QckRead

A Machine Learning Approach to Automatic News Article Summarization

Nikhil Ambha Madhusudhana	NAM603
Ashwin Shanmugasundaram	ASS211
Imran Shaikh	INS359

Abstract -

In this project we aim to build a machine learning approach to summarize a news article. The approach begins by first classifying the article based on its genre or news category. It then uses the information accumulated by the initial classification, to accumulate the greatest amount of information in the least amount of time. The method utilizes an approach we have designed where it looks for similar flow of ideas in sentences with reference to the title to identify important aspects and then weights sentences based on its importance and 'idea' content. The concept of an idea is to understand the general meaning of a word and look for similar words which convey similar ideas rather than look for naive word counts. We then assign weights to these 'ideas' based on their relevance to the genre of the article to extract the information most relevant to understanding an article. We successively apply this to every sentence in order to capture the maximum amount of information in the least amount of sentences.

Introduction -

Many a times readers of news articles find most of the news verbose and cumbersome to read at an entire stretch, but still like to get the essence of the news without having to read the entire article. Headlines of the news just provide the user with the subject of article thereby not providing the important details of its causes or effects. So, our project aims at summarizing the news article to a length where the reader is able to gain some important information about the same in 'QCK'-time. This is done by determining the important words and the sentences that incorporate that word.

Our first steps towards this are determining the genre of the article at hand. For this purpose we made use of dataset 20NewsGroup consisting of different number of articles of 20 different genres .i.e. approximately consisting 1000 articles per genre giving us a count of somewhat 20000 news articles.

We then formulated our own vocabulary with the help of English dictionary by referencing the vocabulary from 20NewsGroup thus giving us unique words which will help us determining the articles.

Subsequently after determining the vocabulary, we made use of NaiveBayes classifier to determine the probabilities of the words present over the entire articles of a genre the result of which was used in predicting the genre of that class.

Once the genre of the article is ascertained we then fabricate the important sentences in the article by assigning them weights and deduce the summary with the help of the highly weighted words and sentences.

Dataset –

The datasets used for the purpose of this project are multiple and the data extracted is used to perform specific functions to help provide the end result.

20newsgroup - First to begin with for the purpose of genre classification we have utilized the 20newsgroup dataset. This dataset is freely available for academic use. The dataset consists of around 20 thousand articles from various newsgroup classified into 20 categories with a near even distribution of articles across the categories. The dataset was used to perform and test the genre classification aspect of the project.

English Dictionary - The entire English dictionary has been downloaded and complied into structural data so as to enable us to cross reference words identified in the genre classification and the during summarization to ensure we discard conjunctions and proper nouns and to also establish the correctness of words so that the classifier does not contain garbage words occurring due to erroneous parsing of text files, which reduce its efficiency.

English thesaurus- The entire English Thesaurus is downloaded and structured for the purpose of understanding the 'ideas' behind words. The Thesaurus used in this case is Roget's english thesaurus from Gutenberg.org.

The thesaurus has been divided into buckets of words which portray a similar meaning or idea and each bucket assigned a number. So this is used to reduce the article into a sentence to an array of bucket numbers which is used to compare the similarity of ideas of and meanings of sentences as a whole.

Preliminary Experiments -

For formulating the vocabulary we cross-referenced the words from the English dictionary with the words present in the articles from the data set and removed the words which did not have an impact on either determining the genre of the article or weighting the most important sentences, these words may include all the conjunctions, proper nouns as most of the times proper nouns may result in different outcome based on the context of the article. We also extracted the delimiters from the sentences as they would also not have any impact in prediction of the genre.

Once we got the unique vocabulary as described above, we then build a matrix where each column of the matrix corresponded to the unique vocabulary obtained and each row represented the articles all across the genres, where the value at the cross section of each row and column of the matrix gave us the number of times a particular word from the vocabulary appears in the article.

We then determined the probability of each word present across the article given the genre of the article. We computed these probabilities for each word in each article across every genre as well as determined the probability of a genre across the number of genres present. Once these probabilities are calculated then for each test article we treat every word of it as attribute and locate whether this word is present in the vocabulary or not then for all the words across the test articles that are present in the vocabulary we determine the probability of the test article for being in every genre by multiplying the probability of a word given a genre with the number of times that word occur across the test article. Then we choose the genre of the test article as the one which results in the maximum probability.

We made a confusion matrix after determining the genre of all the articles in the validation which gave us better understanding of the number of articles that were predicted accurately and the number of articles that were wrongly predicted. This method gave us an accuracy of about 90-93%. The confusion matrix is a 20X20 matrix where each row gives the actual genre of the article and each column depicts the genre that the Naive-Bayes predicted, thus the diagonal element gives us the number of rightly predicted genre.

Final Methodology -

After exploring the preliminary methods and testing the results of each method , the final approach uses a combination of the earlier methods in combination with approach designed by us to provide the most relevant summary.

