SEMANTIC DOCUMENT CLASSIFICATION

Nihan Gümüş-Merve Uğur-Mustafa Güngör
Advisor:Erdoğan Doğdu

Çankaya University, Department of Computer Engineering

Abstract

The web is big and there are billions of text documents that cannot be searched efficiently. Most of the current efforts depend of the word vector representation of text documents, and since this does not involve any semantics, the classification methods are not very accurate. In this project we will investigate existing or currently being developed novel methods in semantic classification of text documents, test and extend them, and also develop new semantic methods to classify documents more accurately and compare all methods extensively.

Introduction

Grouping similar features means that the information obtained is valid for that group. The process of partitioning a set of data objects into subsets is Clustering. Document Clustering has wide application fields. Document clustering is used for data mining, information retrieval, and knowledge discovery from a data of different category. Document clustering is an application that allows you to perform cluster analysis on text documents. Document clustering is an application that allows you to perform cluster analysis on text documents. The document clustering process includes and uses identifiers and conventions of these identifiers. The identifiers of the words in the cluster are recognized.

Solution

We have more than one document. Today, text documents; the internet, e-mail and electronic data base a lot on web pages and are stored in the format. Organize and browse through the papers every day. To overcome this problem are being used more than one method. This paper discusses the methods used. The purpose of a common set of semantic document set that contains a single semantic web documents logically is to categorize the document set. Understanding automatic and smart text due to insufficient billions of text that cannot be efficiently searching document. Vector representation of most current efforts depends on the text document and this does not include any semantics, classification methods are not very accurate. Classification of text documents in this project or still present in developing new semantic method are investigated and dialectics

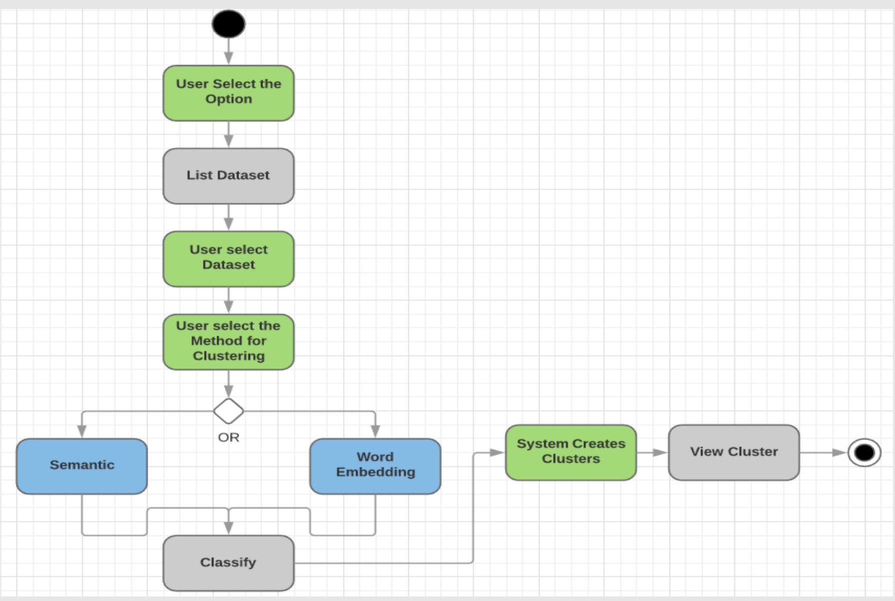
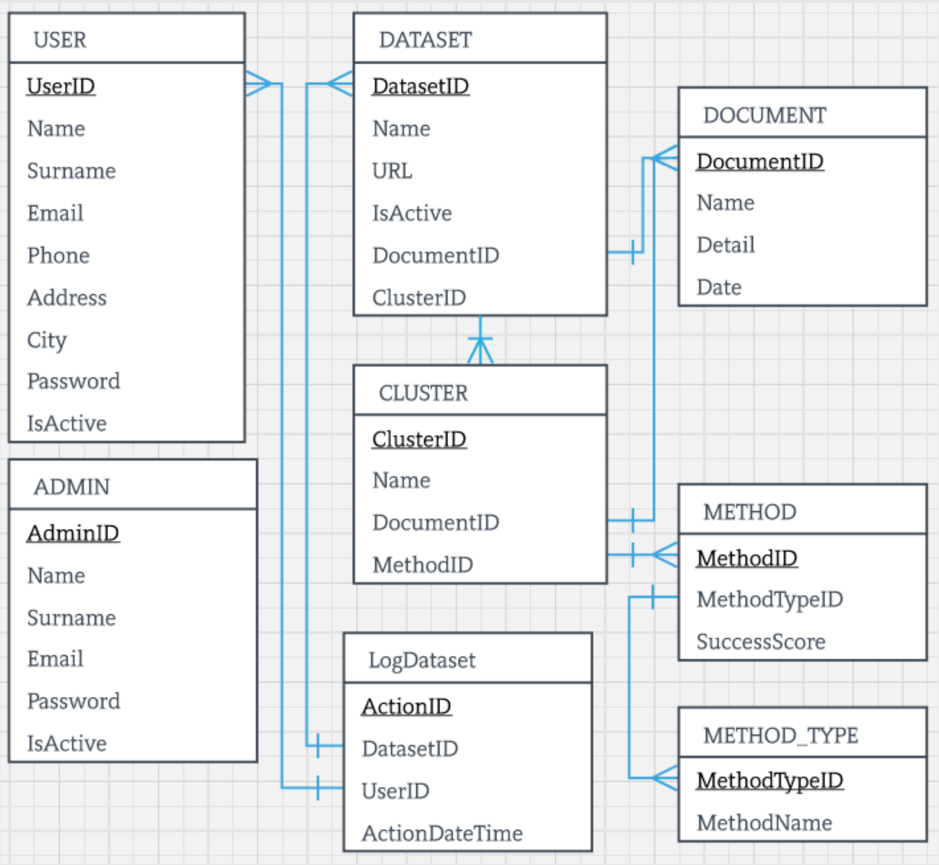


