**CS 5316–Natural Language Processing**

Midterm Exam

March 10, 2019

Duration: 2 hours (10.30 to 12.30)

**Instructions:**

1. Please write legibly. Unreadable answers will NOT be graded. Use a (dark) BALL-POINT and write in a readily readable font size.
2. Write in the spaces provided for the questions. Answers written outside their respective spaces will not be graded without prior approval (signature of invigilator).
3. This is a closed books/notes exam. You may use a help sheet and a calculator.
4. Distribute your time properly as some questions might be more involved than others.
5. There are 14 questions in this exam. The first 10 are worth 4 points each, and the last 4 are worth 15 points each.
6. Briefly describe and/or distinguish between pragmatic and syntactic knowledge in linguistics.
7. Given the dictionary below, segment the sequence a3bba5kaabb3 using max-match algorithm.

a, bb, a3, bba 5kaa

1. Write a regex for dates using one or two digits for day and month, 2 or four digits for year (last component), separator can be dash (-) or slash (/) (e.g., 9/5/18, 09/05/2018, 9-5-18).

1. Consider the following sentences written in informal language. (a) List the normalization steps you will apply before text classification, (b) List the tokens and types after your normalization steps.

The phone looks GREAT!!!!

I like the fon,,, it has gr8 looks, goooood color

Me liking n loving the phone ALOT

(Do this on the back-side of the previous page)

1. Explain briefly (with an example) why MLE is not recommended for language modeling.
2. In Good-Turing estimation, if and , calculate the discounted count of an item that is observed 5 times.
3. Sentiment analysis extracts and aggregates affective states expressed in natural language. In addition to attitude analysis, give examples of two other types of sentiment analyses.
4. Consider the following confusion matrix obtained during the evaluation of a POS tagging algorithm. The rows indicate true tags and the columns indicate predicted tags. Compute the macro averaged precision, recall, and F1-value of the algorithm.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Noun | Verb | Adjective |
| Noun | 550 | 40 | 60 |
| Verb | 100 | 700 | 200 |
| Adjective | 20 | 50 | 300 |

1. Using the IOB encoding, label the following sentence with NER tags: Prime Minister Imran Khan took oath in Islamabad.
2. Write the computational time complexity of solving the decoding problem using (a) Vitterbi algorithm, and (b) brute-force approach. Define the parameters in the expression.
3. 15 points) Determine the minimum edit distance and the corresponding alignment between BARLEY and BEARY. Do this using (a) unit costs for insertion, deletion, and substitution, and (b) insertion has cost 1, deletion has cost 2, and substitution has cost 3.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. ((15points) Consider the following five documents belonging to three classes (d1, d2 is class 1, d3, d4 is class 2, d5 is class 3). For each document, the words and their counts are given.

d1:plot/1, absurd/1, okay/1, bad/2

d2:screenplay/1, terrible/1, novel/1, absurd/1

d3:plot/1, engaging/1, good/2

d4:screenplay/1, excellent/2, okay/1,

d5:plot/1, okay/2, novel/1

* 1. (8 points) Develop a multinomial naïve Bayes classifier for this problem. Show the model and the estimated parameters using add-one smoothing.
  2. (7 points) Determine the best class for a new document: good screenplay, good plot.

1. (15 points) Given the following passage, estimate uusing (a) add-one smoothing, and (b) Kneser-Ney smoothing. Take discount equal to 0.75.

New Delhi's narrative on the India-Pakistan standoff appeared to crumble further on Wednesday as foreign journalists unearthed new details about the events of the preceding weeks.

1. (15 points) Consider an HMM over 2 states and 3 events, and the probability matrices as given below. Compute the likelihood of the event sequence .

0.8 0.2

0.7 0.3

0.4 0.2

0.3 0.2

0.3 0.6

Assume all probabilities from initial state and to final state are uniform.