



Autumn Examinations 2020/2021

| | |
|--|--|
| Course Instance Code(s) Exam(s) | 1CSD1, 1CSD2, 1SPE1, 1MAO2, 1MAI1 MSc in Computer Science (Data Analytics), MSc in Computer Science (Artificial Intelligence), MSc in Computer Science (Artificial Intelligence) - Online |
| Module Code(s) Module(s) | CT5120, CT5146 Introduction to Natural Language Processing, Introduction to Natural Language Processing - Online |
| Paper No. | 1 |
| Repeat Paper | Yes |
| External Examiner(s) | Professor Pier Luca Lanzi |
| Internal Examiner(s) | Dr. Michael Madden *Dr. Paul Buitelaar, Dr. John McCrae |

Instructions: Answer all parts of all questions. There are 4 sections; each section is worth 25 marks (100 marks total). **Use a separate answer book for each section answered.**

| | |
|-------------------------------|---|
| Duration | 2 hours |
| No. of Pages | 5 |
| Discipline(s) | Computer Science |
| Course Co-ordinator(s) | Dr. Frank Glavin, Dr. Matthias Nickles, Dr. James McDermott |

Requirements:

| | | | | |
|----------------------------|------|-------------------------------------|----|-------------------------------------|
| Release in Exam Venue | Yes | <input checked="" type="checkbox"/> | No | <input type="checkbox"/> |
| MCQ | Yes | <input type="checkbox"/> | No | <input checked="" type="checkbox"/> |
| Handout | None | | | |
| Statistical/ Log Tables | None | | | |
| Cambridge Tables | None | | | |
| Graph Paper | None | | | |
| Log Graph Paper | None | | | |
| Other Materials | None | <input type="checkbox"/> | | <input checked="" type="checkbox"/> |
| Graphic material in colour | Yes | | No | |

Introduction to Natural Language Processing

Exam Duration: 2 Hours

You must complete Sections 1 to 4

Section 1: Linguistics; Vector Space Model; Semantics

Instructions: Provide answers for questions 1A, 1B and 1C

Question 1A

5 Marks

Explain the difference between stemming and lemmatization. Give an example of each.

Question 1B

10 Marks

Consider the following sentences.

The man met the woman.

The woman met the man.

The man met the woman and the man.

The woman and the man met the woman.

Define the grammar and lexicon G that can be used to generate these sentences by giving N , Σ and completing the rules P as follows:

$G=(N,\Sigma,P,S)$

N :

Σ :

P : $S \rightarrow NP VP$

Start symbol S

Question 1C

10 Marks

Consider the following frequency vectors. Using cosine similarity, compute the distributional semantic distance between 'dog' and 'cat'.

| | | | | | |
|-----|---|---|---|---|---|
| cat | 0 | 5 | 5 | 2 | 0 |
|-----|---|---|---|---|---|

| | | | | | |
|-----|---|---|---|---|---|
| dog | 0 | 3 | 4 | 2 | 6 |
|-----|---|---|---|---|---|

PTO

Section 2: Language Modeling; Tagging & HMMs; Probabilistic Parsing

Instructions: Provide answers for question 2A, 2B and 2C

Question 2A

10 Marks

Consider a Hidden Markov Model with the following probabilities (Start designates the start state):

| $p(w_i t_i)$ | $w_i = \text{this}$ | $w_i = \text{question}$ | $w_i = \text{is}$ | $w_i = \text{easy}$ |
|--------------|---------------------|-------------------------|-------------------|---------------------|
| $t_i = N$ | 0.4 | 0.4 | 0.1 | 0.1 |
| $t_i = V$ | 0.1 | 0.2 | 0.5 | 0.2 |
| $t_i = O$ | 0.4 | 0.1 | 0.1 | 0.4 |

| $p(t_i t_{i-1})$ | $t_{i-1} = N$ | $t_{i-1} = V$ | $t_{i-1} = O$ | $t_{i-1} = \text{Start}$ |
|------------------|---------------|---------------|---------------|--------------------------|
| $t_i = N$ | 0.2 | 0.2 | 0.7 | 0.1 |
| $t_i = V$ | 0.7 | 0.3 | 0.1 | 0.2 |
| $t_i = O$ | 0.1 | 0.5 | 0.2 | 0.7 |

By using the Viterbi algorithm or otherwise, what is the most likely sequence of tags for the text “this question is easy”?

Question 2B

10 Marks

Given a corpus with a part-of-speech tag values for each word, how would you learn the probabilities for a table such as in question 2A?

Question 2C

5 Marks

How would you modify the Viterbi algorithm in order to produce the probability of the text over all possible combinations of part-of-speech tags?

PTO

Section 3: Information Extraction; Knowledge Graphs & Chatbots

Instructions: Provide answers for question 3A, 3B and 3C

Question 3A

10 Marks

Explain what Hearst patterns are, how they are used in information extraction and give an example of three different Hearst patterns with a corresponding example sentence for each.

Question 3B

5 Marks

Consider the following positive instance (Source Wikipedia) for the relation Play_For between entities of type FOOTBALLER and TEAM:

[FOOTBALLER Kevin De Bruyne] is a Belgian professional footballer who plays as a midfielder for [TEAM Premier League club Manchester City].

Give a negative instance for this relation and these entity types, not involving negation.

Question 3C

10 Marks

Give all taxonomy elements, terms and term pairs, that can be identified in the following text (Source Wikipedia):

A bank is a financial institution that accepts deposits from the public and creates a demand deposit while simultaneously making loans. A bank borrows money by accepting funds deposited on current accounts, by accepting term deposits, and by issuing debt securities such as banknotes and bonds.

PTO

Section 4: Opinion Mining, Ethics & Data Privacy

Instructions: Provide answers for questions 4A, 4B and 4C

Question 4A

10 Marks

What is SentiWordNet and how can it be used in sentiment analysis?

Question 4B

5 Marks

For the following review, identify the sentiment aspects:

Exceptional customer service from all staff we couldn't have asked for a warmer welcome. Food is amazing too and such a great location for travelling around Cork.

Question 4C

10 Marks

Describe in your own words NLP aspects of data privacy.

END