Figure 1 - Flowchart



USER
UserID
Name
Surname
Email
Phone
Address
City
Password
IsActive

ADMIN
AdminID
Name
Surname
Email
Password
IsActive

DATASET
DatasetID
Name
URL
IsActive
DocumentID
ClusterID

CLUSTER
ClusterID
Name
DocumentID
MethodID

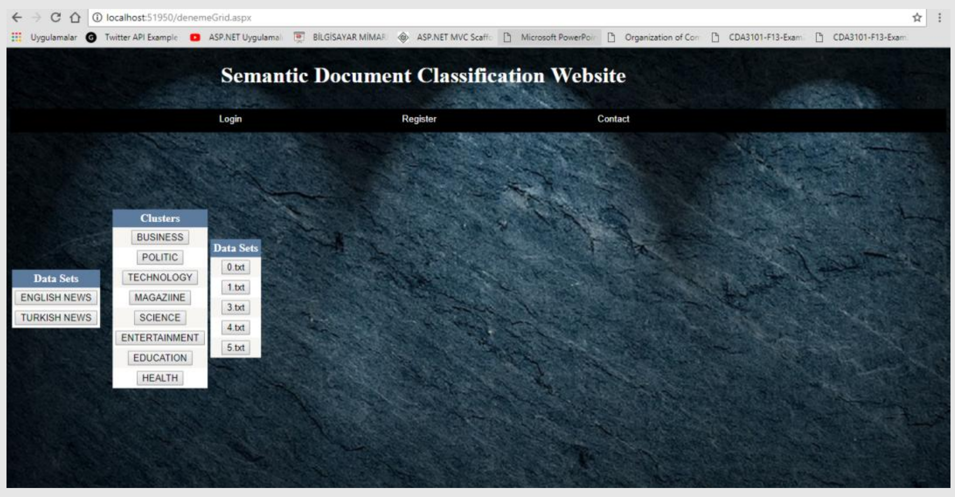
LogDataset
ActionID
DatasetID
UserID
ActionDateTime

