8  **Natural Language Processing (SHT)**

| |
|---|
| Doc 1: |
| ... a large variety of plant life or flora in South Africa... |
| ... an entire plant kingdom inside its borders ... |
| ... more plant species than the entire of the UK... |
| ... related sets of plant and animal life... |
| ... the largest genus of flowering plants... |
| Doc 2: |
| ...tomato processing plant ... |
| ...plant technicians jobs at P&P... |
| ...process operations of the plant... |
| ...plant management... |
| ...CEO of the plant... |
| Doc 3: |
| ...plants' ability to adapt to harsh circumstances... |
| ...half human, half plant... |
| ...traditional food plants... |
| ...tomato plant... |
| ...spectacular plant... |
| ...medicinal plants... |
| Doc 4: |
| ...popular magazine "Plant Life", which humourously describes daily life at the Imperial food processing plant... |
| ...published by a chemical plant process worker... |
| ...small molecular commercial plants... |
| ...pharmaceutical plants... |
| ...heavy plant crossing... |

(a) How many senses of "plant" can you identify in the snippets from the 4 documents given above? Describe each sense by giving a hypernym.   [2 marks]

(b) Describe how Yarowsky's algorithm for word sense disambiguation would process the example texts. Illustrate each stage of the algorithm with an example.
[6 marks]

(c) Under adverse conditions, the algorithm in part (b) can rapidly diverge from a good solution. What are these conditions? Illustrate using the examples above.
[4 marks]

(d) How could you apply the Naive Bayes machine learning algorithm, discussed in the lectures in combination with several tasks, as a method for learning the word senses for an ambiguous word? Give the formula and explain how the necessary parameters can be trained. [8 marks]