**CS 5316 – Natural Language Processing**

**Quiz 1 Solution**

**(Time limit: 12 minutes)**

1. (6 points) Write regular expressions to find the following sets of strings:
   1. Acronyms consisting of two or more uppercase letters followed by a period, e.g., U.S., L.U.M.S.
   2. Integers in the interval [-999, +999]. E.g., +20, 25, -335 (note that the + sign is optional for a positive integer).
   3. HTML tags starting with < and ending with >
   4. Space delimited words at the start of a line starting with a lowercase alphabet

(a)

([A-Z]\.)+

(b)

[+-]?[0-9]{1,3}

(c)

<.\*>

^[a-z]+ (or spaces can be specified by \s)

1. (2 points) List at least two significant NLP tasks performed by a chatbot.

Depending on the granularity, there can be several NLP tasks that are performed by a chatbot.

* Text preprocessing (tokenization, normalization, etc)
* Sentence boundary detection
* Intent classification
* Named entity recognition (e.g., names, dates, locations, etc)
* Information extraction
* Language generation
* Semantic search (e.g., for response retrieval from a database)
* Question answering

1. (2 point) Determine the number of types (words) and tokens in the following corpus (a) before lemmatization and (b) after lemmatization:

He went to see his doctor

He saw his doctors yesterday

He is going to consult a physician

(a)

Types:

he went to see his doctor saw doctors yesterday is going consult a physician

14

Tokens: 18

(b)

Types:

he go to see his doctor yesterday is consult a

10

Went, going 🡪 go; see, saw 🡪 see; doctor, doctors, physician 🡪 doctor

Tokens: 18