ÇANKAYA UNIVERSITY
FACULTY OF ENGINEERING
COMPUTER ENGINEERING DEPARTMENT

Project Report

Version 2

CENG 408

Innovative System Design and Development I

Semantic Document Classification

Nihan Gümüş
201311023
Merve Uğur
201211051
Mustafa Güngör
201211026

Advisor: *Prof.Dr.Erdoğan Doğdu*

Table of Contents

Abstract

The web is big and there are billions of text documents that cannot be searched efficiently because of the inadequacies of the current methods to classify, relate or understand the text automatically and intelligently. 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. some methods are mentioned below. These are separated into semantic and word embedding. The semantic approach involves linked data. Word embedding approach includes WordNet and DBpedia methods.

Key words:

Information extraction, text clustering, natural language processing, machine learning, semantic

Özet:

Web büyüktür ve metnin sınıflandırılması, ilişkilendirilmesi veya metnin otomatik ve akıllıca anlaşılmasının yetersiz olması nedeniyle verimli bir şekilde araştırılamayan milyarlarca metin dokümanı bulunmaktadır. Mevcut çabaların çoğu, metin belgelerinin vektör gösterimine bağlıdır ve bu herhangi bir semantik içermez, sınıflandırma yöntemleri çok doğru değildir. Bu projede, metin belgelerinin semantik sınıflandırılmasında mevcut veya halen gelişmekte olan yeni yöntemleri inceleyeceğiz, bunları test edecek ve genişlettireceğiz ve ayrıca belgeleri daha doğru sınıflandırmak ve tüm yöntemleri kapsamlı bir şekilde karşılaştırmak için yeni semantik yöntemler geliştireceğiz. bazı yöntemler aşağıda belirtilmiştir. Bunlar semantik ve kelime yerleştirmeye ayrılmıştır. Semantik yaklaşım, bağlantılı verileri içerir. Kelime gömme yaklaşımı, WordNet ve DBpedia yöntemlerini içerir.

Anahtar Kelimeler:

Bilgi çıkarma, metin kümeleme, doğal dil işleme, makine öğrenimi, anlamsal

1. Introduction

1.1 Problem Statement

There are billions of text documents that can not be searched efficiently. The purpose of a semantic document classification is to logically categorize web documents containing a common semantic classification into a single semantic document classification. Earlier methods to classify text documents utilized different “word embedding” techniques. These techniques depend on the syntactic similarities of words in the documents. This does not work in many cases. For example a document talking about the “president” of a company and another document talking about the “president” of a country may be classified as similar incorrectly. Or a document talking about Trump and another document talking about Putin should be classified as similar (talking about presidents of countries), but they are not, due to word representations does not capture the semantic similarities.

1.2 Solution Statement

We can use the Semantic Document Classification for these problem solving. Semantic relationships between words are considered. The higher the number of words and phrases with a similar meaning, the cluster grows. There are classifications for documents to be installed on the system. In this system, the user has options such as uploading documents to the system, searching, choosing the classification method, seeing the success rates of the classification methods, and testing. We will investigate the most recent methods developed for semantic document classification, then develop and compare them, and finally extend and develop new methods that can perform better than the existing ones.

1.3 Contribution

We are a group of students in computer engineering department who are interested in Semantic Web. Now, the data become “big data”. The big data learning and the data reasoning are very important for the future success of the Internet of Things (IoT). As a group, we researched Semantic web for a better understanding the most recent protocols, languages, technologies and tools that are used to build the Semantic Web. Also We have reviewed “big data” studies. We aimed to combine the Semantic Web technologies in this project. We have chosen the python language that interpreted, object-oriented, high-level programming language with dynamic semantics to develop our project. For including

Semantic Web technologies, we have acquired and we have read documents for how to design Semantic Document Clustering.

2. Literature Search

The purpose of a semantic document classification is to logically categorize web documents containing a common semantic classification into a single semantic document classification. News aggregation is particularly important for applications such as Web search engines. In recent years news and media organs have gone to the Web, and Web-based news distribution is growing rapidly. clustering of classification or news articles is therefore important for many people. Recent studies in text clustering, external knowledge bases such as WordNet, to understand concepts for documents and / or to identify named entities began to use as reference, [11], [12], [13]. In this approach, the documents are searched for and used in specific assets or sources of external information banks, documents are presented and then considered as documents documents, with regard to law same or similar personalities and relationships included, are measured to a number of metrics. “Linked Data” is therefore a first type of external reference source. Linked data was first defined by Tim Berners-Lee and introduced “structured data published on the Web and best practices for publication” [14], [15].

Recent work in the text clusters has begun to focus on the use of external knowledge bases such as WordNet or DBpedia. WordNet is a lexical database of English words grouped in synonymic concept sets [16]. These clusters are connected by meaningful phrase links that help to find similarities between documents that use different terminologies but use similar meanings. DBpedia is one of the largest linked data sources in the Linked Open Data (LOD) cloud. It is extracted from Wikimedia data and used in various studies [18].

2.1. Document Clustering

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. Document clustering is also thought of as a bag of document words. Semantical meanings of the words are ignored. The meaning of the words is important to cluster the right document. Document clustering can be done by semantic approach because it looks at the semantic relation between words.

2.2 Word Embedding

Earlier methods to classify text documents utilized different “word embedding” techniques. These techniques depend on the syntactic similarities of words in the documents. This does not work in many cases. For example a document talking about the “president” of a company and another document talking about the “president” of a country may be classified as similar incorrectly. Or a document talking about Trump and another document talking about Putin should be classified as similar (talking about presidents of countries), but they are not, due to word representations does not capture the semantic similarities.

2.3 Semantic Document Clustering

Text document clustering or clustering, defining content or text documents by subject, and it is the duty of the grouping. This is plenty of text documents and continued to increase constantly growing is particularly important for the Web. According to the article the article clustering documents subject categories, which is a special task in this field. This problem has been before to solutions, “bag of words” approach on the words and documents; documents used frequency is represented and clustering with task, it measures the similarity of documents using a representation. However, this approach consider the importance and the meaning of words and the uncertainty of words in çözümlenmemektedir. Documents for information to represent bases, especially “linked data” using the document or news articles to cluster new and offer an approach. The words or phrases in the documentation for the article, “linked data” we match real-world counterparts inside. Information such as DBpedia bases and they have linked data entities represent the document. Then, the semantic similarity between documents by using a method that entities and relationships by calculating documents using group together. Early evaluation results, perform much better with regard to Word package approach.

Word Embedding-based Document Clustering

Placing a word is a learned representation for text that has a similar representation of words with the same meaning. It is this approach to represent words and documents that are considered one of the cornerstones of in-depth learning on difficult natural language processing problems. Vocabulary assignments are in fact a technical class in which individual words are represented as real valued vectors in a predefined vector space. Each word is associated with a vector and the vector values are learned in a neural network-like manner, so that the technique is usually incorporated into the field of deep learning. The key to the approach is the idea of using dense distributed representations in every word. Each word is represented by a real valued vector. This is opposite to the size required for infrequent word representations. Distributed representation is learned based on the use of words. This causes similarly used words to have similar representations that naturally capture their meanings. Unless explicitly governed, it can be compared to a crisp but fragile display in a packed word pattern where different words have different expressions, regardless of how they are used. [20] This vector representation has two important and advantageous features:

1. Dimension Reduction - a more efficient representation
2. Contextual similarity - more expressive representation

Word Embedding is used for meaning semantic separation to make meaning from the text to provide natural language understanding. In order for a language model to predict the meaning of the text, it must be aware of the contextual similarities of the words.

Word2Vec: Word2Vec is a statistical method for efficiently learning an independent vein buried in a text corpus. Word2vec is a embedding method that enters words as an input. It produces vectors as output. In addition, the study includes the analysis of learned vectors and the investigation of vector maths on their presentations. Two different learning models have been introduced that can be used as part of the Word2Vec approach to learning word embedding. CBOW model: The CBOW model learns to predict and submit the present word based on its context. Skip-gram model: The continuous skip-gram model learns by predicting the words in the given language.

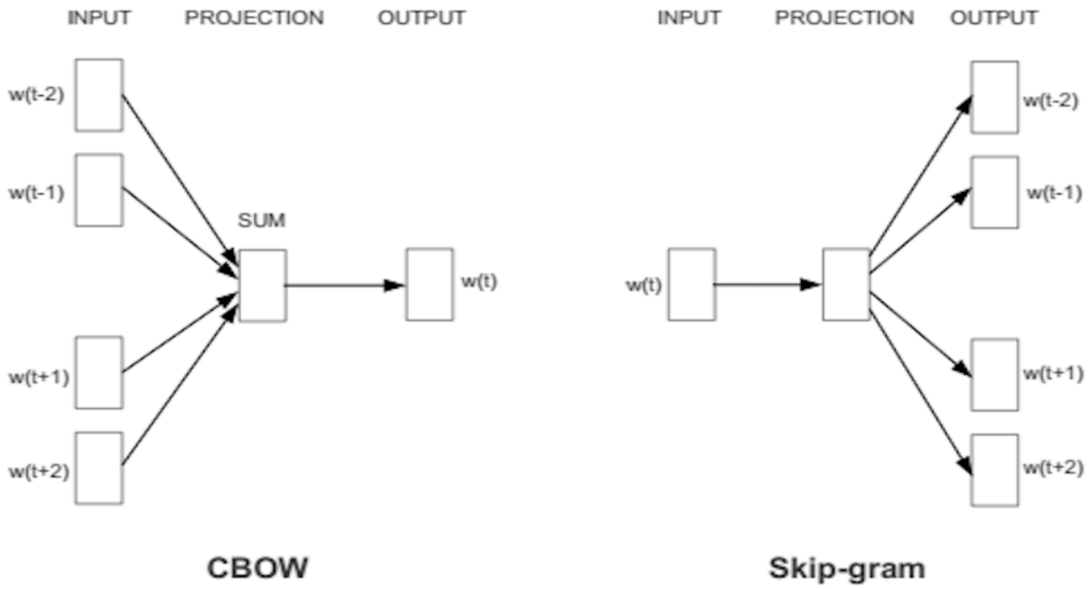


Figure1. Vektor Representation

Both models focus on learning the words given in the context of local use, which is defined by a window of contexts of neighboring words. This window is a configurable parameter of the model. The most important benefit of the approach is that it allows high-quality word placements to be learned efficiently, allowing larger burieds to be learned in larger texts(Figure1).

Doc2vec: Doc2vec is the extension of word2vec to learn about document placements. There are two approaches to this. Doc2vec: Dbow and Dmpv. Dbow works in the same way as skip-gram. While Dbow is a simpler model and ignores word order, dmpv is a more complex model with more parameters. It has been found that Dmvp is a better model as an independent method and others give conflicting results. It has been found that Doc2vec is performing well when using models trained in large external companies and can be further improved using pre-trained word placements. Pre-educated word placementsand hyper-parameter setting, Doc2vec has performed a very strong performance compared to the latest technology document placement approaches, as well as a simple word that embeds both the average and n-gram baseline, and that Doc2vec is particularly powerful was found to perform well.[19] In general, we have seen that Doc2vec performs well and that Dbow is a better model than Dmpv. We found that even when Doc2vec is trained using large external corpus, it performs well and benefits from pre-trained word placements.

2.2.1 WordNet

Ontologies such as WordNet have been used to improve clustering and semantic relationships. WordNet-based clustering methods are often based on the analysis of single-word texts. They do not do sentence-based analysis [2]. Some methods rely on NER-based entity matching techniques for word relationships (such as WorNet associations) or semantic document classification. In previous methods, various methods of “embedding words” were used to classify text documents.

Semantic document similarity-based classification is a technique in intelligent document classification. Some methods utilize word relationships (like WordNet relationships), or named-entity recognition (NER)-based entity matching techniques for semantic document classification. Many methods for considering the semantics of the document use WordNet synonymously. However, WordNet-based clustering methods are based solely on one-time text analysis, not sentence-based analysis. In addition, these methods use synonyms to define concepts, to calculate concept frequencies, and to discover only hypernymia; they do not think about other semantic relationships. To solve these problems [3], authors have combined the perception of namespaces and WorNet. This integration helps to discover documents more semantically for cluster purposes [4]. in Figure1, shows a part of such a hierarchical semantic knowledge base [6]. Document clustering deals with the problem of identifying groups of similar documents without human supervision. Unlike most existing solutions that perform document clustering based on keyword matching, we propose an algorithm that accounts for the meaning of the terms in documents. For example, a document containing multiple words such as “dog” and “cat” may be placed in the same categorization as a document containing the word “pet”, but both documents contain only common noise words. Our semantic clustering algorithm is based on a similarity graph that stores the degree of semantic relation (extracted from WordNet) between expressions in which a term can be a word or a sentence[17].

Semantic Document Clustering

Numerous sources, medical reports, economic analysis, scientific journals, news, blogs, etc. It generates valuable information in the text. It is very difficult to maintain and access these documents without proper classification. These problems can be overcome by appropriate document classification. Only a few documents have been classified. All of them have a needs classification and they are uncontrolled. In this context, clustering is the only solution. Traditional clustering techniques and text clustering have some differences. Relationships between words are very important for the formation of clusters. Semantic clustering has proven to be a more appropriate clustering technique for texts [7].

Document clustering and subject modeling are two interrelated tasks that can be mutually beneficial. Topic modeling can reflect documents into a subject area that facilitates effective document clustering. Cluster tags discovered with the document cluster can be included in topic models to extract local topics specific to each common and common topics shared by all clusters. we propose a multi-grain clustering topic model (MGCTM) that integrates document clustering and subject modeling into a unified framework and performs two tasks in common to achieve the best overall performance. Our model often combines two components: a mix component used to discover hidden groups in the document aggregate, and a topic model component used for mining multi-part topics, including common topics shared between each cluster-specific local topics and clusters. We use approximate variational inference and learn the posteriori and model parameters of hidden variables. The experiments in the two data sets show the effectiveness of our model.

2.3. 1. Problems Of Semantic Document Classification

Semantic document classification is a difficult problem due to the wide variety of documents on the web, and the difficulty in intelligently understanding and semantically classifying documents. This problem requires solving many issues in utilizing natural language processing techniques in cleaning and structuring documents, finding named- entities, word relationships and so on. There is a need to utilize many methods in a hybrid manner as well. Testing is also difficult due to the lack of good datasets. In this approach, semantic relationships between words are considered. Documents that are semantically related are grouped in the same group, and documents that are not semantically related are grouped into another group. The semantic approach can also help to determine the connotation of a community. The semantic approach focuses on the meanings of words, and thus the semantic approach generally uses dictionaries to find relationships between terms to generate keywords [1].

2.3.1.1. Linked Data

The connected data correlates the data and removes the existing interference between the data. Semantic questions can be answered by giving meaning to data in Linked Data. With Linked Data, data is shared in a format that computers can automatically understand from texts that people can understand [8].

There are general web-based browsers (eg Disco, Marbles, Tab, and Zitgist) for various linked data, and various data warehouses work with data associated with arbitrary domain names (eg FactForge, Falcons, Sindice, Sig.ma, Swoogle, SWSE, and Watson), but a real scanner has not been developed yet [9]. Connected data is an approach to data integration. It uses ontologies, terminologies, URIs, and RDF to access information about a piece of semantic and semantic web. Linked records are published because they provide the following potential use cases and the like [10].

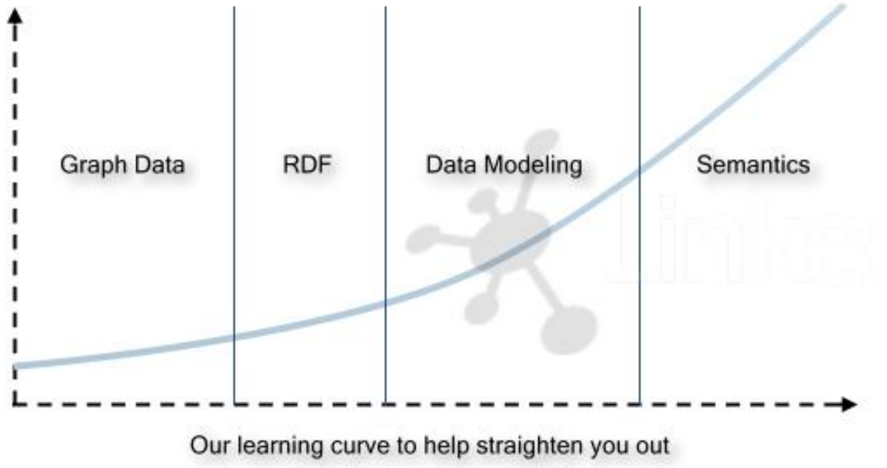


Figure2. Semantic Learning Curve

1. Shareability: A data recorder or publisher standards, a program interface such as SPARQL or resolvable URIs facilitate access to existing data. A scientist wants to have priority access to the database in the research network.
2. Integration: A developer wants to be able to create and update a list of links between different RDF datasets and easily query between these datasets.
3. Semantic Normalization: A computer scientist deals with indexing an existing RDF dataset using a set of common ontologies. The data set is then queried using ontological terms.

