Natural Language Processing

Assignment 10

Type of Question: MCQ

Number of Questions: 10 **Total Marks:(10×1)= 10**

==========

Question 1.

Different phases of entity linking are -

A) Candidate Selection -> Reference Disambiguation
B) Reference Disambiguation -> Candidate Selection -> Mention Identify
C) Mention Identify -> Candidate Selection -> Reference Disambiguation
D) All of the above

Answer: A, C

Solution: Theory. Slide 7; Lecture 1, Week 10 (Entity Linking 1).

==========

Question 2.

The text span s="river" occurs in 700 different Wikipedia articles.

c1 223
c2 161
c3 78
c4 31
No Link 207

Calculate the keyphraseness of "Sea".

A) 0.232
B) 0.886
C) 0.688
D) 0.704

Answer D

Solution: $CF(s_i) / CF(s) = 223 +161+78+31 / 700 = 493 / 700 = 0.704$

==========

Question 3.
What is the commonness of (s, c3) in the above question?

A) 0.765
B) 0.389
C) 0.158
D) 0.910

Answer C

Solution: 78/(223+161+78+31) = 78/493 = 0.158

==========

Question 4.

Relevant feature/s for a supervised model for predicting the topics to be linked is/are:

A) Disambiguation Confidence
B) Relatedness
C) Link Probability
D) All of the above

Generality
Location and Spread

Answer: D

Solution: Theory. Entity Linking Lecture II [Lecture Video 47]

==========

Question 5.

Which of the following problem exists in bootstrapping technique for Information extraction are:

A) Sensitiveness towards the seed set
B) High precision
C) Less manual intervention
D) All of the above

Answer: A

Solution: Theory

==========

Question 6.

Which of the following is an advantage of unsupervised relation extraction:

A) Can work efficiently with small amount of hand-labeled data
B) Not easily generalizable to different relations
C) Need no training data.
D) Always perform better than supervised techniques.

Answer: C

72, 126, 132

Solution: Theory

==========

Question 7.

Which of the following is not a Hearst's Lexico Syntactic Patterns for automatic acquisition of hyponyms -

A) X or other Y
B) X and other Y
C) Y including X
D) X but not Y

Answer: D

Solution: Theory. Lecture Video 48 : Information Extraction - Introduction, Week 10, Lecture 3 - Slide 15/18.

==========

Question 8.

Advantage of Distant supervision over bootstrapping method

A) Need more data
B) Less human effort
C) Can handle noisy data better
D) No Advantage

Answer: C

Solution: Theory

==========

Question 9.

Consider a dataset with a very low number of relations - all of which are very important. For a relation extraction task on that dataset, which of the following is the most useful metric

A) Precision
B) Recall
C) Accuracy
D) F1-Score

Answer: B

Solution: Theory

A) Precision
Explanation: Precision measures how many of the predicted relations are actually correct. It focuses on minimizing false positives.

Usefulness in this context: While precision is important, it is not the most useful metric in this case, as it does not account for the possibility of missing important relations (false negatives).

B) Recall
Explanation: Recall measures how many of the actual important relations were successfully identified by the model, emphasizing minimizing false negatives.

Usefulness in this context: Since the dataset contains a very low number of critical relations, missing even one of them would have a significant impact. Therefore, recall is the most important metric here—it ensures that all key relations are captured.

C) Accuracy
Explanation: Accuracy is the percentage of correctly predicted relations (both true positives and true negatives) out of all predictions made.

Usefulness in this context: Accuracy can be misleading in cases with imbalanced datasets (like this one), as it does not differentiate between the importance of relations. It is not a good choice here.

==========

D) F1-Score
Explanation: F1-Score is the harmonic mean of precision and recall, combining both metrics into a single measure of performance.

Usefulness in this context: Although F1-Score is useful for balancing precision and recall, it is still secondary to recall in this specific case, where identifying all critical relations is paramount.

Question 10.

What is KeyPhraseness (wikipedia)?

A) Number of articles that mention a key phrase divided by the number of wikipedia articles containing it.
B) Number of Wikipedia articles that use it as an anchor, divided by the number of articles that mention it at all.
C) Number of articles that mention a key phrase times by the number of wikipedia articles containing it.
D) Number of Wikipedia articles containing the key phrases times by number of articles mentioning it.

Answer: B

Solution: Theory.

==========