DOCUMENT
DocumentID
Name
Detail
Date

METHOD
MethodID
MethodTypeID
SuccessScore

METHOD_TYPE
MethodTypeID
MethodName

Figure 2 – Database



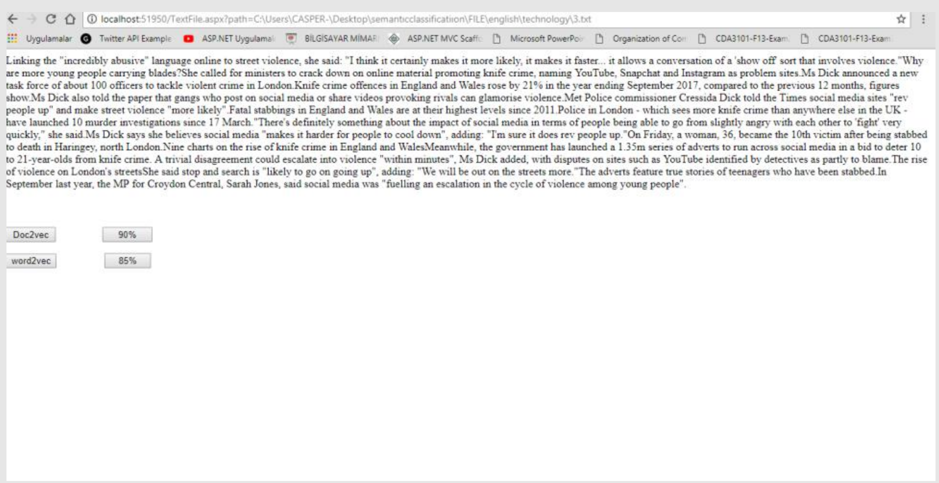
Semantic Document Classification Website

Data Sets
ENGLISH NEWS
TURKISH NEWS

Clusters
BUSINESS
POLITIC
TECHNOLOGY
MAGAZIINE
SCIENCE
ENTERTAINMENT
EDUCATION
HEALTH

Data Sets
0.txt
1.txt
3.txt
4.txt
5.txt

Figure 3 – Finished Product



Linking the "incredibly abusive" language online to street violence, she said: "I think it certainly makes it more likely, it makes it faster... it allows a conversation of a 'show off' sort that involves violence."Why are more young people carrying blades?She called for ministers to crack down on online material promoting knife crime, naming YouTube, Snapchat and Instagram as problem sites.Ms Dick announced a new task force of about 100 officers to tackle violent crime in London.Knife crime offences in England and Wales rose by 21% in the year ending September 2017, compared to the previous 12 months, figures show.Ms Dick also told the paper that gangs who post on social media or share videos provoking rivals can glamorise violence.Met Police commissioner Cressida Dick told the Times social media sites "rev people up" and make street violence "more likely".Fatal stabbings in England and Wales are at their highest levels since 2011.Police in London - which sees more knife crime than anywhere else in the UK - have launched 10 murder investigations since 17 March."There's definitely something about the impact of social media in terms of people being able to go from slightly angry with each other to 'fight' very quickly," she said.Ms Dick says she believes social media "makes it harder for people to cool down", adding: "I'm sure it does rev people up."On Friday, a woman, 36, became the 10th victim after being stabbed to death in Haringey, north London.Nine charts on the rise of knife crime in England and WalesMeanwhile, the government has launched a 1.35m series of adverts to run across social media in a bid to deter 10 to 21-year-olds from knife crime. A trivial disagreement could escalate into violence "within minutes", Ms Dick added, with disputes on sites such as YouTube identified by detectives as partly to blame.The rise of violence on London's streetsShe said stop and search is "likely to go on going up", adding: "We will be out on the streets more."The adverts feature true stories of teenagers who have been stabbed.In September last year, the MP for Croydon Central, Sarah Jones, said social media was "fuelling an escalation in the cycle of violence among young people".

Doc2vec	90%
word2vec	85%

Figure 4 – Finished Product

Results & Conclusion

We have more than one document. Today, text documents; the internet, e-mail and electronic data base a lot on web pages and are stored in the format. Organize and browse through the papers every day. To overcome this problem are being used more than one method. This paper discusses the methods used. The purpose of a common set of semantic document set that contains a single semantic web documents logically is to categorize the document set. Understanding automatic and smart text due to insufficient billions of text that cannot be efficiently searching document. Vector representation of most current efforts depends on the text document and this does not include any semantics, classification methods are not very accurate. Classification of text documents in this project or still present in developing new semantic method are investigated and dialectics.

Acknowledgement

At first we offer our thanks to Erdoğan DOĞDU, our consultant. We thank all our friends who helped us with the feedback. We give thanks to everyone who contributes to the work.