4. Discoverability: A biologist would like to find out what is appropriate both as published results, as raw data, and as tissue libraries in a semantic web that relates to a variety of proteins, genes, or chemical components.
5. Association: A pharmaceutical company wants to use SPARQL to obtain data from sources scattered from one end of the enterprise to another.

This section describes the methodology used in this study and the test environment described in detail. Our approach includes the following steps:

1. Named entities, types, linked data entries (in this case links to DBpedia).
2. using linked data categories for in DBpedia's assets expand hierarchically.
3. each document vector of of words, represented as entities, linked data entities.
4. calculate the similarity between Entities.
5. Being the distance between documents using the similarity matrix math.
6. distance (similarity) apply hierarchical clustering and cluster news matrix documentation.

The proposed method are summarized in Algorithm 1. For a given set D, D before each document in Alchemy are described with API6. Defines the named entities in the document description process and, at the same time, the corresponding DBpedia has a corresponding element in the linked data for organizations (DBpedia) URIs. The result is given in an XML document per document request. Algorithm, then to find the linked data URIs parses XML documents. Each URI then kolette: broader and dc: DBpedia DataSet by using the terms predicate types (an ending point 7) 5 level hierarchically by querying is expanded.

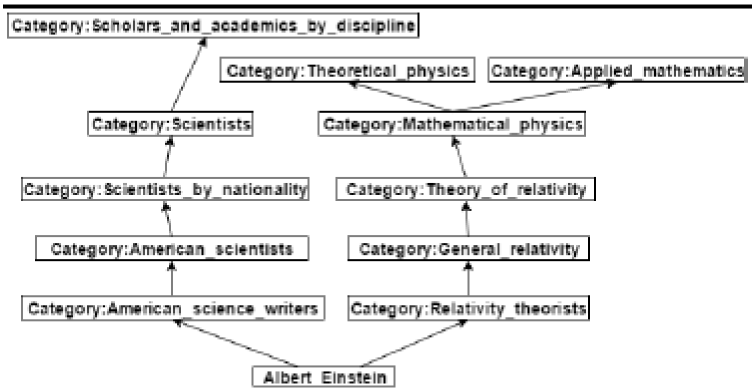

Category:Scholars_and_academics_by_discipline
Category:Theoretical_physics
Category:Applied_mathematics
Category:Scientists
Category:Mathematical_physics
Category:Scientists_by_nationality
Category:Theory_of_relativity
Category:American_scientists
Category:General_relativity
Category:American_science_writers
Category:Relativity_theorists
Albert_Einstein

Figure3. Albert Einstein's category

This expansion is an example, in Figure3, Albert Einstein's category 5 level categories. Were represented in DBpedia.

The Semantic Web should embrace these issues and Linked Data researchers. On the contrary, even in such cases we will need to find good performing technical and methodological solutions. And at the same time, we need to begin to cultivate the best practices that will ease some problems in data publishing and facilitate the development of technical solutions. Moreover, previously described then argumentation, linked data sets Semantic Web and overall computer / information to communicate in an understandable way for researchers and practitioners outside of science, it is clear that we need a way to images and documents.

Summary

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. Then expanding them on a test and classify documents more accurately and in a comprehensive manner all the methods to develop new semantic methods compare. As mentioned earlier, the above method to classify documents using a software tool.

3.Software Requirements Specification

3.1 Introduction

3.1.1 Purpose

This document describes the Semantic Document Classification system. The purpose of a semantic document classification is to logically categorize web documents containing a common semantic classification into a single semantic document classification. There are classifications for documents to be installed on the system. In this system, the user has options such as uploading documents to the system, searching, choosing the classification method, seeing the success rates of the classification methods, and testing. This document explains information on the following subjects. the topics covered include the details of the work to be done. The SRS document describes the functions and activities of people who use the semantic document classification system. This document describes how stakeholders' needs are addressed.

3.1.2 Scope of Project

There are billions of text documents that cannot be searched efficiently because of the inadequacies of the current methods to classify, relate or understand the text automatically and intelligently. Users want to see how the documents they upload to the system are classified. They may want to see it in addition to the previous data. A system was developed to classify documents in an easy and understandable way so that millions of document problems that can not be non-searchable could be addressed.

There are two user entries in the semantic document clustering system. One of them is the user who logs in to the system, the other is the administrator who controls the system and the users. There are different options for users and admin. Admin can access the information of the users entering the system, access the uploaded file, add and delete files, edit the entered data. Users can view existing documents, add new documents, choose a method to classify documents, test documents according to methods, see success rates, search documents. However, users can search for documents by word-based searching.

3.1.3 Glossary

Term	Definition
Participant	The user who interact with the documents classified by the semantic approach.
Stakeholders	Any person who has contribution in the project.
Web Site Environment	The loaded documents, the results of the tested data, and the environment in which access to the classified data is provided.
Administrator	Admin can access the information of the users entering the system, access the uploaded file, add and delete files, edit the entered data.

3.1.4 Overview of Document

The second part of the document describes functionalities of Semantic Document Classification project. Informal requirements are described and it is a context for technical requirement specification in the Requirement Specification part. Requirement Specification part is written for software developers and details of the functionality of the project are described in technical terms.

3.2 Overall Description

3.2.1 Product Perspective

The purpose of a semantic document classification is to logically categorize web documents containing a common semantic classification into a single semantic document classification. There are classifications for documents to be installed on the system. In this system, the user has options such as uploading documents to the system, searching, choosing the classification method, seeing the success rates of the classification methods, and testing. Participant choose the one method in the methods for clustering document. These methods semantic clustering or word embedding.

3.2.2 Development Methodology

For developing the project, we have planned to use Scrum which is an agile software development methodology. Thanks to the Agile method, project productivity, ability of the project to adapt to the variables in a fast way, project quality, excessive interaction among the team members and the delivery speed of the project increases. At the first appearance Scrum is an administrative model which has simple rules. It is applied for software projects which are open to changes and in which the requirements are not clearly specified.

Scrum includes Daily meetings in its scale and these meetings takes maxiumum 15 minutes. Project process are divided into sprints and these sprints has to be at fixed length. Scrum demonstrates the progress of the project in a clear and continuous way. At the end of each sprint some part of the project is completed and delivered to the client. The most significant advantage is the short sprints and dealing with the changes thanks to the feedbacks in an efficient way.Because of these, Scrum is going to be suitable for our project.

3.2.3 User Characteristics

3.2.3.1 Participants

- Participant must connect to Internet.
- Participant must read and understand English language.

3.2.3.2. Admin

- Admin must read and understand English language.
- Admin must connect to Internet.
- Admin must know the system because he/she can change the information or do edit in the document.
- Admin must know the semantic document clustering method and word embedding method for classification. If the admin don’t know these methods, he/she can’t does the edit.

3.3 Requirements Specification

3.3.1 External Interface Requirements

3.3.1.1 User interfaces

The user interface will be worked on web application.

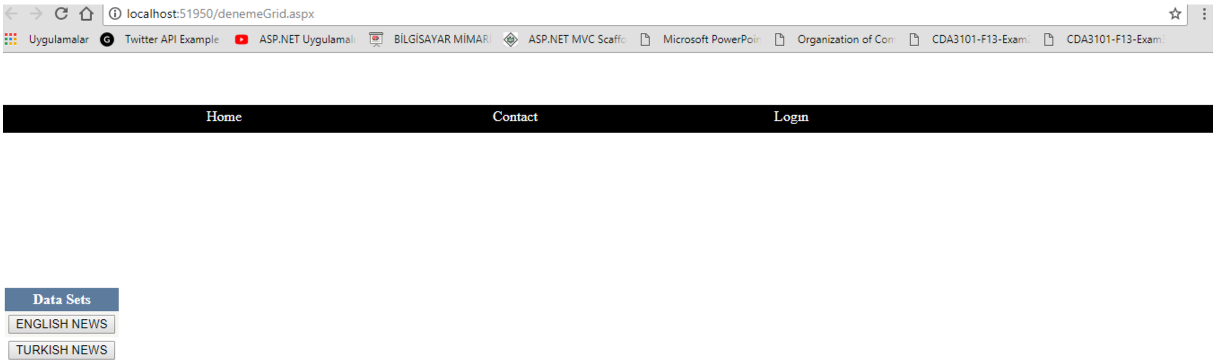


Figure 1. Homepage Interface

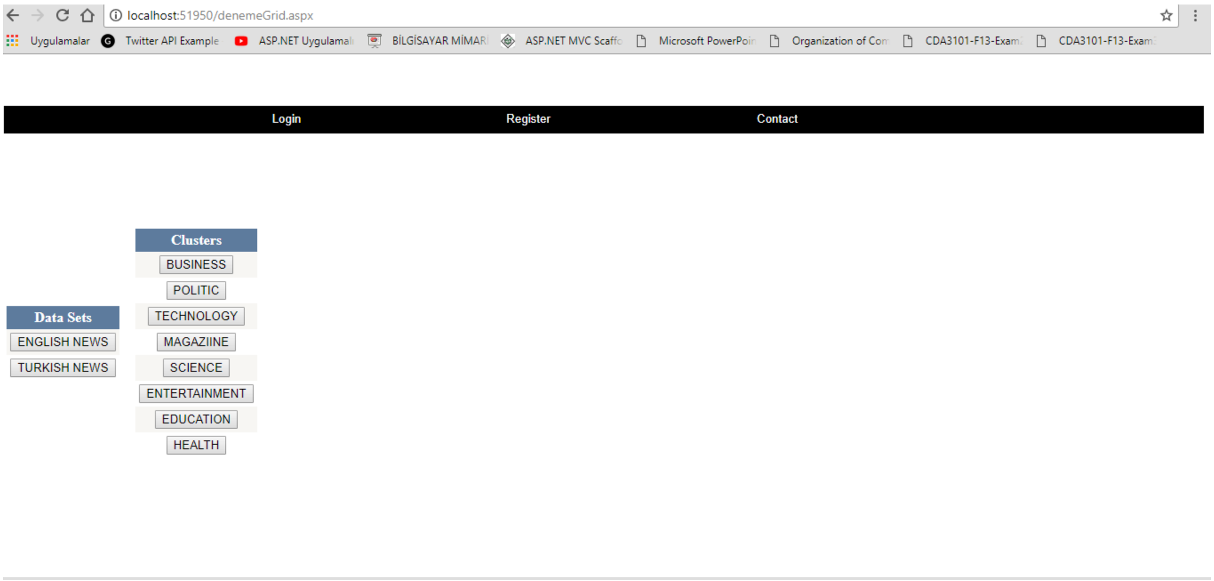


Figure 2. Click the Dataset Button

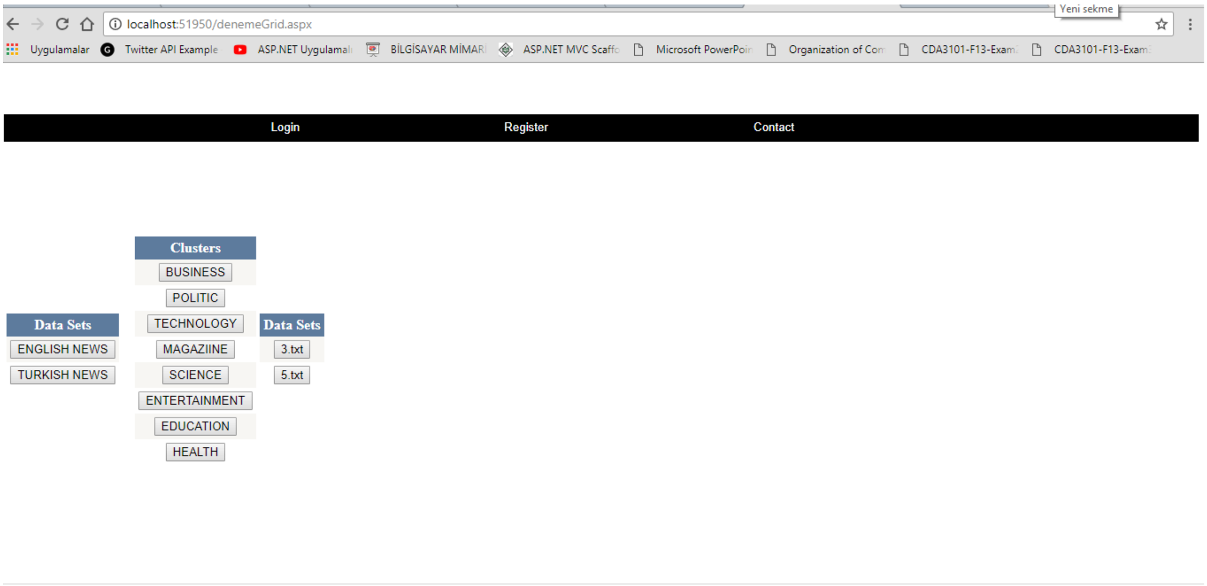
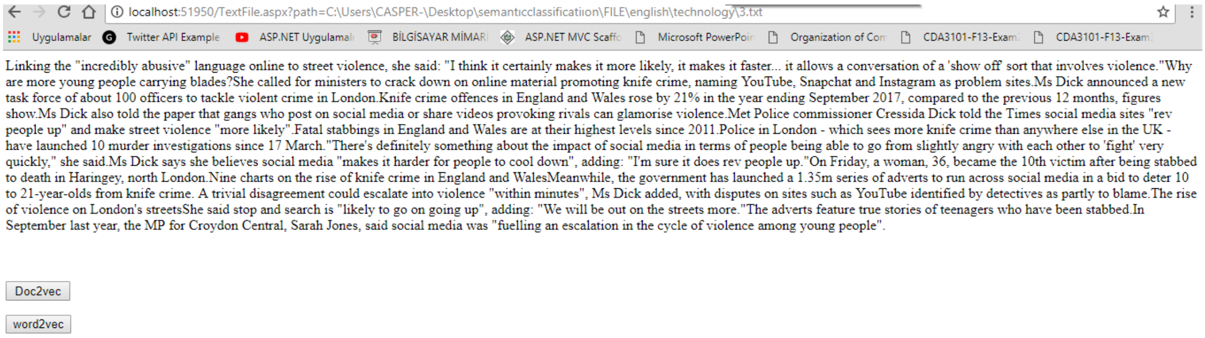
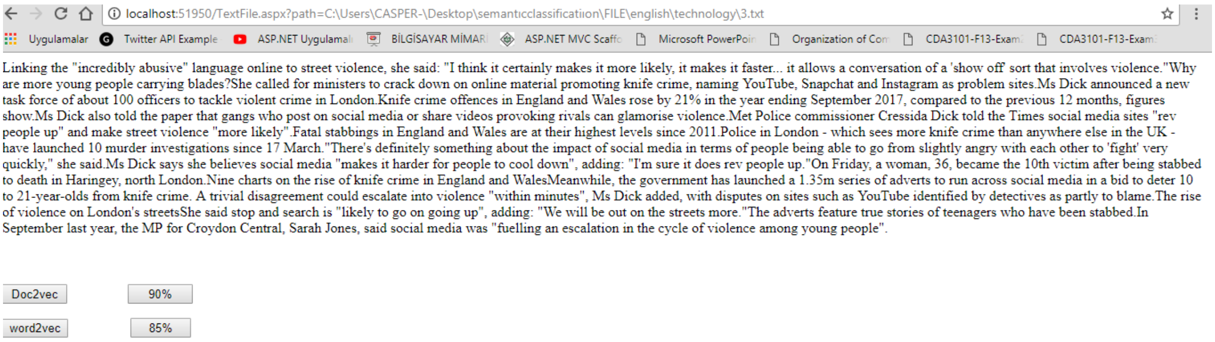


Figure 3. Click the Cluster Button



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".