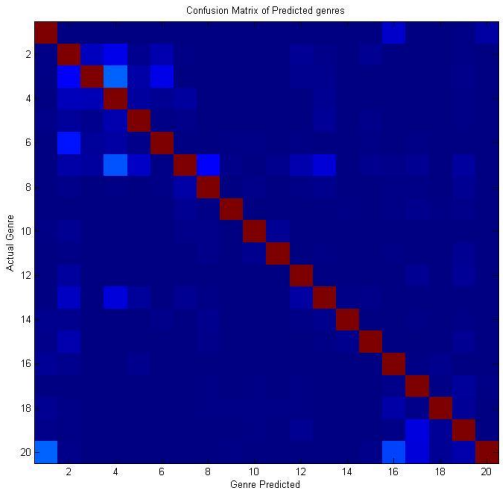

Fig 1 - The confusion matrix showing the articles predicted correctly and incorrectly, where the diagonal elements are of highest intensity showing that it predicted the articles correctly the y-labels been the actual genres and the x-label depicting the value that the algorithm predicted.

The algorithm begins by training the genre classifier using the naïve bayes technique with the use of the 20newsgroup dataset. The dataset for this purpose can be replaced by any other dataset of classified news articles and the algorithm will still work. After the classification of the news article and building the bayes classifier the data is cross referenced with the dictionary to eliminate conjunctions and proper nouns. This ensures that all words are meaningful and get a hit when we look for their buckets in the thesaurus. The thesaurus is then read in and structured into buckets to so as to be readily available when summarizing the article and so that the article can be easily converted into the bucket format. The article to be summarized is read in and split into independent sentences. Once this is done the sentences are converted into bucket arrays. This provides the basis for estimating the similarity of ideas within sentences.

The title of the article is used in the beginning as reference to establish the important aspects of the article. Now we look for a similar flow of ideas in sentences with respect to the title by looking for a sentence with the longest common subsequence with respect to the title. This establishes the sentence that provides the supporting argument to the title hence providing a deeper understanding of the article. Now we remove the sentence and insert it into the article. The other 'ideas' present in the sentence not part of the subsequence are now taken care of by assigning them weights based on the probabilities of similar bucket words used to classify the genre of the article. If the words do not exist in the current vocabulary their weights are only assigned with respect to the current article and its occurrences in the important sentences classified. The weights of the words which were not present in the vocabulary but appeared in the summary of the article are added to the vocabulary with low threshold weights. These words will be reinforced by summarizing more articles and similarly modifying word weights based on whether or not they appeared in the summary of the article. This provides for continuous automatic updation of the knowledge base so as to improve the summarization at every successive iteration.

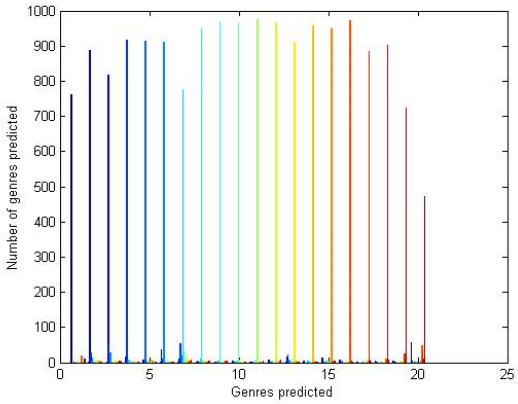

Fig 2 - This graph gives for each class the number of articles it has predicted incorrectly and number of articles it has predicted correctly, the bar with the maximum height is the final class predicted.

Conclusion -

News article summarization does not assure that the summary generated is 100% accurate as summarization depends on every individual reader but still can generate significant results giving us the gist of the entire news article in QCK-time.

Future Work -

For the future work we will try to implement the summarization technique directly on the website so that the news article can be summarized in place as well as we will try to include conjunctions and proper nouns in the dictionary according to the context of the article in which they appear, so that it will generate better summarization results.

References –

Amari, S.-I. and Nagaoka, H. (2001). Methods of Information Geometry (Translations of Mathematical Monographs). Oxford University Press.

Aone, C., Okurowski, M. E., Gorlinsky, J., and Larsen, B. (1999). A trainable summarizer with knowledge acquired from robust nlp techniques. In Mani, I.
and Maybury, M. T., editors, Advances in Automatic Text Summarization, pages
71{80. MIT Press.

Barzilay, R. and Elhadad, M. (1997). Using lexical chains for text summarization.
In Proceedings ISTS'97.

Lin, C.-Y. and Hovy, E. (2002). Manual and automatic evaluation of summaries. In
Proceedings of the ACL-02 Workshop on Automatic Summarization, pages 45{51,
Morristown, NJ, USA.

McKeown, K. R. and Radev, D. R. (1995). Generating summaries of multiple news
articles. In Proceedings of SIGIR '95, pages 74{82, Seattle, Washington.