LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer’s Name: Dr James Connolly

Assessment Title: Natural Language Processing CA 2 Submission 2

Work to be submitted to: Blackboard

Date for submission of work: 15-04-2020

Place and time for submitting work: Letterkenny (Home), 00:00

To be completed by the Student

Student’s Name: Pathan Faisal Khan

Class: Big Data and AI Group A

Subject/Module: Artificial Intelligence

Word Count (where applicable):

I confirm that the work submitted has been produced solely through my own efforts.

Student’s signature: Faisal Date: 15-04-2020

|  |
| --- |
| **Notes**  **Penalties:** The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero. [Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute’s Assessment Regulations.]  **Plagiarism:** Presenting the ideas etc. of someone else without proper acknowledgement (see section L1 paragraph 8).  **Cheating:** The use of unauthorised material in a test, exam etc., unauthorised access to test matter, unauthorised collusion, dishonest behaviour in respect of assessments, and deliberate plagiarism (see section L1 paragraph 8).  **Continuous Assessment:** For students repeating an examination, marks awarded for continuous assessment, shall normally be carried forward from the original examination to the repeat examination. |

CA 2 Submission

Pathan Faisal Khan (L00151142), BDA & AI Group A- AI 2 (NLP)

# Q 1. Text Classification

We found out using GridSearch that best parameters for LDA is {n\_components- 10, perplexity- 0.9}. We used CountVectorizer as it works on Probabilistic model which is also the underlying logic of LDA. We are considering that a word should come in atleast 2 documents (min\_df) and should not come in more than 90% of the documents (max\_df). We are also removing stop words in this process. With this configuration, we got 50,470 words from 2,00,000 documents/rows/questions.

We then did NMF with 10 components/topics and default 0.7 perplexity. We used TfidfVectorizer which gives better result as compared to CountVectorizer as Tfidf takes in account words in all documents. We are considering that a word should come in atleast 2 documents (min\_df) and should not come in more than 90% of the documents (max\_df). We are also removing stop words in this process. With this configuration, we got 27,884 words from 2,00,000 documents/rows/questions.

# Q 2. Supervised Learning

We have selected NMF classification data as it more accurately classified topics as compared to LDA. We found that the probability of words in a few topics were quite low based on the observations of the graphs.

We got the following accuracies:

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy (%)** |
| Logistic Regression | 87% |
| Navie Bayes | 71% |
| Random Forest | 90% |
| Support Vector Classifier | 21% |

We have noticed that Random Forest has the highest accuracy with 90%.