Figure 4. Click the Do test Button



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".

Figure 5. Click the See Success Score Button

3.3.1.2 Hardware interfaces

There is no external hardware interface requirement.

3.3.1.3 Software interfaces

There is no external software interface requirement.

3.3.1.4 Communications interfaces

There is no external communication interface requirement.

3.3.2 Functional Requirements

USE CASE DIAGRAM:

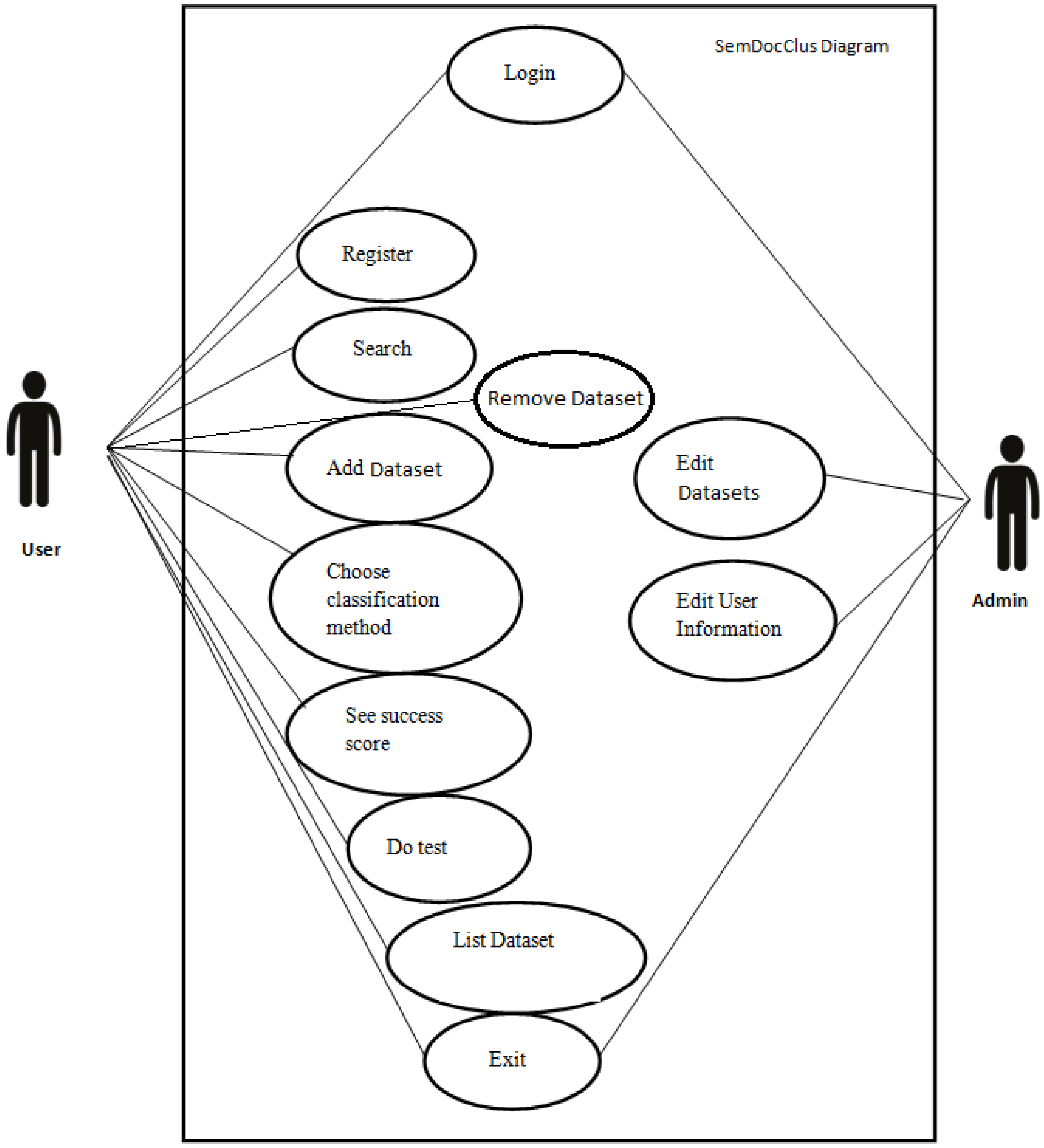
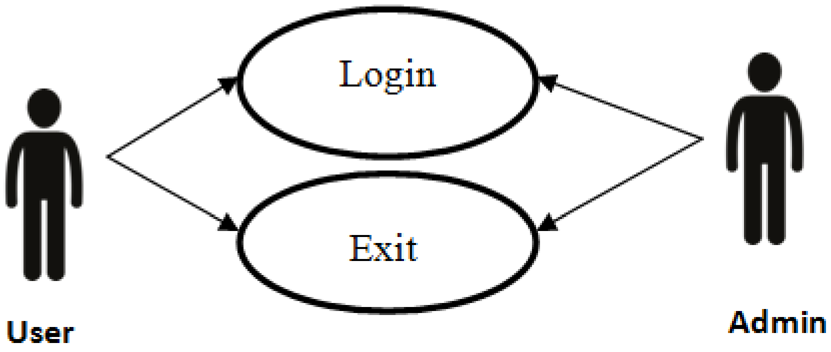


Figure7.Use case Diagram

In *Figure7*,There are two user entries in the semantic document clustering system. One of them is the user who logs in to the system, the other is the administrator who controls the system and the users. There are different options for users and admin.

Use Case: Login and Exit

Diagram:



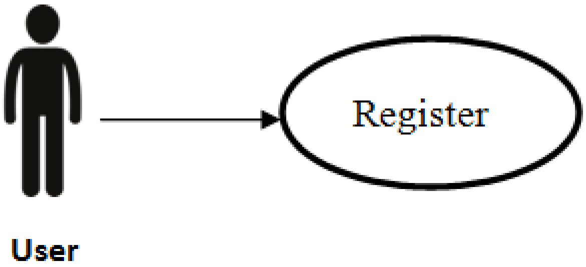
Briefly Description: Describes the action that are admin and users who wish the use the system can performed. It can use both admin and user Login and Exit function. In addition, the administrator can use the login function of the diagram as admin.

Inıtial Step-By-Step Description

1. User should start with Login.
2. User mst Login using email and password.
3. Admin must Login using email and password.
4. Admin and user can Exit from system.

Use Case: Register

Diagram:



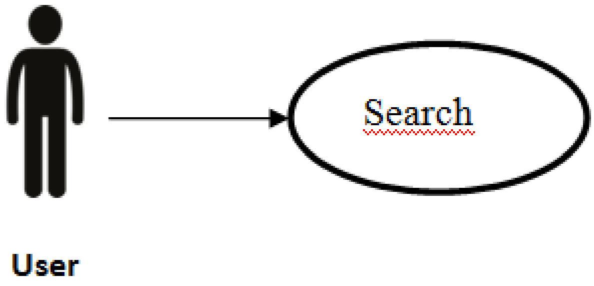
Briefly Description: Contains operations that the user of the system can perform.

Inıtial Step-By-Step Description

1. The user indicates the personal information like as name, surname, email, password, city, phone number.
2. The user clicks request new account button.
3. The system sends on email for validation to user.
4. The user click validation link.
5. The user access to web system.

Use Case: Search

Diagram:



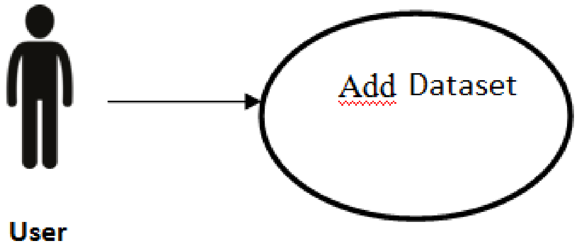
Briefly Description: The user after selecting search buttoni user can searching by keywords. If the user click the search button, he/she can access all document that contains searched keyword.

Inıtial Step-By-Step Description

1. The user select the search button to search for words.
2. After clicking on the button, the user will see all relevant document.
3. The user sees the words he/she searched as highlighted in the document.

Use Case: Add Dataset

Diagram:



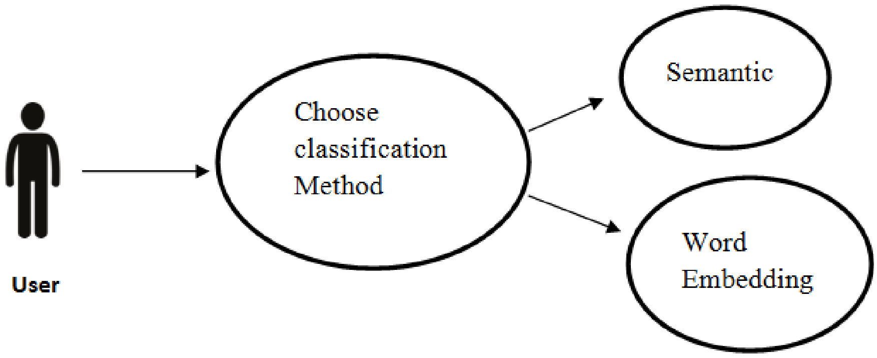
Briefly Description: This function allows users to add dataset. In this way, user can access all datasets that are not classified or want to be classified.

Inıtial Step-By-Step Description

1. The user click the Add Dataset button for do testing or training with the other datasets.
2. User can select the Dataset.
3. User can load the Document in Cluster.
4. User can upload the Dataset into the Datasets.

Use Case: Choose Classification Method

Diagram:



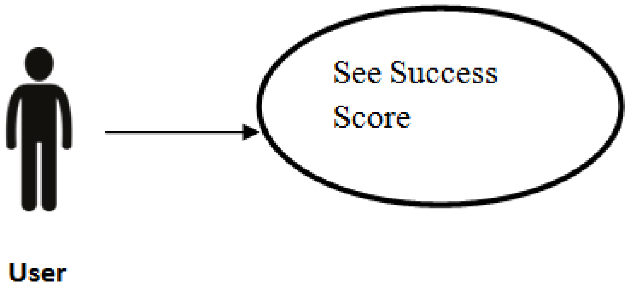
Briefly Description: When the user clicks on this button, two options will be available. The first option Semantic is Word Embedding. With these methods, they will be able to classify the documents they upload.

Inıtial Step-By-Step Description

1. User Click the button and then see the two options.
2. If the user choose the Semantic option, documents cluster using Semantic method.
3. If the user choose the Word embedding option, documents cluster using Word embedding method.
4. The system cluster the added document.

Use Case: See Success Score

Diagram:



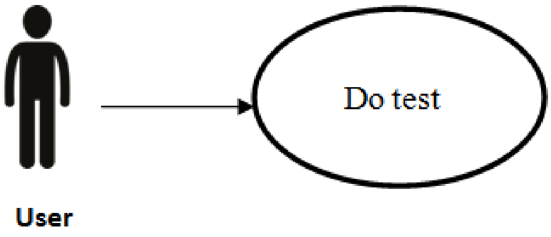
Briefly Description: As a result, the user has used the method of classification which made her success will be able to see the percent.

Inıtial Step-By-Step Description

1. User clicks “see succes score” button.
2. Two methods to see them to the user; semantic and word embedding methods.
3. The classification made by the method in which the user selects see the percentage of success.

Use Case: Do Test

Diagram:



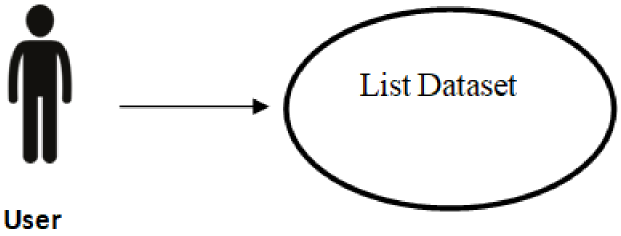
Briefly Description: User classification do here how accurate tests whether.Success in all methods of classification sees the percentages.

Inıtial Step-By-Step Description

1. User clicks the “do test” button.
2. When the user loads a new document, it compares it with other classifications and tests the success.
3. If the user has not loaded a new document, it tests the classification success by comparing the classifications made on the site.

Use Case: List Dataset

Diagram:



Briefly Description:

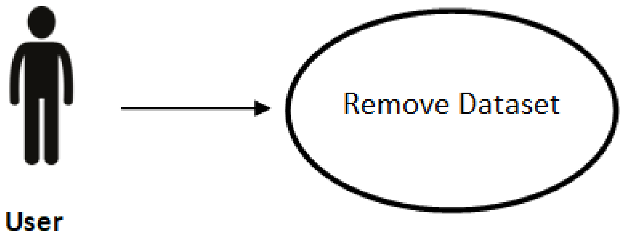
The system offers one options for the user. This option is List Datasets. If the user chooses which one to follow, the system follows step by step and reaches the relevant part.

Inıtial Step-By-Step Description

1. The user clicks the “List Dataset” button.
2. If user wants to add dataset, he can add whole folder to dataset part.
3. If the user wants to see the clusters of the datasets, he can access the clusters after the Dataset button.
4. If the user click the cluster button, user can access the short file version but, if the user clicks on the file under the cluster, the user can see the entire document in the cluster.

Use Case: Remove Dataset

Diagram:



Briefly Description:

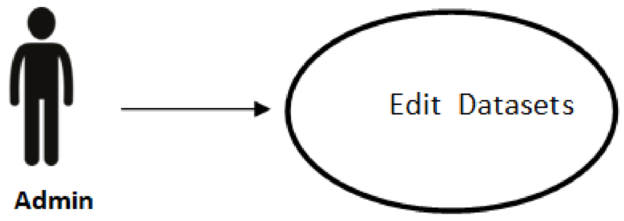
The system offers one options for the user. This option is Remove Datasets. If the user chooses which one to follow, the system follows step by step and reaches the relevant part.

Inıtial Step-By-Step Description

1. The user clicks the “Remove Dataset” button.
2. If user, who added dataset, wants to remove dataset, user remove only own added folder from dataset.
3. When the user lists the datasets, they will no longer see the old datasets.

Use Case: Edit Datasets

Diagram:



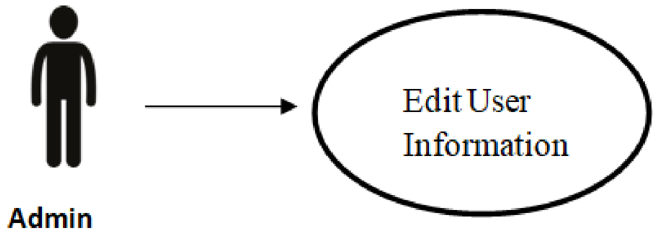
Briefly Description:: The admin here can do what they want to play on all the documents within the site deletes adds new documents it wants.

Inıtial Step-By-Step Description

1. Admin clicks “Edit Dataset” button.
2. Admin can view the documents in dataset on the site and you can add new dataset.
3. The admin can delete all dataset in the website.

Use Case: Edit User Information

Diagram:



Briefly Description: Here all users logged in on the site administrator documentation. How can I see which users have used the method.

Inıtial Step-By-Step Description

1. Admin click “Edit user information” button.

2. The admin can access the information of the users who use the website.
3. The admin can see users which method with classifications they did.
4. If the user enters incorrect information, the user can reach the admin by mail. The admin can make changes to the user information.

3.4 Performance Requirements

- The login information shall be confirmed within a second.
- The app./system should be fast so every developer care the performance of the system and also it should work in a different operating system.

3.5 Design constraints

- Semantic Document Clustering system shall be web application. The system’s user interface should be simple for using because if the system to be simple, users can use and understand the system easily. Moreover, developer should implement simple design because the appearance of the system should be simple.

3.6 Software system attributes

3.6.1 Adaptability

- The application should be adaptable in different web browser.

3.6.2 Usability

- The applcation is intended to be used by people of all ages with a beautiful interface.

3.6.3 Security

- Each user will have their own password that are linked to the mails. For this reason, the system will give an error message when an unexpected condition is encountered in a system. Such that no unauthorized user can access unprivileged pages.

