $\mathbf{w}$ for a word w, describe the pseudo-code of finding the outlier word.

Task: Describe the pseudo-code of this algorithm. Clearly state your assumptions and introduce your notations in the algorithm.

[5]

One solution to the word outlier detection problem that does not require learning any parameters (via gradient descent) is the distributional semantics vector approach, where each word is represented via a bag-of-words vector of contexts. Now, **answer the following:**

(b) The window size, k, used to define the contexts for each word is an important parameter of this approach. What happens if k is too large or too small? [2]

(c) Describe the pseudo-code of this approach that requires only a single pass through a collection (clearly describe the data structures for an efficient solution). [5]

(d) Discuss (with an example) why the vectors of function words (frequent words, such as 'of', 'the' etc.) obtained with this approach are not reliable. [2]

2. Question on word frequencies and language model (Total marks: 15)

An alien probe crashes to Earth containing a short passage of alien text. The alien text uses a five letter alphabet: [a, b, c, d, e] with no punctuation or spaces. Below is a short section of the text:

abcaedabccbaedabceda

(a) Using character n-grams, write out all of the trigrams that appear more than once with their frequency for the sample text above. [3]

(b) Provide the theoretical maximum number of character n-grams for the alien probe full text for n = 1, 2, 3, 4 and 5. The full text found in the probe is 593 characters long. [3]

(c) A linguist makes a breakthrough in understanding the tokens used in the alien text. She provides two possible ways to tokenize the sample text.

(i) In plain English, explain a single rule that could reproduce this first tokenization

a	bca	eda	bccba	eda	bceda

[1]

(ii) In plain English, explain a single rule that could reproduce this second tokenization

ab	caedab	ccbaedab	ceda

[1]

(d) More alien probes crash land in different parts of the world. Scientists want to measure the similarity between the text found in each probe. Here are two tokenized probe texts fragments.

Probe Text A:

a	eda	bceda	eda	bcda	bce

Probe Text B:

ca	eda	bcba	eda	bceda	eda	bce

(i) Calculate the Sørensen–Dice Coefficient and Jaccard Similarity between the two probe texts. Show your work.

[4]

(ii) Calculate the similarity between the third probe text (Probe Text C below) and the two prior probe texts using the Sørensen–Dice Coefficient. Using these results, show that the Sørensen–Dice Coefficient is a semi-metric as it breaks the triangle inequality.

Probe Text C:

beda	bceda	bceca	ebeda	bceda	b

[3]

3. This question is about Natural Language Processing (Total marks: 14)

You just landed an awesome job at the Intellectual Property Office. As your first project, you have been tasked with automatically classifying submitted patent applications into one of the eight broad International Patent Classification sections, as shown here:

Section	Subject Matter	Section	Subject Matter
A	Human necessities	E	Civil engineering; Building accessories
B	Performing operations; Transporting	F	Mechanics; Lighting; Heating
C	Chemistry; Metallurgy	G	Instrumentation
D	Textiles; Paper	H	Electricity

(a) You start by applying a typical pre-processing pipeline that consists of case normalisation and a stemmer. Within the context of patent classification application, clearly justify these two pre-processing stages and provide an example that shows why it could lead to improved classification performance.

[4]

(b) You recall from Text as Data that NLP features, such as parts of speech, are often helpful for classification tasks. Within the context of patent classification, provide and justify a specific example where considering a word along with its part-of-speech may help distinguish between two of the above sections. [3]

(c) Armed with the above intuition, you select an off-the-shelf part-of-speech tagger (based on a Hidden Markov Model) that reports 97% accuracy and apply it to some sample patents to ensure that it produces reasonable part-of-speech tags. To your dismay, you find that it frequently makes mistakes. On closer inspection, you observe that the errors are usually on specialised, domain-specific language in the patents. Explain why this problem arises and what you could do to fix it. [4]

(d) You want to identify whether two systems (called System A and System B) are better than a baseline method at the classification task. The following table shows intrinsic evaluation metrics obtained over the classification on the train and test sets:

	Train Set		Test Set	
	Precision	Recall	Precision	Recall
Baseline	0.61	0.42	0.56	0.50
System A	0.62*	0.43*	0.58*	0.51
System B	0.67*	0.48*	0.51	0.42

* statistical significance w.r.t. baseline (t-test with p-value < 0.05)

Discuss the effectiveness (e.g., generalizability, overfit/underfit, performance on training/test sets etc.) of the models A and B in comparison to the 'Baseline' method. [3]