# SEMANTIC DOCUMENT CLASSIFICATION

**Prepared By:** Nihan GÜMÜŞ
Merve UĞUR
Mustafa GÜNGÖR

**Advisor:** Prof. Dr. Erdoğan DOĞDU

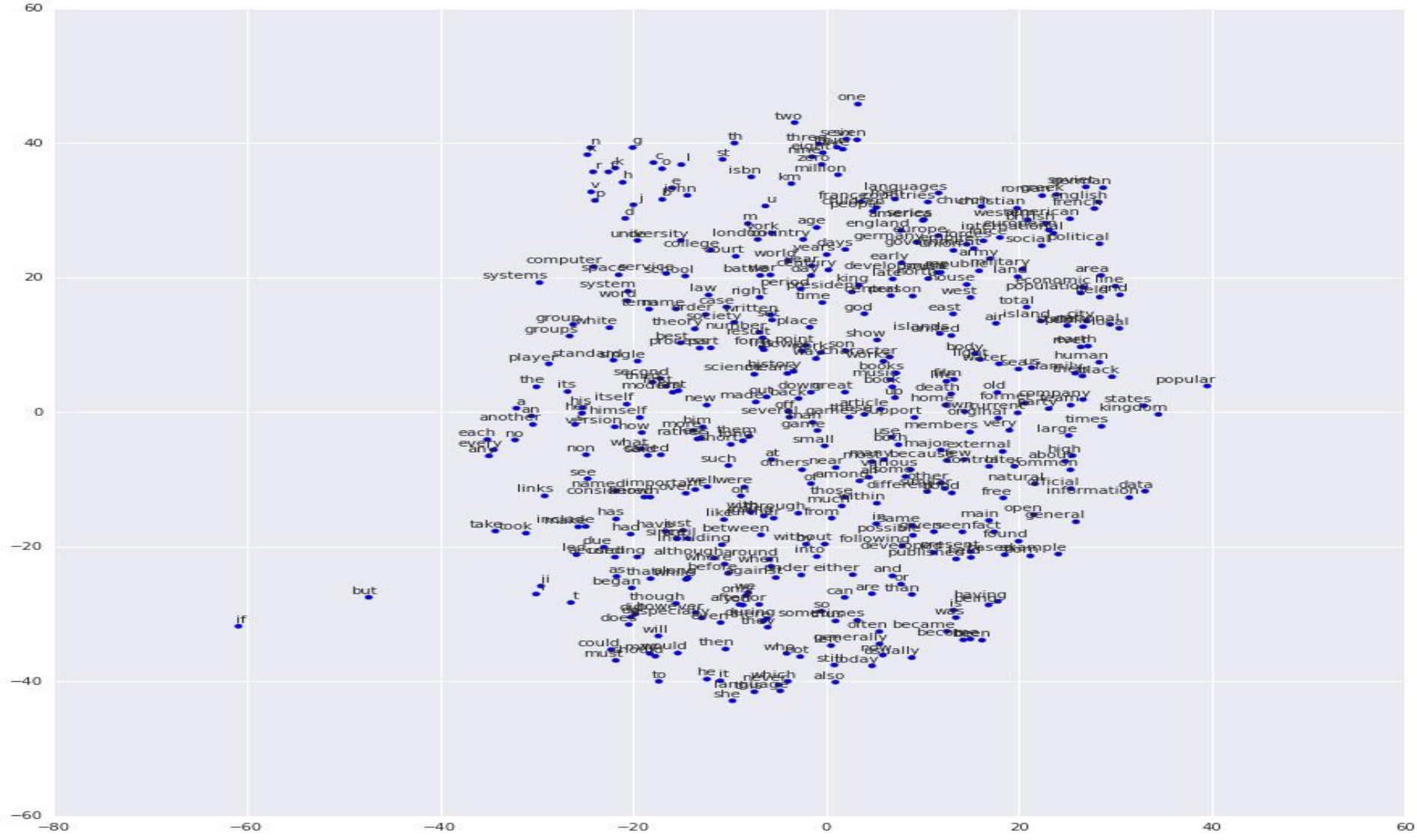
# Contents

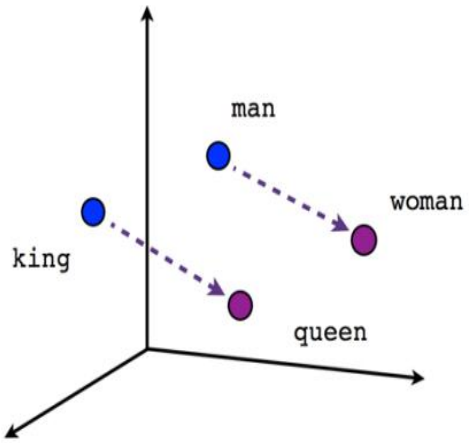
1. Motivation
2. Definition of the Problem
3. Analysis
4. Solution
5. Results and Conclusion
6. References

