Project Name **Automatic Author Recognition for Turkish Texts**

Project Group Hasan ÇAKAR
Assist.Prof.Banu DİRİ (Advisor-2009)

Abstract

The interest of natural language processing and researches for it are rising these days. Because of these researches, the theories of linguistics can be tested more comprehensively and rapidly by creating an area of experiment for them. Thanks to these improvements on the field of natural language processing, the process on classifying of millions of electronic documents, especially on the Internet, has gained momentum.

The aim of this project work is to end disorder that untidy document cause and to arrange the documents by finding the authors of the text whose authors are unknown. In this project, recognition operation is made for maximum 10 authors. The authors are chosen between the people write about politics, sports, art-magazine, economics etc. The system is trained for each author by using 50 documents and occurred training set which is used the author recognition operation. Two methods, Hold out Method and Rotation Method, are used by producing the training set. Two methods are used author recognition and text classification works. Naive Bayes and k-Nearest Neighbor methods are used for finding the author of given test texts whose author is found. At the end of this project work %86 ratio of success for recognition of 10 authors has been caught, so a system which put an end to the confusion which appeared by the document number ratio that increase day by day has been developed.

Özet (In Turkish)

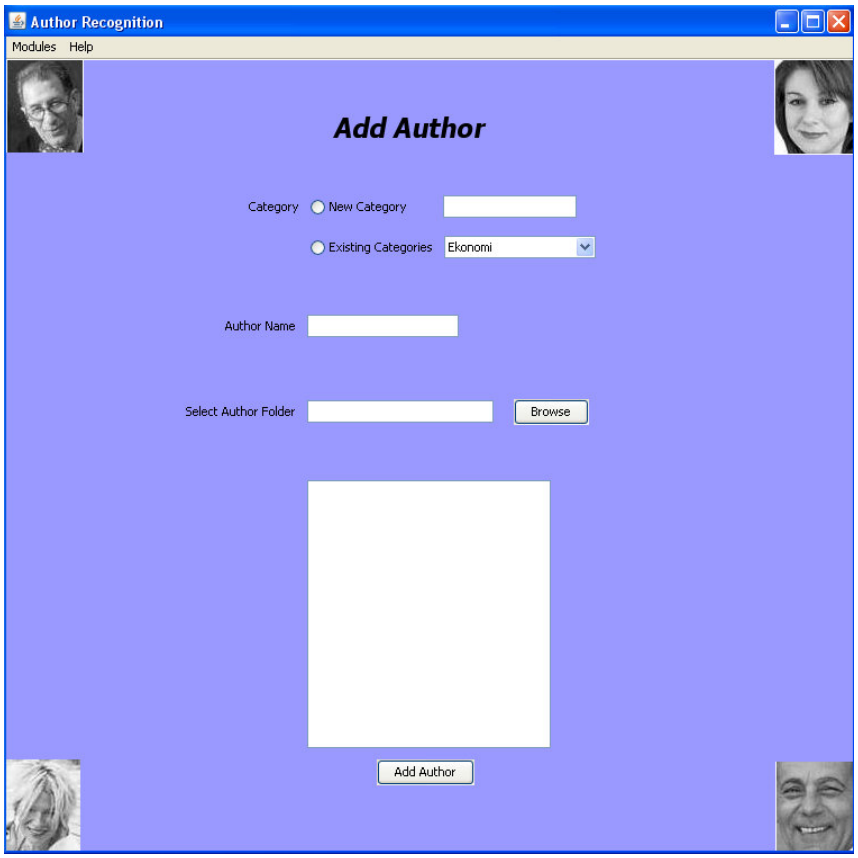
Günümüzde doğal dil işleme alanına olan ilgi ve yapılan çalışmalar giderek artmaktadır. Bu çalışmalar neticesinde dil bilim kuramlarına deney ortamı oluşturularak daha kapsamlı ve hızlı bir şekilde test edilmesi sağlanabilmektedir. Doğal dil işleme alanındaki bu gelişmeler sayesinde özellikle Internette bulunan milyonlarca elektronik dokümanın etkin şekilde sınıflandırılması üzerine yapılan çalışmalar hızlanmıştır.