3.6.4 Maintainability

- The system must be easy to maintain.

4. Software Design Description

4.1 Introduction

4.1.1 Purpose

The purpose of this Software Design Document is providing the details of project titled as “Semantic Document Classification”.

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 process deals with grouping of an unclassified collection of documents into semantically or word embedding groups. The document clustering process includes and uses identifiers and conventions of these identifiers. The identifiers of the words in the cluster are recognized. Document clustering is also thought of as a bag of document words. Semantical meanings of the words are ignored.

Millions of complex documents are used for classification. One of these classification types is Semantic Document Clustering. The meaning of the words is important to cluster the right document. Semantic Document clustering allows users to discover the meanings of the sentences to enrich the experience on the document. Semantic Clustering groups semantically equivalent search queries word,sentences into clusters based on meaning. The semantic document classification is to logically categorize web documents containing a common semantic classification into a single semantic document classification. The higher the number of words and phrases with a similar meaning, the cluster grows. There are classifications for documents to be installed on the system. In this system, the user has options such as uploading documents to the system, searching, choosing the classification method, seeing the success rates of the classification methods, and testing. Text document clustering or clustering, defining content or text documents by subject, and it is the duty of the grouping.

This is plenty of text documents and continued to increase constantly growing is particularly important for the Web.[21]

In order to provide a better comprehension, this SDD includes various diagrams such as UML diagram of the project, activity diagram and block diagram.

4.1.2 Scope

This document contains a complete description of the design of Semantic Document Classification. There are billions of text documents that cannot be searched efficiently because of the inadequacies of the current methods to classify, relate or understand the text automatically and intelligently. Users want to see how the documents they upload to the system are classified. They may want to see it in addition to the previous data. A system was developed to classify documents in an easy and understandable way so that millions of document problems that can not be non-searchable could be addressed.

There are two user entries in the semantic document clustering system. One of them is the user who logs in to the system, the other is the administrator who controls the system and the users. There are different options for users and admin. Admin can access the information of the users entering the system, access the uploaded file, add and delete files, edit the entered data. Users can view existing documents, add new documents, choose a method to classify documents, test documents according to methods, see success rates, search documents. However, users can search for documents by word-based searching.

4.1.3 Glossary

Term	Definition
BLOCK DIAGRAM	The type of schema which the components in the system are displayed in blocks.
ADMINISTRATOR	Admin can access the information of the users entering the system, access the uploaded file, add and delete files, edit the entered data.

WEB SITE ENVİRONMENT	The loaded documents, the results of the tested data, and the environment in which access to the classified data is provided.
STAKEHOLDERS	Any person who has contribution in the project.
PARTICIPANT	The user who interact with the documents classified by the semantic approach.
SDD	Software Design Document.
UML DIAGRAM	It is a modelling language which is used in Software Engineering.

4.1.4 Overview of document

The remaining chapters and their contents are listed below.

Section 1 included Describe the Project.

Section 2 is the Architectural Design which describes the project development phase. Also it contains class diagram and database diagram of the system and architecture design of the web application which describes actors, exceptions, basic sequences, priorities, pre-conditions and post conditions.

Section 3 is Use Case Realization. In this section, a block diagram of the system, which is designed according to use cases in SRS document, is displayed and explained.

Section 4 is related to Interface Design. In this section, we have shown the user Interface Design of Web Application from web site.

4.1.5 Motivation

We are a group of students in computer engineering department who are interested in Semantic Web. Now, the data become "big data". The big data learning and the data reasoning are very important for the future success of the Internet of Things (IoT). As a group, we researched Semantic web for a better understanding the most recent protocols, languages, technologies and tools that are used to build the Semantic Web. Also We have reviewed "big data" studies. We aimed to combine the Semantic Web technologies in this project. We have chosen the python language that interpreted, object-oriented, high-level programming language with dynamic semantics to develop our project. For including Semantic Web technologies, we have acquired and we have read documents for how to design Semantic Document Clustering.

4.2 Architecture design

4.2.1 Semantic Document Approach

For developing the project, we have planned to use Scrum which is an agile software development methodology. Thanks to the Agile method, project productivity, ability of the project to adapt to the variables in a fast way, project quality, excessive interaction among the team members and the delivery speed of the project increases. At the first appearance Scrum is an administrative model which has simple rules. It is applied for software projects which are open to changes and in which the requirements are not clearly specified.

Scrum includes Daily meetings in its scale and these meetings takes maxiumum 15 minutes. Project process are divided into sprints and these sprints has to be at fixed length. Scrum demonstrates the progress of the project in a clear and continuous way. At the end of each sprint some part of the project is completed and delivered to the client. The most significant advantage is the short sprints and dealing with the changes thanks to the feedbacks in an efficient way.Because of these, Scrum is going to be suitable for our project.

4.2.1.1 Class Diagram

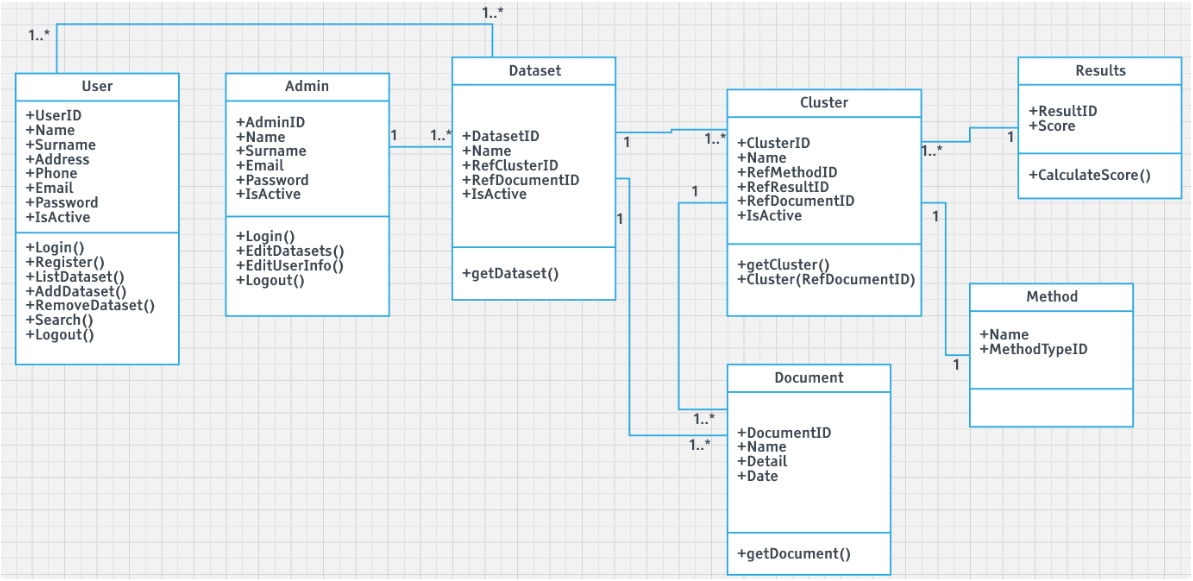


Figure 1. Class Diagram of Semantic Document Classification System

As figure 1, represents the static structure of the system. This diagram shows the structure of the system; expresses the classes in the system, the qualities of the classes and the relations between the classes. There are two users in this section. the first is the user, the other is the administrator. view user datasets, list datasets, search, add dataset, login, register and logout. Admin can edit users’ data, edit datasets, login and logout. Tables contain functions. Some of these functions allow us to access the table data. Some of them make calculations within themselves and we see the result.

4.2.1.2 Database Diagram

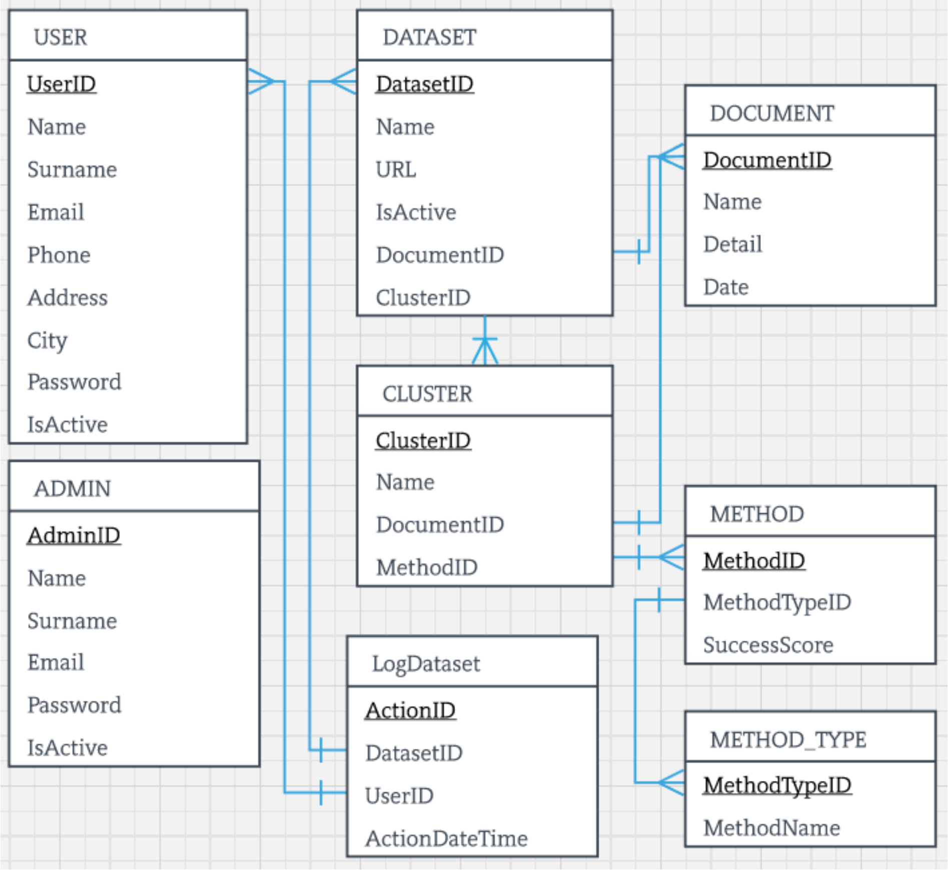


Figure 2. Database Diagram of Semantic Document Classification System

As mentioned in the Figure 2, there are related tables. the relations between them are provided according to the desired characteristics. In our system, a user has an admin on it, so there are two actor tabs. the dataset, cluster, and document tables that are outside of them are tied together. When the user inserts a dataset, he has to enter the information in the DATASET Table and we can see this information again from the DATASET Table. The DATASET Table depends on the CLUSTER Table. When the user adds a dataset, the information is held in the cluster table after the added data is clustered. The documents are kept in the DOCUMENT Table. If the user wants to see the documents, first the dataset and then the clustered folder reaches the document's data. System retrieves this data from the database. The user who wants to clustering selects the method. The user chooses the method name from page and the data is kept as an enum in database. The dialed data is stored in the id, but the user only selects the method name.

4.2.1.3 Sequence Diagram

Login Web Application

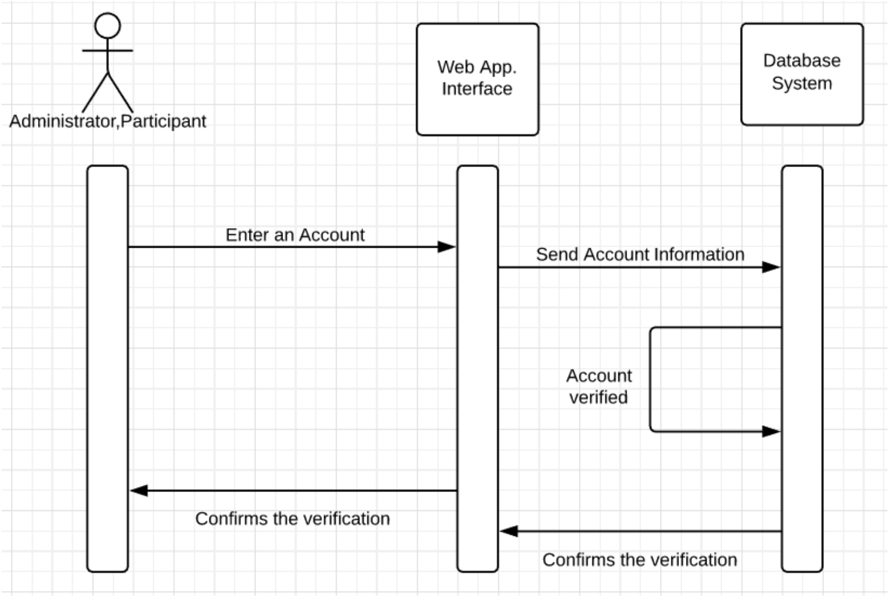


Figure 3. Login to Web App. Interface for Admin and Participant

In figure 3, Admin and User Enter Information in Login Page, then the system sends this information to database system. If the information mathes in databae system, User can login to web page succesfully.

Register Web Application

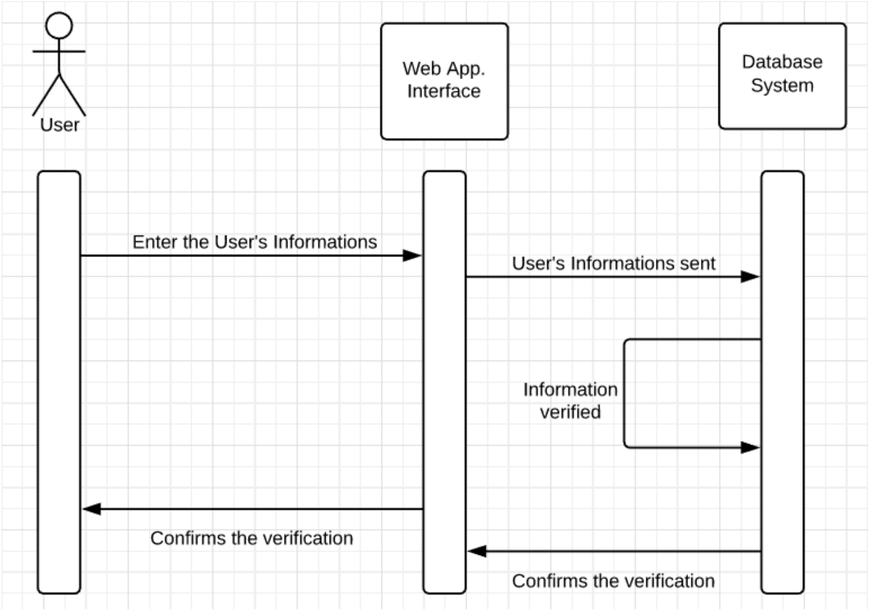


Figure 4. Register to Web App. Interface

As Figure 4, User enters the personal information like as name, surname, e-mail, password, phone.. etc.

Add Dataset to Web Application

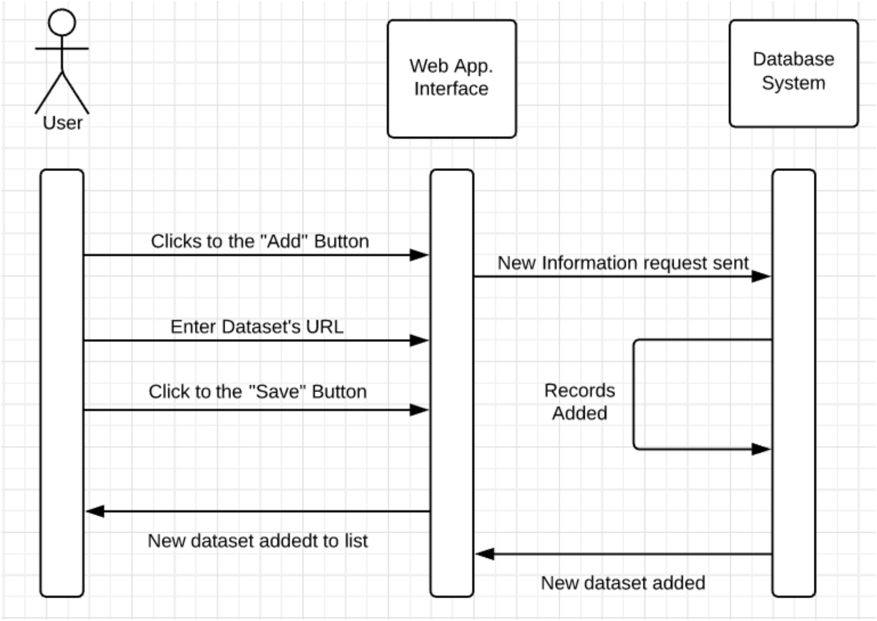

User
Web App. Interface
Database System
Clicks to the "Add" Button
New Information request sent
Enter Dataset's URL
Click to the "Save" Button
Records Added
New dataset addedt to list
New dataset added

Figure 5. Add Dataset to Web App. Interface

As Figure 5, The user enters the URL of the new dataset in the allocated space and clicks the add button. The system automatically creates new Dataset and insert it into the database.

