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 250. What can be the best guess for the rank of a word with frequency 125?

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

Answer: 3

Solution:
frequency * rank =k [by Zipfs law]
250*4 = 125*r
r = 8

Question 2:

In the sentence, “In Delhi I took my hat off. But I can’t put it back on.”, total number of word tokens and word types are:

1. 14, 13
2. 13, 14
3. 15, 14
4. 14, 15

Answer: a) 14, 13.
Solution: Here, the word “I” is repeated two times so type count is one less than token count.

Question 3:

Let the rank of two words, w1 and w2, in a corpus be 1600 and 100, 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: 1

Solution:

m1/m2 = sqrt(rank2)/sqrt(rank1) = sqrt(100)/sqrt(1600) = 1:4

Question 4:

What is the valid range of type-token ratio of any text corpus?

1. TTR∈(0,1] (excluding zero)
2. TTR∈[0,1]
3. TTR∈[−1,1]
4. TTR∈[0,+∞] (any non-negative number)

Answer: 1.

Solution: Number of unique words or type ≤ Total number of tokens in text, and both are greater than 1

Question 5:

If first corpus has $TTR_1 = 0.06$ and second corpus has $TTR_2 = 0.105$, where TTR_1 and TTR_2 represents type/token ratio in first and second corpus respectively, then

1. First corpus has more tendency to use different words.
2. Second corpus has more tendency to use different words.
3. Both a and b
4. None of these

Answer: b
Solution: Second corpus has more tendency to use different words. If TTR scores are higher then there is more tendency to use different words.

Question 6:

Which of the following is/are true for the English Language?

1. Lemmatization works only on inflectional morphemes and Stemming works only on derivational morphemes.
2. The outputs of lemmatization and stemming for the same word might differ.
3. Output of lemmatization are always real words
4. Output of stemming are always real words

Answer: 2, 3
Solution: *Stemming* usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. *Lemmatization* usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma* .

Question 7:

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 on 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 a heuristic process that sort-of lops off the ends of words.

Question 8:

Which of the following are not instances of stemming? (as per Porter Stemmer)

1. are -> be
2. plays -> play
3. saw -> s
4. university -> univers

Answer: 1, 3
Solution: Stemming cannot convert are->be as it can only convert or chop off word suffixes. Also Porter Stemmer wouldn't chop off if the final outcome is of length 1 as in saw -> s.

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

---

**************END*********