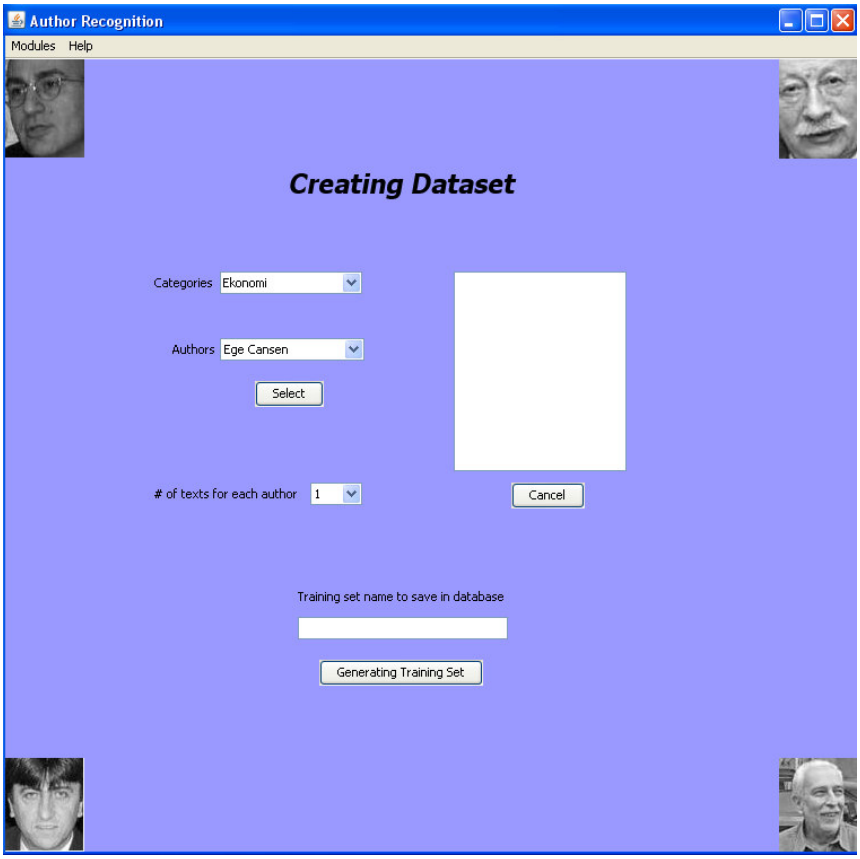
Bu projenin amacı düzensiz halde bulunan bu dokümanların yarattığı kargaşaya, yazıların yazarlarını bularak son vermektir. Projede, yazar tanıma işlemi en fazla 10 farklı yazar ile yapılmaktadır. Bu yazarlar politika, spor, sanat-magazin, ekonomi gibi alanlarda yazı yazanların arasından seçilmiştir. Sistem, her bir yazarın 50 yazısı ile eğitilir ve yazar tanıma işlemi için kullanılan eğitim seti çıkarılır. Eğitim seti oluşturulurken Hold out ve k-Cross Validation Method adında iki ayrı metot kullanılmıştır. Yazar tanıma ve doküman sınıflandırma için ise iki ayrı metot kullanılır. Naive Bayes ve k-Nearest Neighbor metotları verilen test dokümanlarının sahibinin bulunması işlemi için kullanılmıştır. Bu çalışma sonucunda 10 yazar için yapılan tanıma işlemi %86 oranında bir başarı oranı elde edilmiştir.