List Dataset to Web Application

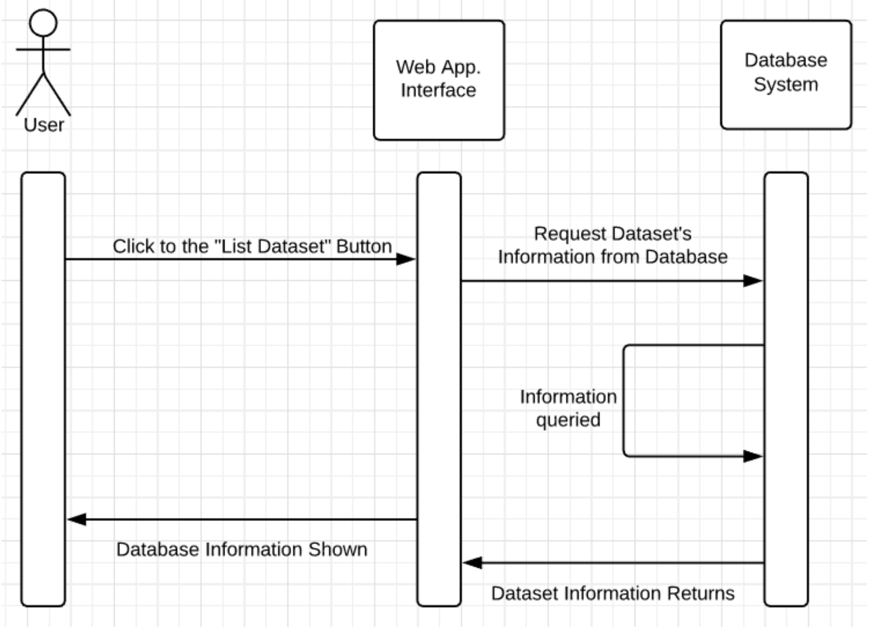

User
Web App. Interface
Database System
Click to the "List Dataset" Button
Request Dataset's Information from Database
Information queried
Database Information Shown
Dataset Information Returns

Figure 6. List Dataset to Web App. Interface

As Figure 6, The user click the “List Dataset” button and datasets are showed by the system.

Remove Dataset to Web Application

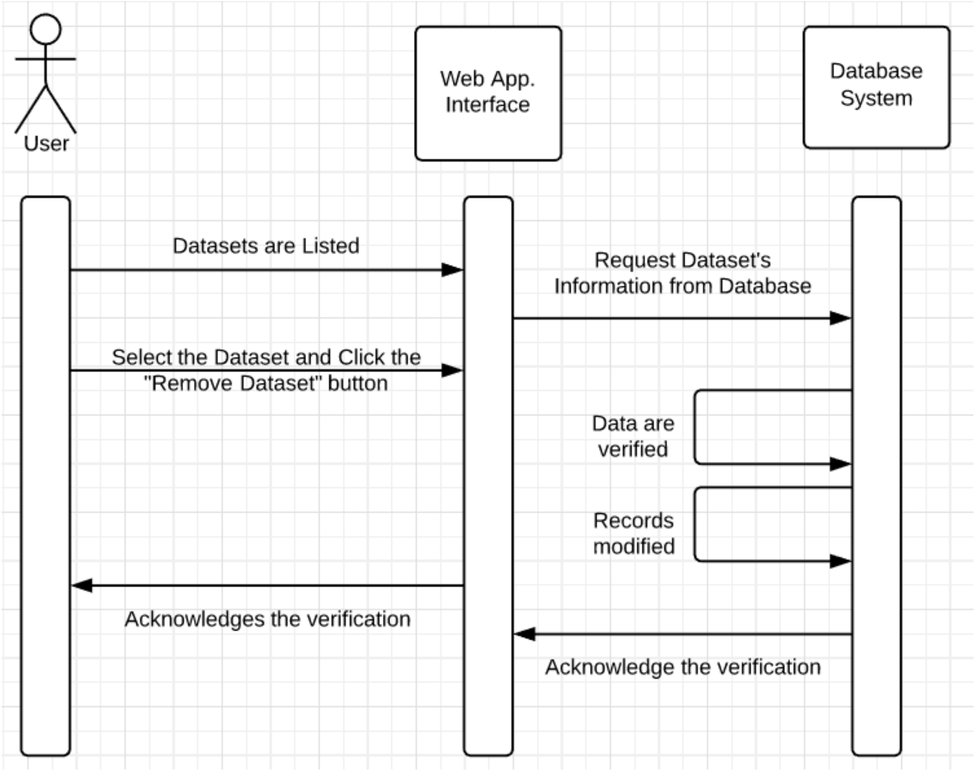

User
Web App. Interface
Database System
Datasets are Listed
Request Dataset's Information from Database
Select the Dataset and Click the "Remove Dataset" button
Data are verified
Records modified
Acknowledges the verification
Acknowledge the verification

Figure 7. Remove Dataset to web App. Interface

As Figure 7, The user select Dataset and then click the “Remove Dataset” button and dataset is removed by system.

Search Word From Documents to Web Application

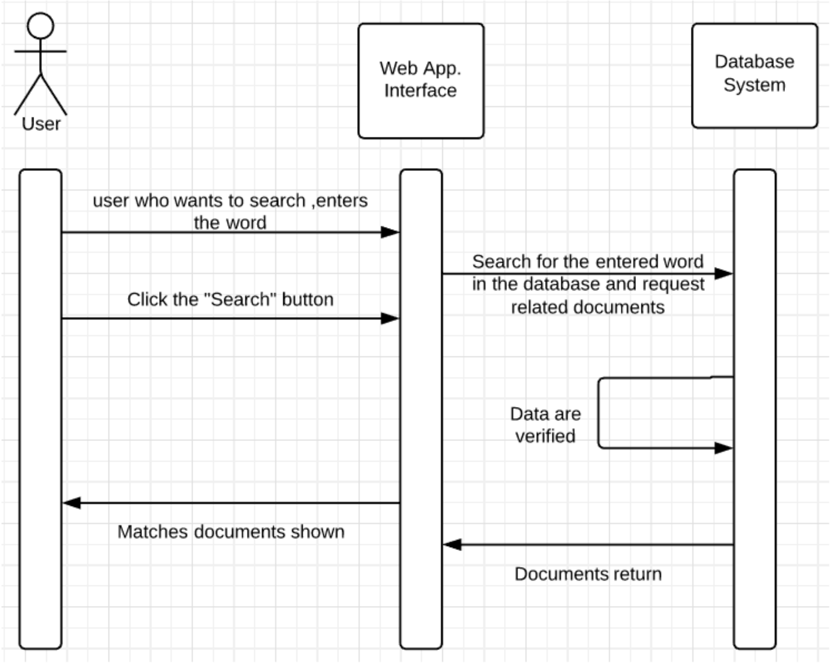

User
Web App. Interface
Database System
user who wants to search ,enters the word
Search for the entered word in the database and request related documents
Click the "Search" button
Data are verified
Matches documents shown
Documents return

Figure 8. Search Word from Documents to Web App. Interface

As Figure 8, The user who want to search ,enters word. System searchs in Database for entered word. Documents match and system shows the page these releated documents.

Edit Dataset to Web Application

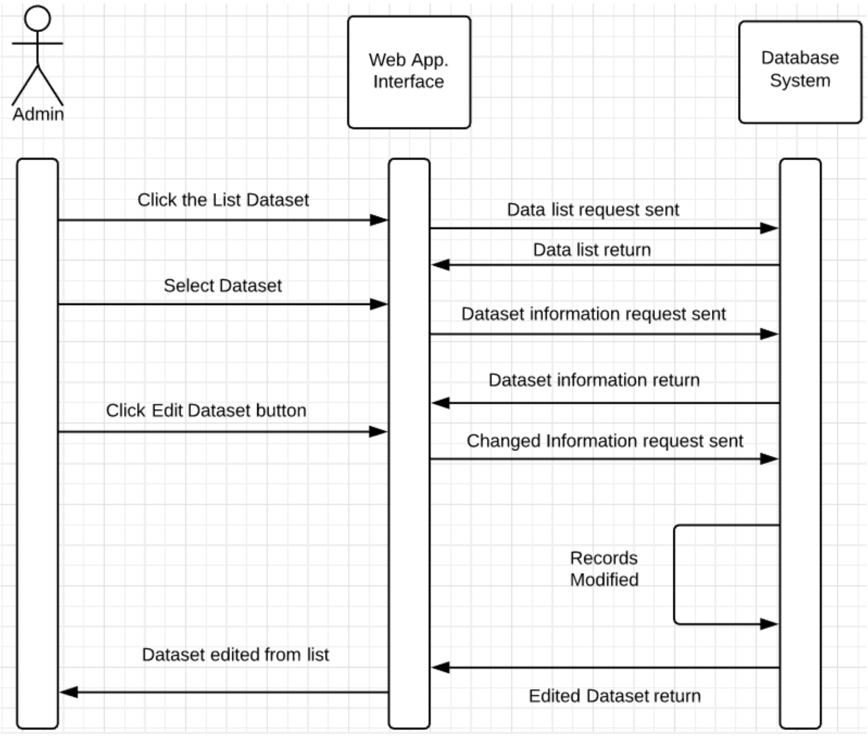


Figure 9. Edit Dataset to Web App. Interface

As Figure 9, The admin can edit all Dataset Information. Admin can Edit Dataset, Remove Dataset, Add Dataset…etc in this part. Admin who want to any meyhods can select and change the Information.

Edit User Information to Web Application

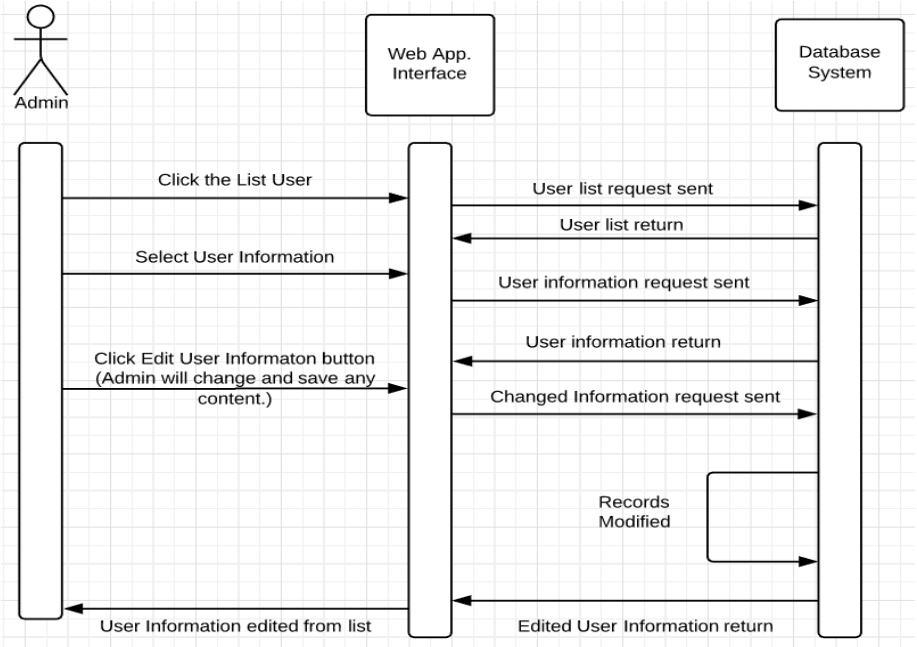

Admin
Web App. Interface
Database System
Click the List User
User list request sent
User list return
Select User Information
User information request sent
User information return
Click Edit User Informaton button (Admin will change and save any content.)
Changed Information request sent
Records Modified
User Information edited from list
Edited User Information return

Figure 10. Edit User Information to Web App. Interface

As Figure 10, The admin can edit all User Information, If the user send email to Admin for change own information. The admin edit the relevant fields at the user's request and saves them to the system .

4.3 Architecture Design of Web Application of Semantic Document Classification

4.3.1 Login Page

Summary: This web page is for administrators and participant to login and access to the system.

Actor: Participant, admin

Pre-condition: Registered in database system.

Post-condition: Access to the system.

Basic Sequence:

1. System Administrator or User enter the email and password.
2. The system Administrator or participant clicks on the "Login" button.
3. Confirmation of email and password between system and database is done.
4. If Login operation is authenticate, the administrator who logs in to the system will be directed to the admin page.
    - 4. 1. The participant who enters the system is directed to the participant page.
5. If not verified, the system will give an error message.
6. The system redirect to main page.

Exception: Database connection can be failed.

Post Conditions: None

Priority: Medium

4.3.2 Register Page

Summary: this page is for participants.Participant create an own account for using the system.

Actor: Participant

Pre-condition: User must have entered the register page. And user must have a current account.

Post-condition: A new user creates membership.

Basic Sequence:

1. The participant types the information like name,surname,email password into textboxes.
2. The system checks information for verification from database.
3. If validation is completed,the system redirect to Login page and user's information is added to database by system.

Exception: Database connection can be failed.

Post Conditions: None

Priority: Medium

4.3.3 Main Page

Summary: This pages is for adminstrator and participants. In this page, participants can view datasets, clusters of all documents, add documents, classify documents according to a desired method and see the results. The user can edit the documents on this page and access the information of the users.

Actor: Participant, admin

Pre-condition: Logged into the homepage on his/her own account.

Post-condition: Documents are clustered by system. Participants and Admin select what they want from the guidelines on the main page and see their results.

Basic Sequence:

1. The system administrator clicks to the ” Edit User Information” button.
    - 1. 1 The administrator selects user information for editing.
    - 1. 2 The administrator edits the user information on the system at the user’s request. Admin clicks to the “save” buton.
    - 1. 3 Information is edited by Administrator.
    - 1. 4 Sytem sends details of the changed information to the user by email.
2. The system adminstrator click to the “Edit Dataset buton.
    - 2. 1 All datasets are listed by the system.
    - 2. 2 Admin selects which dataset to change.
    - 2. 3 The system lists the clusters of the dataset selected by the Admin..
    - 2. 4 Admin clicks on the cluster button and the documents are listed by the system..
    - 2. 5 Admin will select and edit which document he wants to edit and then click on the “Save” button.
    - 2. 6 Changed Informations are saved by the system.
3. The user click to the “List Dataset” button.
    - 3. 1 User select the dataset in the computer and then click the Add button. Datasets are added by user in the system.
4. The use click to the “List Dataset” buton.
    - 4. 1 User see thee all uploaded Dataset List.
    - 4. 2 If the user want to see the clusters of the dataset, he/she can Access the clusters after the dataset buton.
    - 4. 3 If the user want to see the result of the datasets in the classification techniques, he/she click the Dataset buton. And then, user can see all results.
5. The user click to the “Remove Dataset” button.
    - 5. 1 User want to remove the attached dataset. Select the dataset.
    - 5. 2 User click the remove buton.

- 5. 3 Dataset removed by the system.

6. User click the “Cluster Dataset” button.
    - 6. 1 The user can see the success percentages of clustering methods for all datasets loaded or installed. User can choose method.
    - 6. 2 The user selects the dataset.
    - 6. 3 User selects the clustering method for the selected dataset.
    - 6. 4 If the user wants to clustering with the semantic method, he selects the dataset and clicks on the “semantic” button.
    - 6. 5 The system shows the success percentage of the clustering result.
    - 6. 6 If the user wishes to make a cluster with Word Embedding method, he selects Dataset and clicks “Word Embedding” button.
    - 6. 7 The system shows the success percentage of the clustering result.

