NLP End Term exam

**The English language has 26 characters. The unigram probabilities of the characters A, N and T are 0.1, 0.05 and 0.01. What will be the perplexity of the sequences “NAN” and “ANT”. How will the perplexity change for the two sequences if the probabilities were equal for all the 26 characters? (4 marks)**

Solution:

Perplexity (“NAN”) = P(“NAN”)-1/3= (0.05 \* 0.1\* 0.05)-1/3

Perplexity (“ANT”) = (0.1 \* 0.05\* 0.01)-1/3

If the probabilities were the same, then

Perplexity (“NAN”) = ((1/26)\*(1/26)\*(1/26)-1/3

Perplexity (“ANT”) = ((1/26)\*(1/26)\*(1/26)-1/3

**A sentiment classifier predicts the labels as given below. Find the precision, recall and accuracy of the model. (3 marks)**

|  |  |
| --- | --- |
| **Expected** | **Predicted** |
| Positive | Negative |
| Positive | Positive |
| Negative | Negative |
| Positive | Positive |
| Negative | Positive |
| Negative | Negative |
| Negative | Negative |
| Positive | Positive |
| Positive | Negative |
| Negative | Negative |

Solution:

|  |  |  |
| --- | --- | --- |
|  | Pos Actual | Neg Actual |
| Pos Pred | 3 | 1 |
| Neg Pred | 2 | 4 |

Accuracy = 7/10

Precision = ¾

Recall = 3/5

**Find the tf-idf vectors and cosine similarity between the following documents.**

**d1= The best team plays the finals**

**d2= India won a medal in the finals (7 marks)**

Solution:

TF (2 marks) IDF (2 marks) TF-IDF (1 mark), Cosine Similarity (2 marks)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | d1 | d2 | IDF | TF-IDF d1 | TF-IDF d2 |
| the | 0.33 | 0.14 | 0.00 | 0.00 | 0.00 |
| best | 0.17 | 0.00 | 1.00 | 0.17 | 0.00 |
| team | 0.17 | 0.00 | 1.00 | 0.17 | 0.00 |
| plays | 0.17 | 0.00 | 1.00 | 0.17 | 0.00 |
| India | 0.00 | 0.14 | 1.00 | 0.00 | 0.14 |
| finals | 0.17 | 0.14 | 0.00 | 0.00 | 0.00 |
| won | 0.00 | 0.14 | 1.00 | 0.00 | 0.14 |
| a | 0.00 | 0.14 | 1.00 | 0.00 | 0.14 |
| medal | 0.00 | 0.14 | 1.00 | 0.00 | 0.14 |
| in | 0 | 0.14 | 1 | 0.00 | 0.14 |

Cosine Similarity = 0

**For the sentence, “Play this year’s French radio-hit pop songs”, a chatbot determines the following slots, Genre: Pop, Language: French, Year: 2020. Determine the intent accuracy and slot error rate. (1 mark)**

Solution:

Slot Error Rate = ⅓

Intent Accuracy = ⅔