

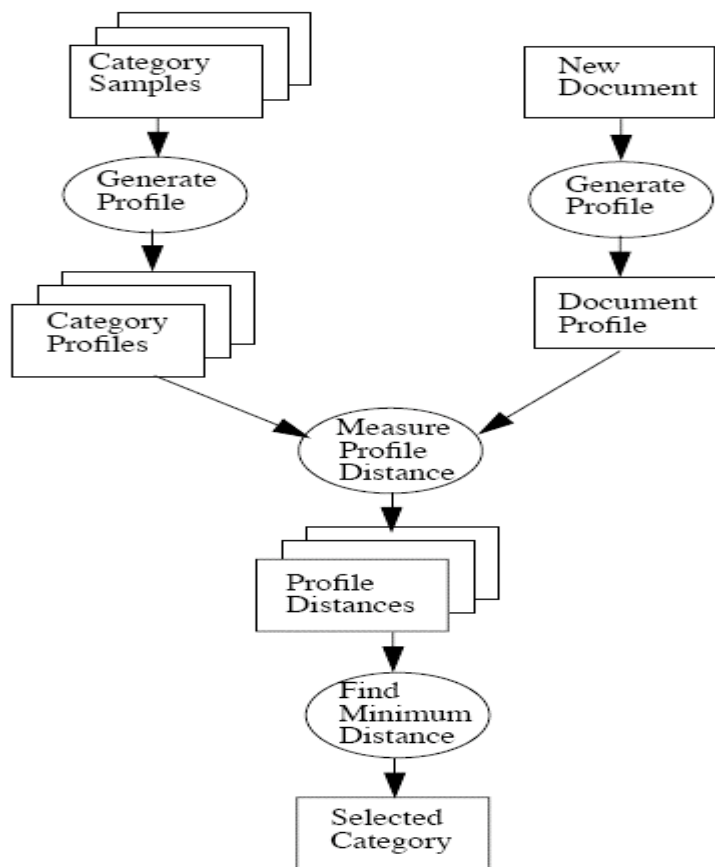
**Problem Statement:**

Implement and analyze n-grams based text categorization for Bangla language.

**Literature Review:**

Text categorization is a fundamental task in document processing, allowing the automated handling of enormous streams of documents in electronic form. In text categorization an incoming document is assigned to some pre-existing category.

This project was an generic implementation of the paper “**N-Gram-Based Text Categorization** “ by William B. Cavnar and John M. Trenkle. The flowchart of the algorithm that was followed in given bellow:



**Figure 1 . Flowchart of the n-gram base text categorization algorithm.**

Here category sample are golden standard which works as the base of categorization. Profiles were n-gram frequency. After that the profiles were categorized. For testing a text its own profile was generated and then that profile was compared with the existing profiles of golden standard. The profile distance was basically the absolute value of the frequency difference of the test document and documents of the golden standard

document. The least scoring distance was the answer for which category the test document belonged.

### **My solution:**

This algorithm was implemented in java and the Prothom-Alo 1 day corpus was used for categorization.

### **Analysis and comparison with other solution:**

The result of the categorization was not satisfactory. The reason are the following:

- The corpus size was small.
- The golden standards were not acceptable enough to have a good categorization.

### **Future work:**

The future works are the following:

- IDF ( Inverse Document Frequency ) is to be introduced to the implementation so that better performance can be achieved.
- Incorporate 1 year Prothom-Alo corpus. This will increase the golden standard for better text categorization.
- Text clusterization for Bangla Language.

### **Reference:**

1. Classification Algorithms on Text Documents.  
<http://www.cs.utexas.edu/users/hyukcho/classificationAlgorithm.html>
2. William B. Cavnar and John M. Trenkle , N-Gram-Based Text Categorization  
<http://www.citeseer.ist.psu.edu/68861.html>
3. N-gram based Text Categorization, by Peter Náther.
4. Text Categorization Using Acquaintance by Jonas Gustavsson  
<http://www.f.kth.se/~f92-jgu/C-uppsats/Cupps.e.html>

### **Acknowledgement:**

Special thanks to Dr. Mumit Khan Sir and Mr. Naushad UzZaman for their support.