Exception: None

Post Conditions: Changes made by admin will be saved within related table. Changes made by user will be saved within related table.

Priority: High

4.4 Activity Diagram

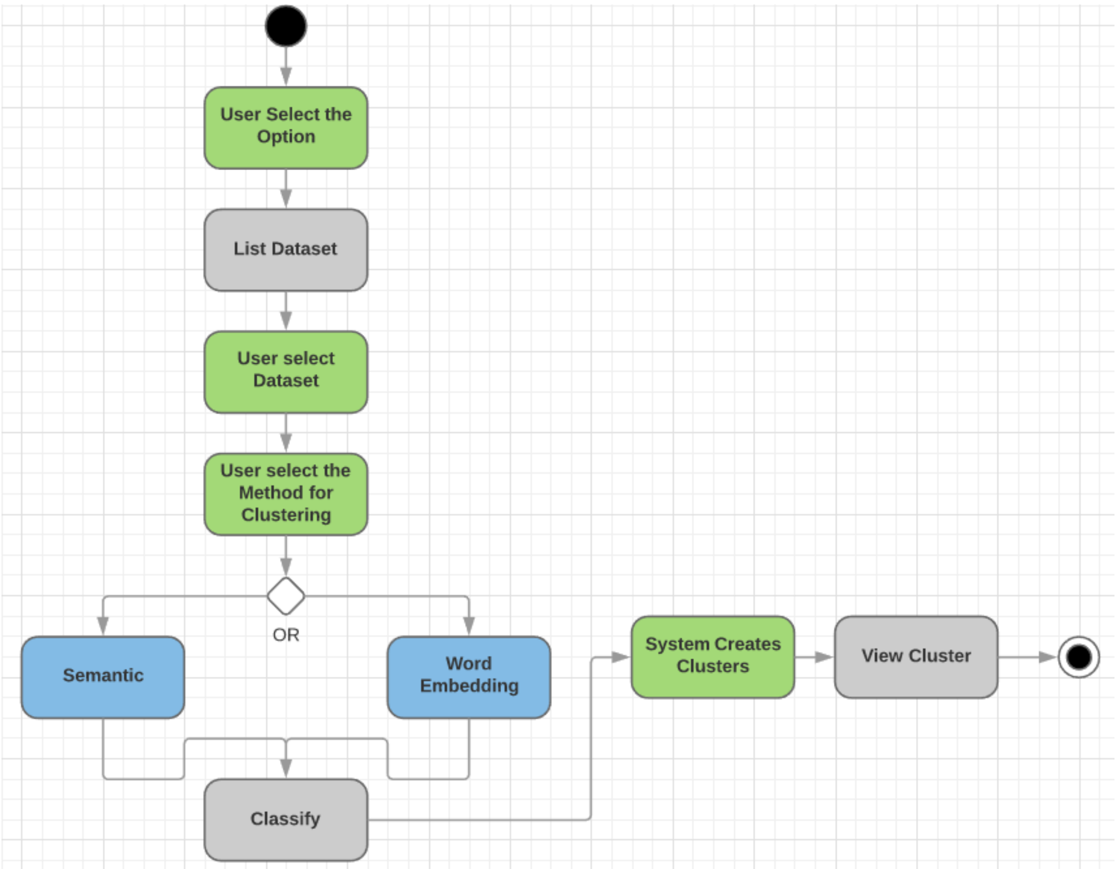


Figure 11. Activity Diagram For Semantic Document Classification

As figure 11, the program works as follows. the user logs in to the system. He/she clicks on the Corpus Button and sees the dataset. User clicks the “List Dataset” button and select dataset. Then the user chooses which method he wants to classify. These methods are semantic or Word embedding. The system cluster the datasets according to the method selected by the user. Then the user wants to see these datasets. Finally, user can view the Dataset.

4.5 Use case realizations

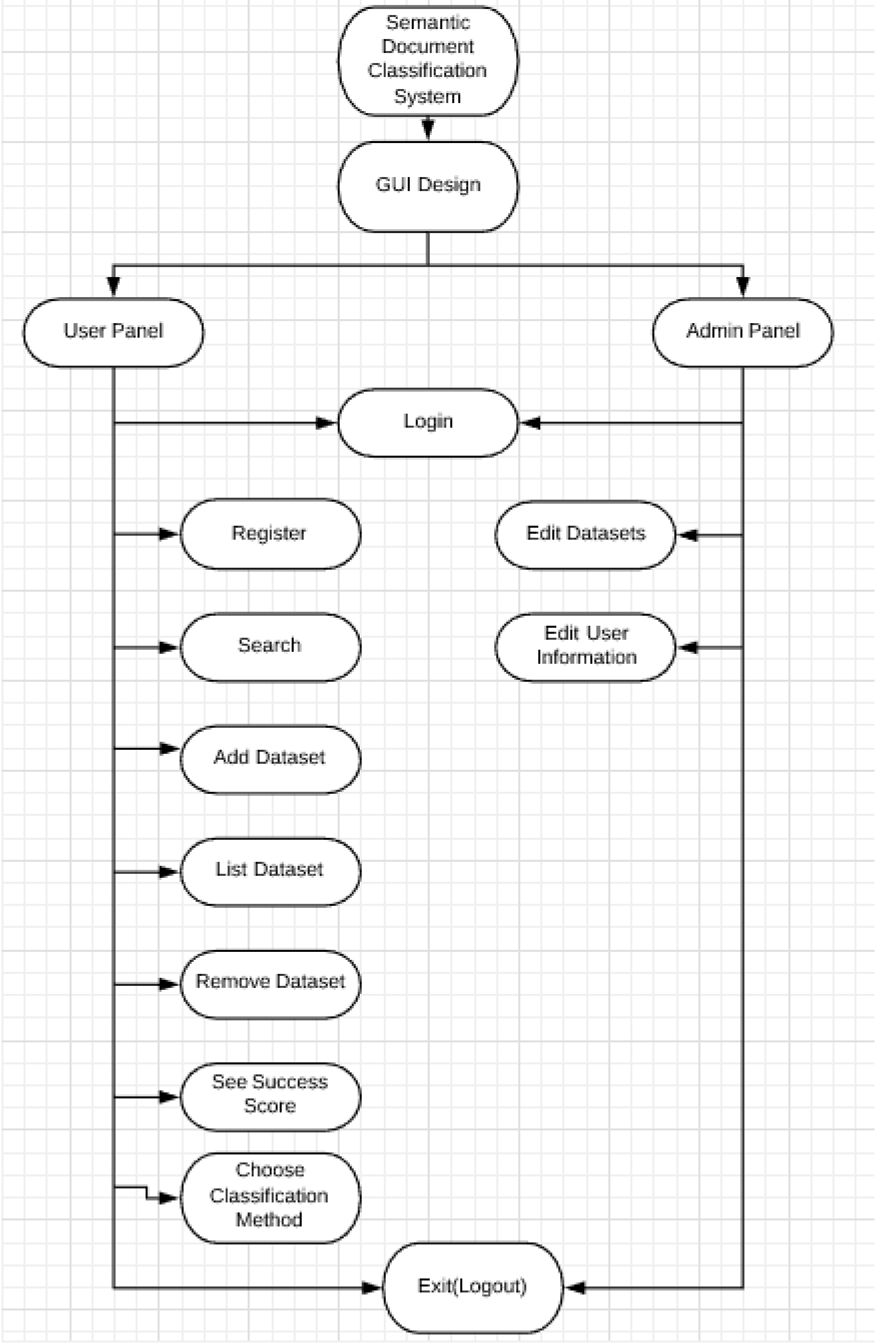


Figure 12. Semantic Document Classification Block Diagram

4.5.1 Brief Description of Figure 12

Components of the Semantic Document Classification Project are shown in the Figure 12. All designed systems of the semantic are displayed in the block diagram in the figure. There are two main components of the system which have their own sub-systems.

4.5.1.1 User Panel Design

In student panel, user be able to interact with system. System has own user main page and it contains four sub-system; List Dataset, Add Dataset, Remove Dataset, Search Word. The user enters the word he / she wants to search and the system searches for words in all documents. Finally, the system brings the documents it finds to the screen. User may want to see datasets. To do this, click the List Dataset button. The system automatically brings up the information of all documents and documents. The user sees all the documents uploaded on the system and the success of the previous classification methods. User can add Dataset, If the user clicks Add Dataset button. User who want to add document selects the document set from the computer and loads dataset into the system. It can categorize the dataset it loads. If the user clicks Choose Classification Method, there are two options. first semantic second word embedding. the user can classify the documents into the system by selecting the desired method. Datasets are loaded by system. the user, who added the dataset, can delete the added dataset , if user clicks the Remove Dataset button. The system displays the datasets added by the user himself. user makes selection and deletes. The system deletes all information from the deleted dataset from the system.

4.5.1.2 Admin Panel Design

In admin panel, admin be able to interact with system. System has own admin main page and it contains two sub-system; Edit Datasets, Edit User Information.The administrator can check and see everything that has been added, deleted, changed, added to the system. The administrator can delete the dataset information by clicking the Edit Dataset button. It displays all the information of the datasets on the screen. change the location of the selected dataset and save it. The system saves changes. If the admin wants to delete the dataset, Admin makes false the Dataset’s IsActive and closes the Dataset’s access to the user. If the user wants their information to change, mail to admin. The requested information is opened on the admin page and the admin can easily update the information.