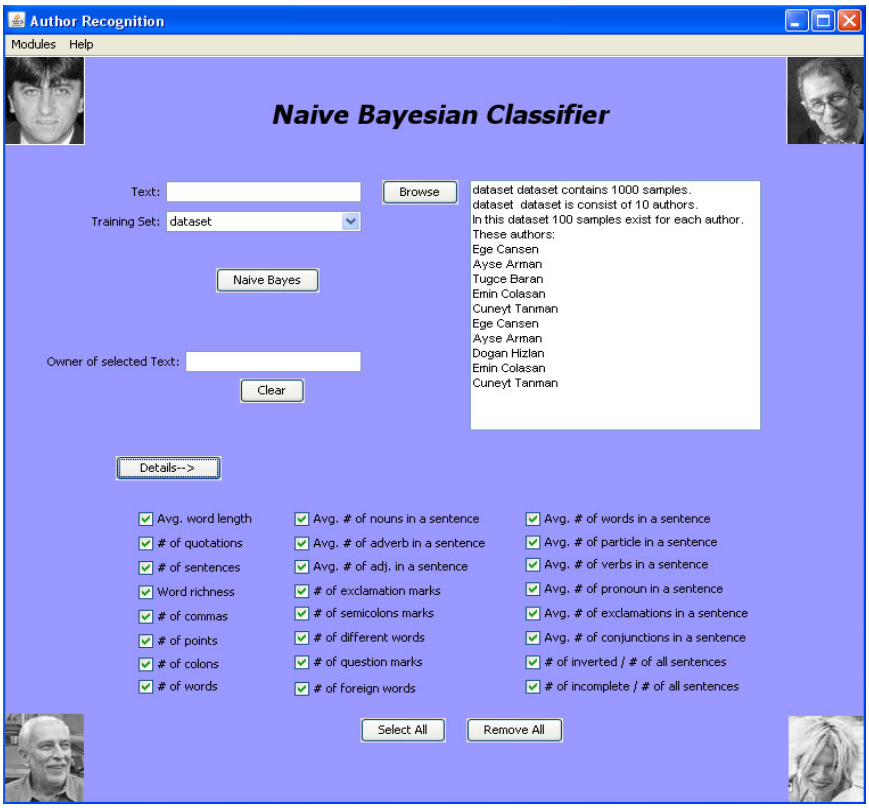
Requirements

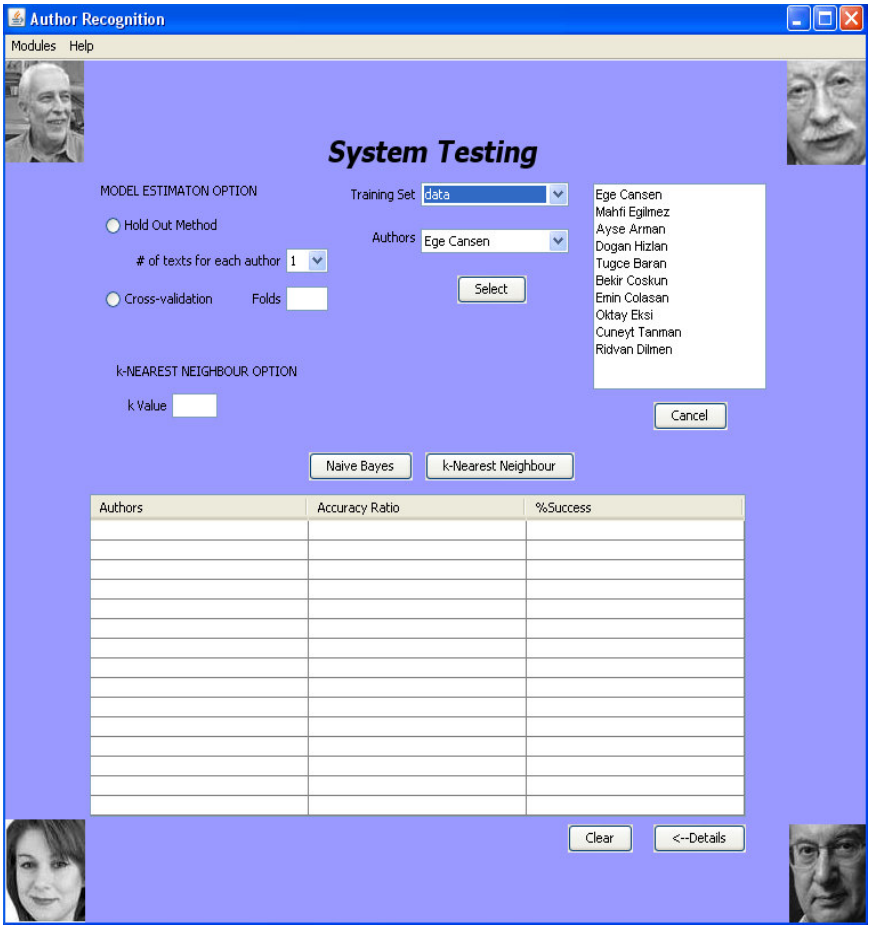
During the implementation of the project, two main tools are used Java programming language and its IDE called NetBeans 6.5. MySQL is chosen as database. Turkish NLP library called Zemberek-2.1.1 is used in this project.

