Natural Language Processing

Assignment- 1

TYPE OF QUESTION: MCQ

Number of questions: 10 **Total mark: 10 X 1 = 10**

Question 1: In a corpus, you found that the word with rank 4th has a frequency of 600. What can be the best guess for the rank of a word with frequency 300?

1. 2
2. 4
3. 8
4. 6

Answer: 3

Solution:
frequency * rank =k [by Zipfs law]
600*4 = 300*r
r = 8

Question 2: In the sentence, "The only thing we have to fear is fear itself", the ratio between total number of word tokens and word types is :

1. 8/10
2. 10/11
3. 10/10
4. 10/9

Answer: 4
Solution:
Count the number of word tokens and word types. # word token = 10, # word type =9

Question 3: Let the rank of two words, w1 and w2, in a corpus be 1600 and 400, respectively. Let m1 and m2 represent the number of meanings of w1 and w2 respectively. The ratio m1 : m2 would tentatively be

1. 1:4
2. 4:1
3. 1:2
4. 2:1

Answer: 3

Solution:
m1/m2 = sqrt(rank2)/sqrt(rank1) = sqrt(400)/sqrt(1600) = 1:2

Question 4: Which of the following is/are true?

1. Ambiguity can appear in Tokenization steps
2. Ambiguity will not appear in Sentence segmentation step
3. Function word is generally more frequent in a text than any content word.
4. Output of lemmatization are always real words

Answer: 1, 3

Solution:
In sentence segmentation there can be ambiguity. The output of lemmatization sometimes don't lead to real words

---

Question 5: If first corpus has $TTR_1 = 0.085$ and second corpus has $TTR_2 = 0.78$, where TTR_1 and TTR_2 represents type/token ratio in first and second corpus respectively, then Which of the following is /are false?

1. First corpus has more tendency to use different words.
2. Second corpus has more tendency to use different words.
3. TTR value sometime can be greater than 1
4. A high TTR indicates a high degree of lexical variation while a low TTR indicates the opposite.

Answer: 1, 3

Solution:

TTR can not be greater than 1 higher TTR indicates more tendency to use different words

---

Question 6: In linguistic morphology, _____________ is the process for reducing inflected words to their root form.

1. Stemming
2. Rooting
3. Text-Proofing
4. Both a & b

Answer: 1

Stemming is used to reduce inflected words to the root form

Question 7: Morphological Segmentation

1. Is an extension of propositional logic
2. Does Discourse Analysis
3. Separate words into individual morphemes and identify the class of the morphemes
4. None of the mentioned

Answer: 3

Solution:

Morphological segmentation separates words into individual morphemes and detects the class of it

---

Question 8: An advantage of Porter stemmer over a full morphological parser?

1. The stemmer is better justified from a theoretical point of view
2. The output of a stemmer is always a valid word
3. The stemmer does not require a detailed lexicon to implement
4. None of the above

Answer: 3

Solution: The Porter stemming algorithm is a process for removing suffixes from words in English. The Porter stemming algorithm was made in the assumption that we don't have a stem dictionary (lexicon) and that the purpose of the task is to improve Information Retrieval performance. Stemming algorithms are typically rule-based. You can view them as heuristic process that sort-of lops off the ends of words.

Question 9: What is natural language processing good for?

1. Summarize blocks of text
2. Automatically generate keywords
3. Identifying the type of entity extracted
4. All of the above

Answer: 4

Solution:
For all the above-mentioned task, NLP can be used

Question 10: What is the size of unique words in a document where total number of words = 12000. K = 3.71 Beta = 0.69?

1. 2421
2. 3367
3. 5123
4. 1529

Answer: 1

Solution:
3.71 x 12000^0.69 = **2421** unique words. Heap’s Law