4.6 User Interface Model

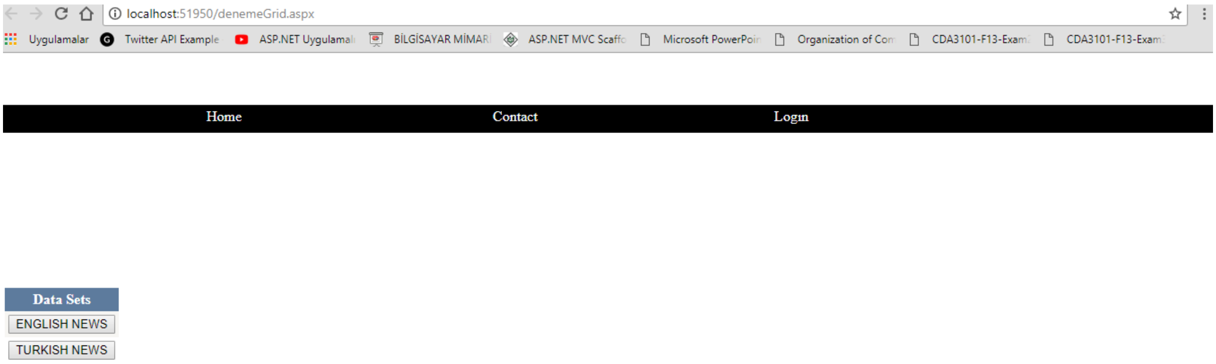


Figure 13. Home Page

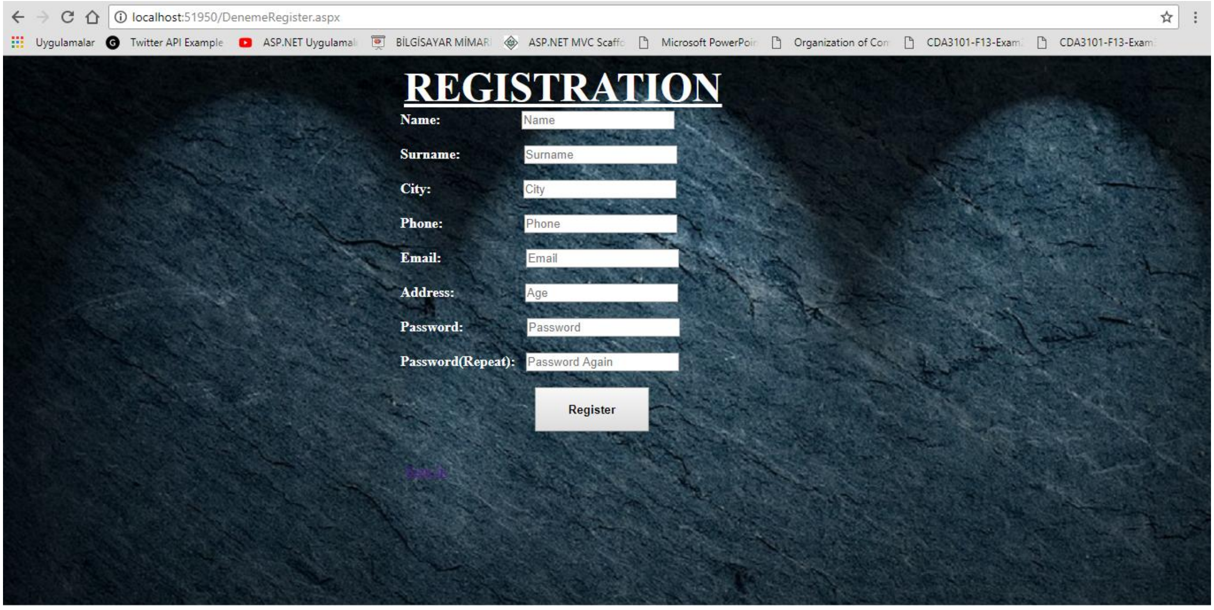


Figure 14. Register Page

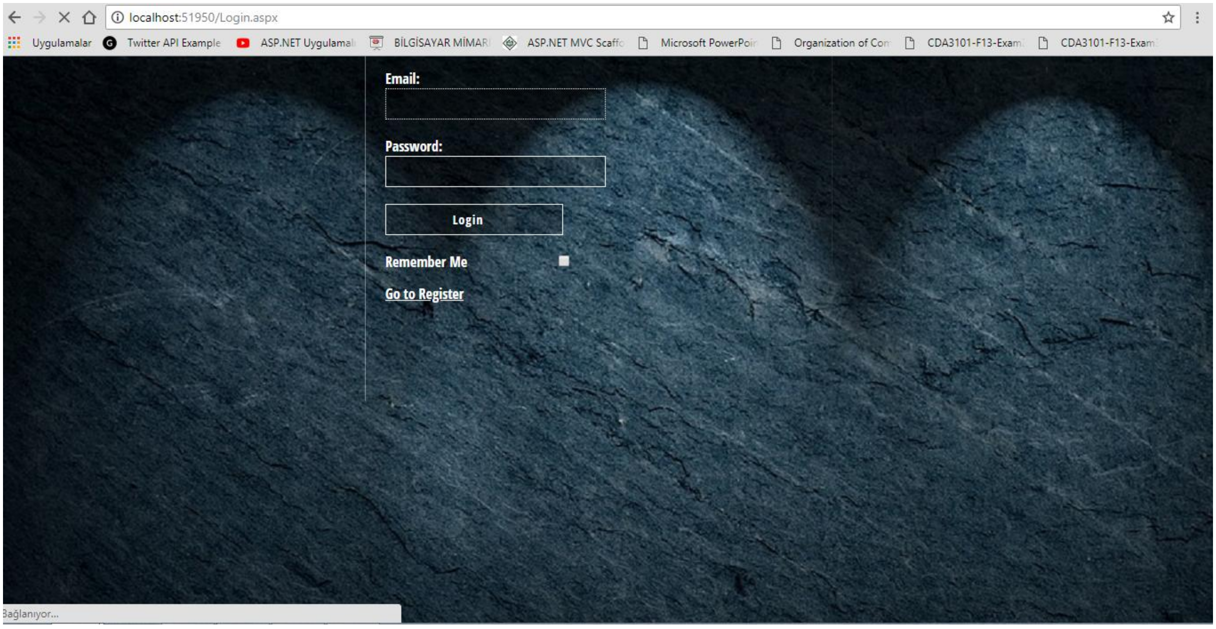


Figure 15. Login Page

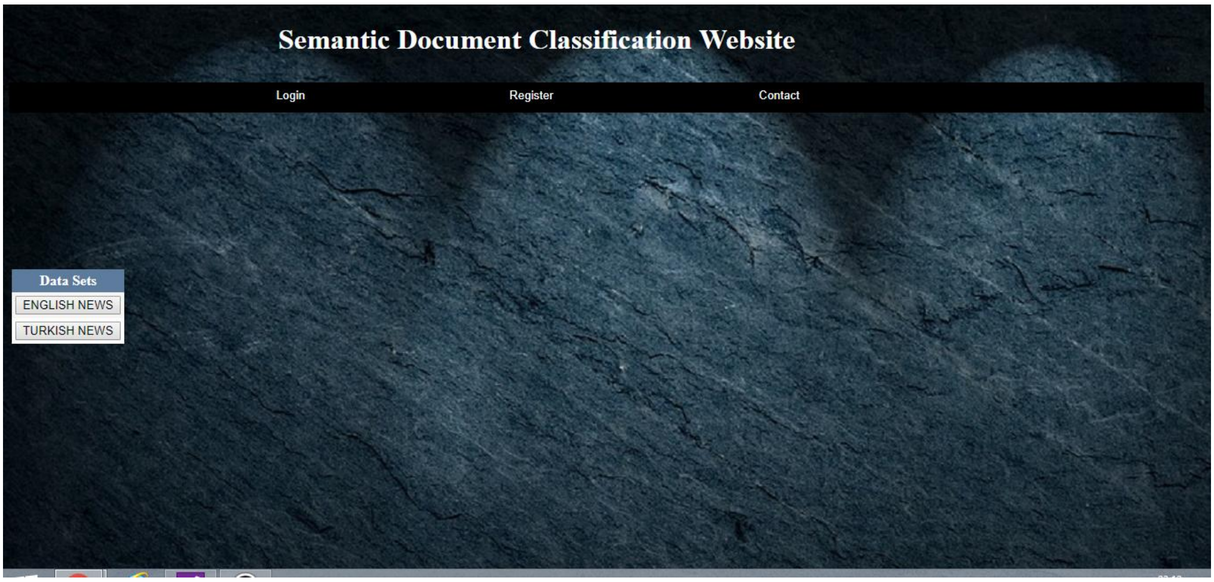


Figure 16. Click the Homepage Button

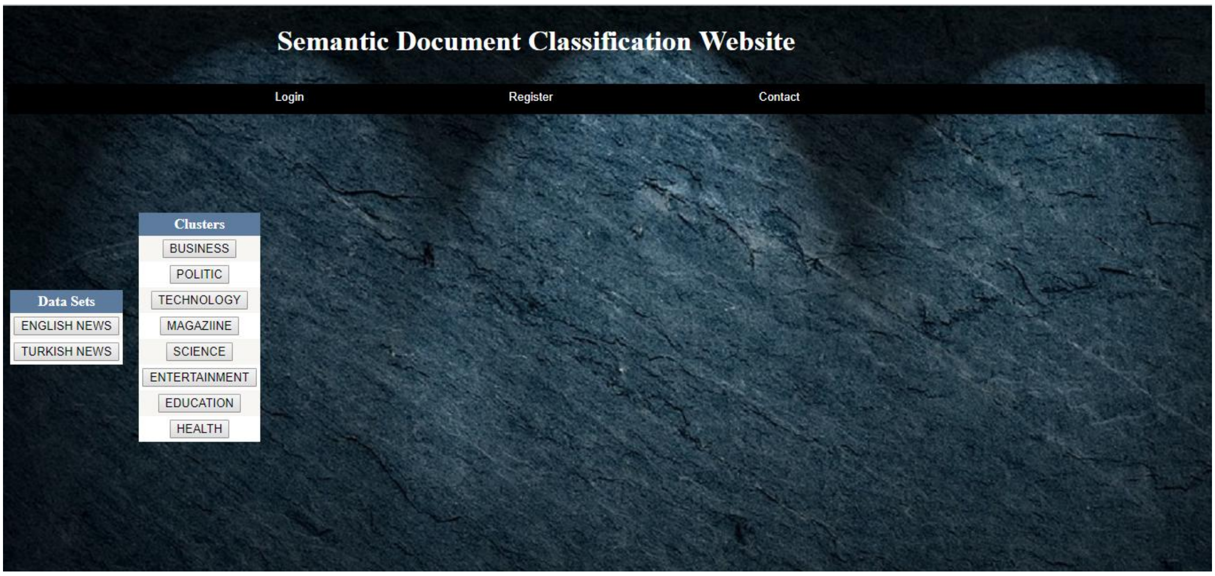


Figure 17. Click the Dataset Button

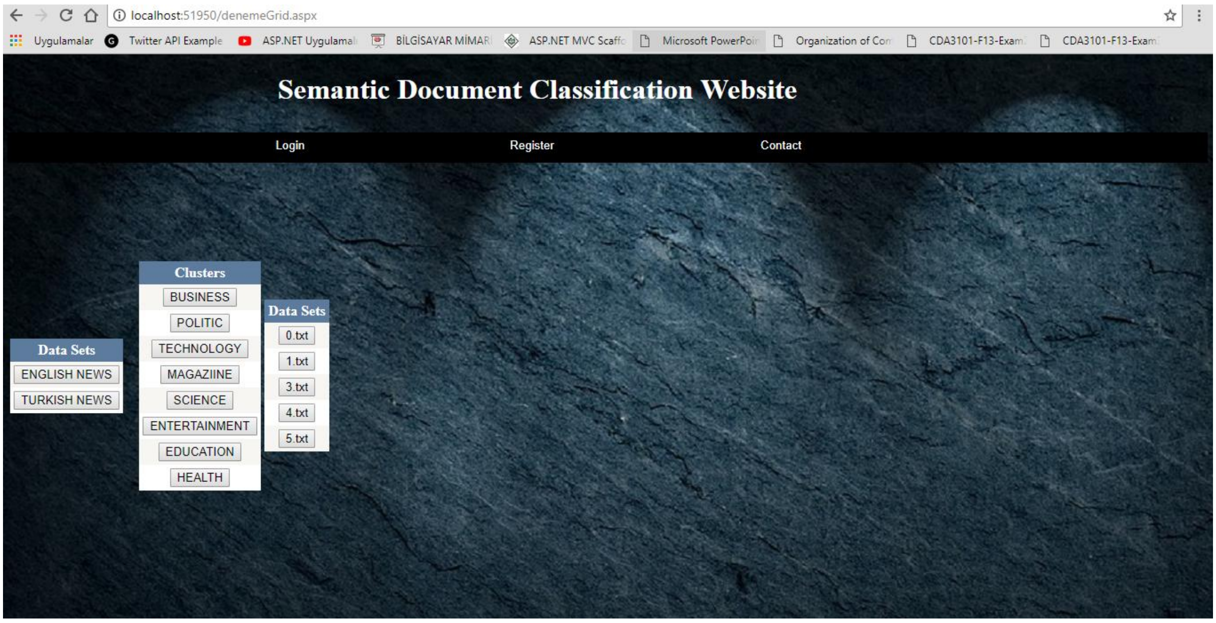
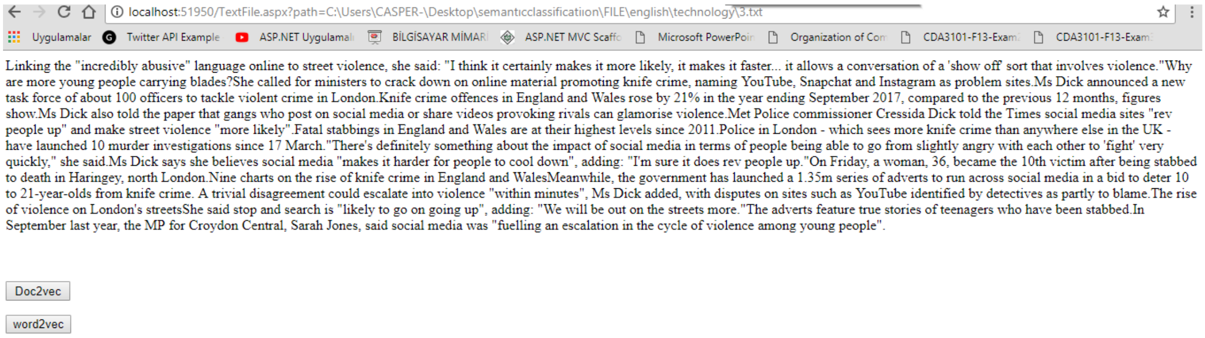
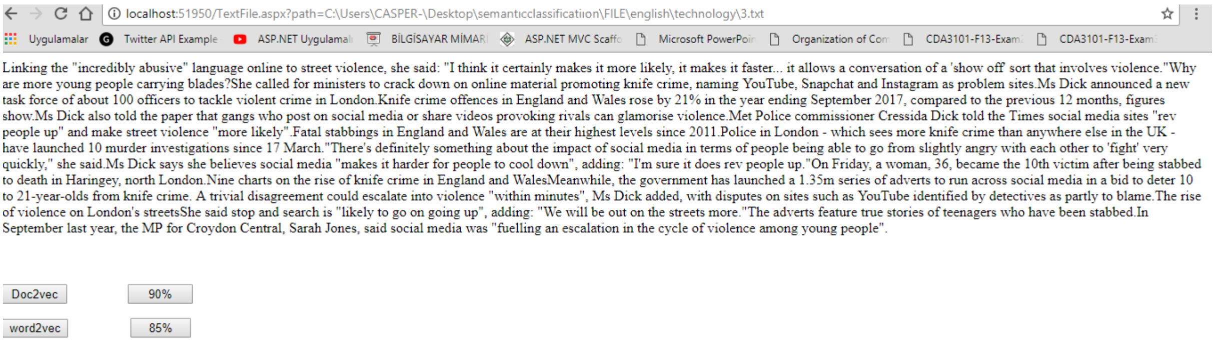


Figure 18. Click the Cluster Button



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".

Figure 19. Click the Do test Button



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".

Figure 20. Click the See Success Score Button for Loading Document

5.Test Plan

5.1 Introduction

5.1.1 Version Control

Version No	Description of Changes	Date
1.0	First Version	Apr 25, 2018
2.0	Second Version	May 22,2018

5.1.2 Overview

The use case of Semantic Document Classification System users namely participant and system administrator which has been determined in SRS document will be tested.

5.1.3 Scope

This document includes the test plan of use cases, test design specifications and test cases correspond to test plan.

Scope of this worked to classify any text document that you have with created dataset from news (BBC News).

Dataset includes categorized news as Economics, Technology, Sport, Science, Politics, Health, Magazine, Entertainments.

5.1.4 Terminology

Acronym	Definition
UWP	User Web Page
SAWP	System Administrator Web Page
Word2Vec	generate representation vectors out of words
Doc2Vec	create a numeric representation of a document

5.2 FEATURES TO BE TESTED

This section lists and gives a brief description of all the major features to be tested. For each major feature there will be a Test Design Specification added at the end of this document.

5.2.1 User Web Page (UWP)

In this project, User Interface constituents are used. The UI part is divided into ten parts which are Register, Login, Add Dataset, List Dataset, Remove Dataset, Do test, Choose Classification Methods, See Success Score, Search and Exit. Every part of the UI also includes smaller parts and includes test of functions of UI components which are used in this project such as panel, button, search, etc. The participant shall sign in to account, if there is no account, they will be registered. Then the participant shall fill in the own information. Then participant who wants to reach anything, clicks on the button . If the dataset wants to load it clicks the corresponding button and the system waits for a data upload. If participant want to see the datasets, click on the list dataset button and the system will show all registered datasets on the screen. Testing of the stated requirements will occur in this document.

5.2.2 System Administrator Web Page (SAWP)

This part includes test cases and test plan of Administrator Interface. It includes Login, Edit Datasets, Edit User Information and Exit. The admin shall edit dataset and edit user information(Remove, change information). Testing of the stated requirements will occur in this document.

5.2.3 Word2Vec

A numeric representation for each word, that will be able to capture such relations as above. Such representations, encapsulate different relations between words, like synonyms, antonyms, or analogies. Thanks to Word2Vec, we create a vectorel space via words. Classify with this method. Word2vec representation is created using 2 algorithms: Continuous Bag-of-Words model (CBOW) and the Skip-Gram model.

5.2.4 Doc2Vec

Doc2vec is to create a numeric representation of a document, regardless of it's length. But unlike words, documents do not come in logical structures such as words, so another method has to be found.

5.3 ITEM PASS/FAIL CRITERIA

Describe the general rule to use to decide when a test case passes and when it fails.

5.3.1 Exit Criteria

Describe under what conditions the testing of the product is considered successful. Some examples are:

- 100% of the test cases are executed
- 95% of the test cases passed
- All High and Medium Priority test cases passed

5.4 REFERENCES

[1] P2018_SRS_V1.1, 13 March, 2018

[2] P2018_SDD_V1.1, 13 March, 2018

5.5 TEST DESIGN SPECIFICATIONS

5.5.1 User Web Page(UWP)

5.5.1.1 Sub features to be tested

5.5.1.1.1 Register (UWP.RG)

The participant can indicate his/her personal information to login the system.

2.1.1.1 Login (UWP.LG)

The participant has to login the web application system by entering email and password.

5.5.1.1.3 List Dataset(UWP.LD)

The participant can list dataset. The system automatically brings up the information of all documents and documents.

5.5.1.1.4 Add Dataset (UWP.AD)

The participant can add dataset. User who want to add document selects the document set from the computer and loads dataset into the system. It can categorize the dataset it loads.

5.5.1.1.5Remove Dataset (UWP.RD)

The participant can remove dataset. The user, who added the dataset, can delete the added dataset.

5.5.1.1.6 See Success Score (UWP.SSS)

The participant can see success score for datasets.

5.5.1.1.7Choose Classification Method (UWP.CCM)

The participant can clicks Choose Classification Method, there are two options. first semantic second word embedding. the user can classify the documents into the system by selecting the desired method.

5.5.1.1.8 Search (UWP.SRC)

The participant can search the word he/she wants to see in all document.

5.5.1.1.9 Exit (UWP.EXT)

The participant can leave this system with Exit button.

5.5.1.2 Test Cases

Here list all the related test cases for this feature

TC ID	Requirements	Priority	Scenario Description
UWP.RG.01	3.1	H	Enter a valid user email and password.
UWP.RG.02	3.1	H	Enter a invalid user email.

TC ID	Requirements	Priority	Scenario Description
UWP.LG.01	3.1	H	Enter a valid user email and password.
UWP.LG.02	3.1	H	Enter a valid email and invalid password.

TC ID	Requirements	Priority	Scenario Description
UWP.LD.01	3.1	H	Dataset list, which mean that dataset's information, appeals dataset and when to loaded and approvals.

TC ID	Requirements	Priority	Scenario Description
UWP.AD.01	3.1	H	Select dataset from computer, After selecting datasets will be uploaded by the system.
UWP.AD.02	3.1	H	Select existing dataset in database from computer, After selecting dataset will not be added the dataset.

TC ID	Requirements	Priority	Scenario Description
UWP.RD.01	3.1	H	Select added dataset by user who log in this system, After selecting dataset will remove.
UWP.RD.02	3.1	H	Select added dataset by user who not log in this system, After selecting dataset will not be remove.

TC ID	Requirements	Priority	Scenario Description
UWP.SSS.01	3.1	H	Click on the button to see the success scores of all the data.
UWP.SSS.02	3.1	H	Click on the button to see the success score of uploaded data.

TC ID	Requirements	Priority	Scenario Description
UWP.CCM.01	3.1	H	Select a method in list. After selecting, system automatically classified these datasets and scores will be displayed.
UWP.CCM.02	3.1	H	Select an empty. After selecting, system not automatically classified these datasets and scores will not be displayed.

TC ID	Requirements	Priority	Scenario Description
UWP.SRC.01	3.1	L	Select search word and write a word. After selecting, relevant articles will be displayed.
UWP.SRC.02	3.1	L	Write incomplete word . After writing and selecting search button, relevant articles will be displayed.
UWP.SRC.03	3.1	L	Leave blank and select search button. After selecting, articles will not be displayed.

TC ID	Requirements	Priority	Scenario Description
UWP.EXT.01	3.1	H	Select completely the Exit Button. User can completely leave this page.

5.5.2 System Administrator Web Page (SAWP)

5.5.2.1 Subfeatures to be tested

5.5.2.1.1 Login (SAWP.LG)

The admin has to login the web application system by entering email and password

5.5.2.1.2 Edit User Information (SAWP.EUI)

The admin can change, remove, add and list user's information.