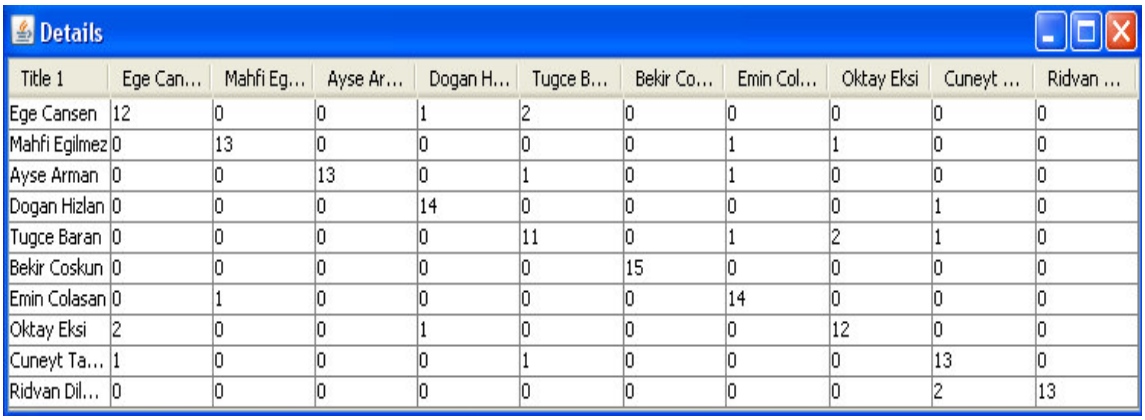
Some of Screenshots






Author Recognition
Modules Help
Naive Bayesian Classifier
Text:
Browse
Training Set: dataset
dataset dataset contains 1000 samples.
dataset dataset is consist of 10 authors.
In this dataset 100 samples exist for each author.
These authors:
Ege Cansen
Ayse Arman
Tugce Baran
Emin Colasan
Cuneyt Tanman
Ege Cansen
Ayse Arman
Dogan Hizlan
Emin Colasan
Cuneyt Tanman
Naive Bayes
Owner of selected Text:
Clear
Details-->
Avg. word length
# of quotations
# of sentences
Word richness
# of commas
# of points
# of colons
# of words
Avg. # of nouns in a sentence
Avg. # of adverb in a sentence
Avg. # of adj. in a sentence
# of exclamation marks
# of semicolons marks
# of different words
# of question marks
# of foreign words
Avg. # of words in a sentence
Avg. # of particle in a sentence
Avg. # of verbs in a sentence
Avg. # of pronoun in a sentence
Avg. # of exclamations in a sentence
Avg. # of conjunctions in a sentence
# of inverted / # of all sentences
# of incomplete / # of all sentences
Select All
Remove All


Author Recognition
Modules Help
System Testing
MODEL ESTIMATON OPTION
Hold Out Method
# of texts for each author 1
Cross-validation
Folds
Training Set data
Authors Ege Cansen
Select
Ege Cansen
Mahfi Egilmez
Ayse Arman
Dogan Hizlan
Tugce Baran
Bekir Coskun
Emin Colasan
Oktay Eksi
Cuneyt Tanman
Ridvan Dilmen
Cancel
k-NEAREST NEIGHBOUR OPTION
k Value
Naive Bayes
k-Nearest Neighbour
Authors
Accuracy Ratio
%Success
Clear
<--Details



Title 1	Ege Can...	Mahfi Eg...	Ayse Ar...	Dogan H...	Tugce B...	Bekir Co...	Emin Col...	Oktay Eksi	Cuneyt ...	Ridvan ...
Ege Cansen	12	0	0	1	2	0	0	0	0	0
Mahfi Egilmez	0	13	0	0	0	0	1	1	0	0
Ayse Arman	0	0	13	0	1	0	1	0	0	0
Dogan Hizlan	0	0	0	14	0	0	0	0	1	0
Tugce Baran	0	0	0	0	11	0	1	2	1	0
Bekir Coskun	0	0	0	0	0	15	0	0	0	0
Emin Colasan	0	1	0	0	0	0	14	0	0	0
Oktay Eksi	2	0	0	1	0	0	0	12	0	0
Cuneyt Ta...	1	0	0	0	1	0	0	0	13	0
Ridvan Dil...	0	0	0	0	0	0	0	0	2	13