5.5.2.1.3 Edit Datasets (SAWP.ED)

The admin can change, remove, add and list added datasets.

5.5.2.1.4 Exit (SAWP.EXT)

The admin can change, remove, add and list added datasets.

5.5.2.2 Test Cases

Here list all the related test cases for this feature

TC ID	Requirements	Priority	Scenario Description
SAWP.LG.01	3.2	H	Enter a valid user email and password.
SAWP.LG.02	3.2	H	Enter a valid email and invalid password.

TC ID	Requirements	Priority	Scenario Description
SAWP.EUI.01	3.2	H	If the user wants to change own information, he/she can contact admin. Then, information is edited by Administrator.
SAWP.EUI.02	3.2	H	If the user forgets password, he/she can contact admin. Then, information is edited by Administrator.

TC ID	Requirements	Priority	Scenario Description
SAWP.ED.01	3.2	H	Admin will select and edit which document he wants to edit and then click on the “Save” button. Changed informations are saved by the system.
SAWP.ED.02	3.2	H	Admin will select and not edit which document he wants to edit and then click on the “Edit Dataset” button. Not Changed informations are not saved by the system because of there is no edit.

TC ID	Requirements	Priority	Scenario Description
SAWP.EXT.01	3.2	H	Select completely the Exit Button. Admin can completely leave this page.

5.5.3 Detailed Test Cases

5.5.3.1 UWP.RG.01

TC_ID	UWP.RG.01
Purpose	Enter a valid user email and password.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Register user test cases pass.
Setup	An user should be created.
Procedure	[A01] Go to login page.
	[A02] Click the "Register to Login " button.
	[A03] Enter name, surname, phone address, city.
	[A04] Enter a valid email and password.
	[V01] Observe that the registered is successful and the admin page appears.
Cleanup	

5.5.3.2 UWP.RG.02

TC_ID	UWP.RG.02
Purpose	Enter a invalid user email and password.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Register user test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to login page.
	[A02] Click "Register to Login" button.
	[A03] Enter name, surname, phone, address, city.
	[A04] Enter an invalid email and password.
	[V01] Observe that the register is unsuccessful and "The email is currently being used" message, and register page refreshes.
Cleanup	

5.3.3.3 UWP.LG.01

TC_ID	UWP.LG.01
Purpose	
Requirements	
Priority	
Estimated Time Needed	
Dependency	Login Web Page should be opened.
Setup	An user should be created.
Procedure	[A01] Go to Login page.
	[A02] Enter a valid user email.
	[A03] Enter a valid password for this user.
	[A04] Click on the "Login" button.
	[V01] Observe that the Login is successfully and the user interface page appears.
Cleanup	

5.3.3.4 UWP.LG.02

TC_ID	UWP.LG.02
Purpose	Enter a valid email and password.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Login Web Page should be opened.
Setup	An user should be created.
Procedure	[A01] Go to Login page.
	[A02] Enter a valid user email.
	[A03] Enter a invalid password for this user.
	[A04] Click on the "Login" button.
	[V01] Observe that the Login is unsuccessful and "The email/password is wrong, try again." message, the Login page refreshes.
Cleanup	

5.5.3.5 UWP.LD.01

TC_ID	UWP.LD.01
Purpose	Dataset List, which mean that dataset's information, appeals dataset and when to loaded and approvals.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	List Dataset test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] Click ”List Dataset” button.

	[V01] Observe that show the Dataset List for information.
Cleanup	

5.5.3.6 UWP.AD.01

TC_ID	UWP.AD.01
Purpose	Select dataset from computer. After selecting dataset will be uploaded by the system.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Add Dataset test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] Click ”Add Dataset” button.
	[A03] Select dataset from computer.
	[V01] Observe that the a new dataset uploaded.
Cleanup	

5.5.3.7 UWP.AD.02

TC_ID	UWP.AD.02
Purpose	Select existing dataset from computer, After selecting dataset will not be added the dataset.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Add Dataset test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] Click ”Add Dataset” button.
	[A03] Select existing dataset from computer.
	[A04] Enter an invalid email and password.
	[V01] Observe that "This dataset already in database" message and go back to interface.
Cleanup	

5.5.3.8 UWP.RD.01

TC_ID	UWP.RD.01
Purpose	Select added dataset by user who log in this system. After selecting dataset will remove.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Remove Dataset test cases should pass.

Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] The user must click on the dataset that the user has added and then click "Remove Dataset" button.
	[V01] Observe that "This Dataset deletes completely" message and go back to interface.
Cleanup	

5.5.3.9 UWP.RD.02

TC_ID	UWP.RD.02
Purpose	Select added dataset by user who not log in this system, After selecting dataset will not be removed.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Remove Dataset test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] The user who have not added the dataset, click on the this dataset and then click "Remove Dataset" button.
	[V01] Observe that "This Dataset is not deleted" message and go back to interface.
Cleanup	

5.5.3.10 UWP.SSS.01

TC_ID	UWP.SSS.01
Purpose	Click on the button to see the success scores of all the data.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	See Success Score test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] Select Dataset and then Click ”See Success Score” button.
	[V01] Observe that show the Success Score List for information.
Cleanup	

5.5.3.11 UWP.SSS.02

TC_ID	UWP.SSS.02
Purpose	Click on the button to see the success score of uploaded data.
Requirements	3.1
Priority	High.

Estimated Time Needed	3 Minutes
Dependency	See Success Score test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] Select Add Dataset, After the addition If the user want to see classification success score , click ”See Success Score” button.
	[V01] Observe that show the Success Score List for information.
Cleanup	

5.5.3.12 UWP.CCM.01

TC_ID	UWP.CCM.01
Purpose	Select a method in list. After selecting, system automatically classified these datasets and scores will be displayed.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Register user test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] Select Dataset or Add Dataset.
	[A03] Click "Semantic" or "Word embedding" method for classification.
	[V01] Observe that the dataset is classified with selected method.
Cleanup	

5.5.3.13 UWP.CCM.02

TC_ID	UWP.CCM.01
Purpose	Select an empty. After selecting, system not automatically classified these datasets and scores will not be displayed.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Register user test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] Select Empty Dataset.
	[A03] Click "Semantic" or "Word embedding" method for classification.
	[V01]] Observe that "This Dataset is not classified " message and go back to interface.
Cleanup	

5.5.3.14 UWP.SRC.01

TC_ID	UWP.SRC.01
Purpose	Select search word and write a word. After selecting, relevant articles will be displayed.
Requirements	3.1
Priority	Low.
Estimated Time Needed	3 Minutes
Dependency	Search test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] Enter the word for see articles.
	[A03] Click "Search" button.
	[V01] Observe that show the word with highlight in articles.
Cleanup	

5.5.3.15 UWP.SRC.02

TC_ID	UWP.SRC.02
Purpose	Write incomplete word . After writing and selecting search button, relevant articles will be displayed.
Requirements	3.1
Priority	Low.
Estimated Time Needed	3 Minutes
Dependency	Search test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] Enter the incomplete word for see articles
	[A03] Click "Search" button.
	[V01] Observe that show the word with highlight in articles.
Cleanup	

5.5.3.16 UWP.SRC.03

TC_ID	UWP.SRC.03
Purpose	Leave blank and select search button. After selecting, articles will not be displayed.
Requirements	3.1
Priority	Low.
Estimated Time Needed	3 Minutes
Dependency	Search test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to User Web Page.
	[A02] Enter the Leave blank.
	[A03] Click "Search" button
	[V01] Observe that the search is unsuccessful and “This

	word is not searched” message, and go back to interface.
Cleanup	

5.5.3.17 UWP.EXT.01

TC_ID	UWP.EXT.01
Purpose	Select completely the Exit Button. Admin can completely leave this page.
Requirements	3.1
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	User Web Page test cases should pass.
Setup	An user should be created.
Procedure	[A01] Go to Login page.
	[A02] Enter a valid user email
	[A03] Enter a valid password for this user.
	[A04] Click on the "Login" button.
	[A05] User interface page is loaded.
	[A06] Click on the "Exit" button for leaving Web Page.
	[V01] Observe that the Exit is successfully the user leave web page.
Cleanup	

5.5.3.18 SAWP.LG.01

TC_ID	SAWP.LG.01
Purpose	Enter a valid email and password.
Requirements	3.2
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Login Web Page should be opened.
Setup	An user should be created.
Procedure	[A01] Go to Login page.
	[A02] Enter a valid user email.
	[A03] Enter a valid password for this user.
	[A04] Click on the "Login" button.
	[V01] Observe that the Login is successfully and the user interface page appears.
Cleanup	

5.5.3.19 SAWP.LG.02

TC_ID	SAWP.LG.02
Purpose	Enter a valid email and invalid password.
Requirements	3.2
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Login Web Page should be opened.

Setup	An user should be created.
Procedure	[A01] Go to Login page.
	[A02] Enter a valid user email.
	[A03] Enter a invalid password for this user.
	[A04] Click on the "Login" button.
	[V01] Observe that the Login is unsuccessful and "The email/password is wrong, try again." message, the Login page refreshes.
Cleanup	

5.5.3.20 SAWP.EUI.01

TC_ID	SAWP.EUI.01
Purpose	If the user wants to change own information, he/she can contact admin. Then, information is edited by Administrator.
Requirements	3.2
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	System Administrator Web Page test cases should pass
Setup	An User should be created.
Procedure	[A01] Go to request list at Edit User Information Page.
	[A02] Select a user who wants to changes information.
	[A03] Edit user's information.
	[A04] Click on the "Edit" button.
	[V01] Observe that the Edit is successful and "Informations are edited" message shown and Changing information message is sent to user's email by the system.
Cleanup	

5.5.3.21 SAWP.EUI.02

TC_ID	SAWP.EUI.02
Purpose	If the user forgets password, he/she can contact admin. Then, information is edited by Administrator.
Requirements	3.2
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	Login Web Page should be opened.
Setup	An user should be created.
Procedure	[A01] Go to request list at Edit User Information Page.
	[A02] Select a user who want new password.
	[A03] New password is created by system.
	[A04] Click on the "Edit" button.
	[V01] Observe that the Edit is successful and "Informations are send" message shown and Changing information message is sent to user's email by the system.
Cleanup	

5.5.3.22 SAWP.ED.02

TC_ID	SAWP.ED.01
Purpose	Admin will select and edit which document he wants to edit and then click on the “Save” button. Changed informations are saved by the system.
Requirements	3.2
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	System Administrator Web Page should be opened.
Setup	An Admin should be created.
Procedure	[A01] Go to List Dataset Page.
	[A02] Select a dataset.
	[A03] Admin can change uploaded date, name date...
	[A04] Click on the "Edit Dataset" button.
	[V01] Observe that the Edit Dataset is successful and " Dataset is edited" message, Dataset edited from list.
Cleanup	

5.5.3.23 SAWP.ED.02

TC_ID	SAWP.ED.02
Purpose	Admin will select and not edit which document he wants to edit and then click on the “Edit Dataset” button. Not Changed informations are not saved by the system because of there is no edit.
Requirements	3.2
Priority	High.
Estimated Time Needed	3 Minutes
Dependency	System Administrator Web Page should be opened.
Setup	An Admin should be created.
Procedure	[A01] Go to List Dataset Page.
	[A02] Select a dataset.
	[A03] Admin does not edit which document he wants to edit.
	[A04] Click on the "Edit Dataset" button.
	[V01] Observe that the Edit is unsuccessful and " Not Changed informations are not saved by the system because of there is no edit " message, Go to List Dataset Page.
Cleanup	

5.6 Test Results

Models	Number of DATA	Result %
Word2Vec	2000	95.4
Doc2Vec	2000	91.8

The test cases are implemented and 95.4% of the test cases are completed successfully. Software development activities are completed within the anticipated cost.Current stage of the project is available to use. Failed test cases will be solved in time in order to increase the quality.

6. Conclusions

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. Then expanding them on a test and classify documents more accurately and in a comprehensive manner all the methods to develop new semantic methods compare. As mentioned earlier, the above method to classify documents using a software tool. We have studied the requirements given to us in our work. We discussed how we can show them and how we can present them clearly. We're trying to put out some data. We have also examined the methods that have been tried in our research. We've been working on how to solve the problem more easily. We designed interfaces that people could easily understand. We tried to design our system thinking about every age group so that the returns can be solved easily.

Acknowledgement

Our consultant in the project, Erdoğan Doğdu, helped us very much and supported us a lot. Thanks to him, we have made our project a bigger site by further improving. Her ideas, his aids are very important to us. Thank you very much our consultant. We are also grateful to our colleagues for their support while doing the project.

References

[1] N.Y. Saiyad,H.B. Prajapati, and V.K. Dabhi. "A survey of Document Clustering using Semantic Approach"International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) - 2016

[2]Biomedical Knowledge Engineering Laboratory, BK21 College of Dentistry, Seoul National University, 28 Yeongeon-dong, Jongro-gu, Seoul 110-810, Republic of Korea

[3] Hai-Tao Zheng, Bo-Yeong Kang, Hong-Gee Kim, "Exploiting noun phrases and semantic relationships for text document clustering" , Journal of Information Sciences, Vol. 179, Issue 13, pp. 2249-2262, Jun. 2009

[4] N.Shah and S. Mahajan "Semantic based Document Clustering: A Detailed Review",International Journal of Computer Applications (0975 – 8887) Volume 52– No.5, August 2012

[5] S Yumusak, E Dogdu, H Kodaz, A Kamilaris, PY Vandenbussche IEICE TRANSACTIONS on Information and Systems 100 (4), 758-767

[6] Wei Song, Cheng Hua Li, Soon Cheol Park, "Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity," Journal of Expert Systems with Applications, Vol. 36, Issue 5, pp. 9095-9104, Jul. 2009

[7] Fahad, S. A., &amp; Yafooz, W. M. (2017). Review on Semantic Document Clustering. International Journal of Contemporary Computer Research, 1(1), 14-30.

[8] T. Berners-Lee, "Linked Data," The World Wide Web Consortium (W3C), 27 07 2006. http://www.w3.org/DesignIssues/LinkedData.html.

[9] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres ve S. Decker, "An empirical survey of Linked Data conformance," Web Semantics: Science, Services and Agents on the World Wide Web, no. 14, pp. 14-44, 2012.

[10] M. S. Marshall, R. Boyce, H. F. Deus, J. Zhao, E. L. Willighagen, M. Samwald, E. Pichler, J. Hajagos, E. Prud'hommeaux ve S. Stephens, "Emerging practices for mapping and linking life sciences data using RDF — A case series," Web Semantics: Science, Services and Agents on the World Wide Web, cilt 14, pp. 2-13, 2012.

[11] C. Bouras and V. Tsogkas. W-kmeans: clustering news articles using wordNet. Knowledge-Based and Intelligent Information and Engineering Systems, pages 379–388, 2010.

[12] C. Bouras and V. Tsogkas. A clustering technique for news articles using WordNet. Knowledge-Based Systems, 36:115 128, 2012.

[13] C.-W. Kim and S. Park. Enhancing text document clustering using nonnegative matrix factorization and wordnet. Journal of information and communication convergence engineering, 11(4):241–246, 2013.

[14] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked data on the web (ldow2008). In Proceedings of the 17th international conference on World Wide Web, pages 1265–1266. ACM, 2008.

[15] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts, pages 205–227, 2009.

[16] S. Yumusak, E. Dogdu, H. Kodaz, and A. Kamilaris. Spend: Linked data sparql endpoints discovery using search engines. arXiv preprint arXiv:1608.02761, 2016.

[17] Stanchev, L. (2016, February). Semantic document clustering using a similarity graph. In Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on (pp. 1-8). IEEE.

[18] M. M.Yücesan and E. Doğdu.(n.d.) New Clustering Using Linked Data Resources and Their Relations, pages 2.

[19] J. H. Lau, and T. Baldwin. “An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation”. Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany, pp. 78–86, 2016.

[20] X. Dai, M. Bikdash and B. Meyer. “From social media to public health surveillance: Word embedding based clustering method for twitter classification”.2017 IEEE 17. International Conferance. April 2017

[21] M. P. Naik, H. B. Prajapati, V. K. Dabhi. “A survey on semantic document clustering” Retrieved August, 2015 Available: http://ieeexplore.ieee.org/abstract/document/7226036/?reload=true