# Motivation

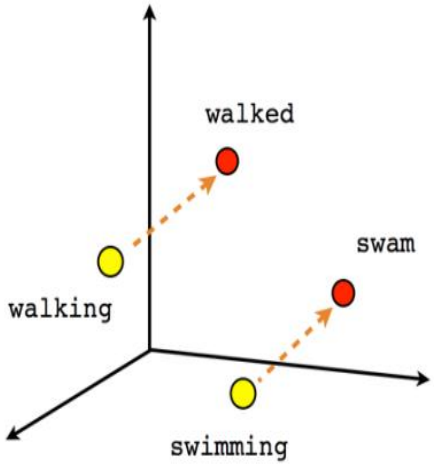
- Design and develop efficient web application.
- Big data learning
- Data reasoning
- Internet of Things (IoT).
- Better understanding the most recent technologies and tools.
- Choose method such as Word embedding,Semantic.

# Word Embedding

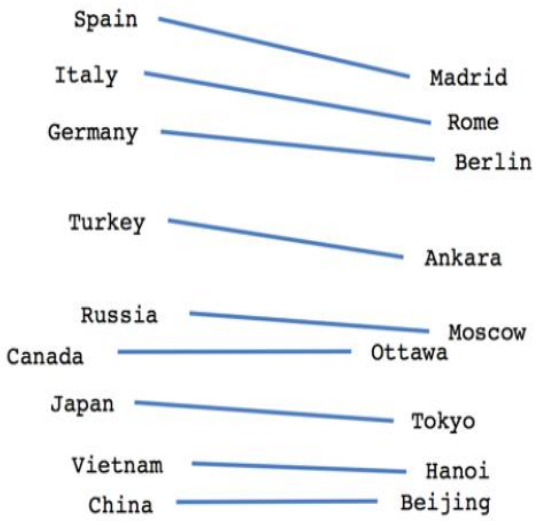





Male-Female




Verb tense

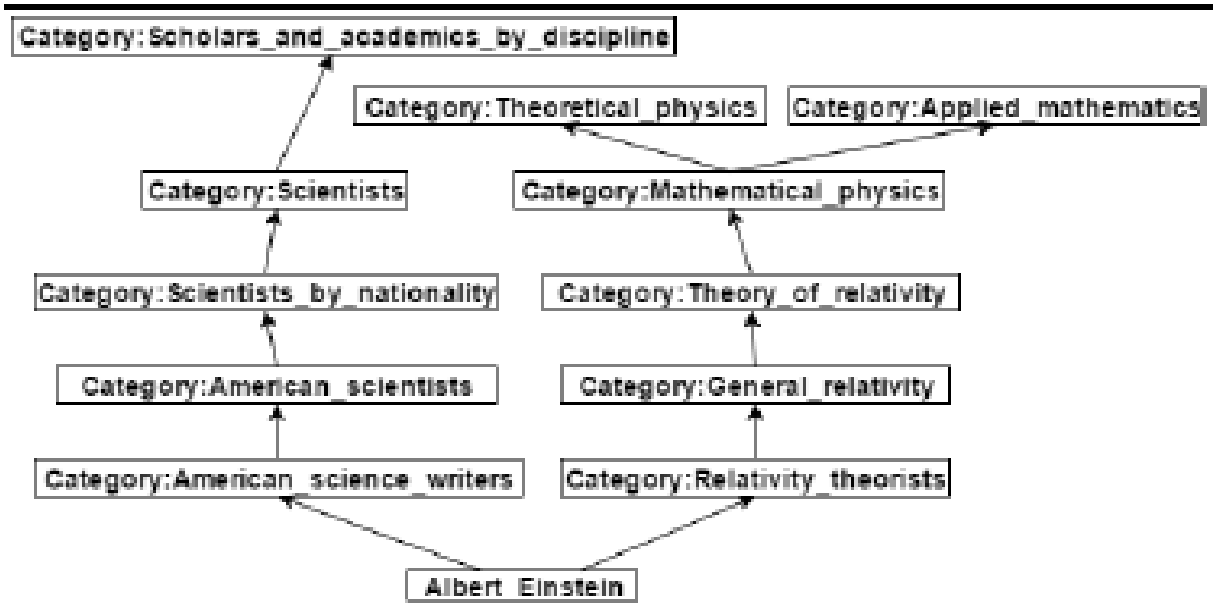



Country-Capital

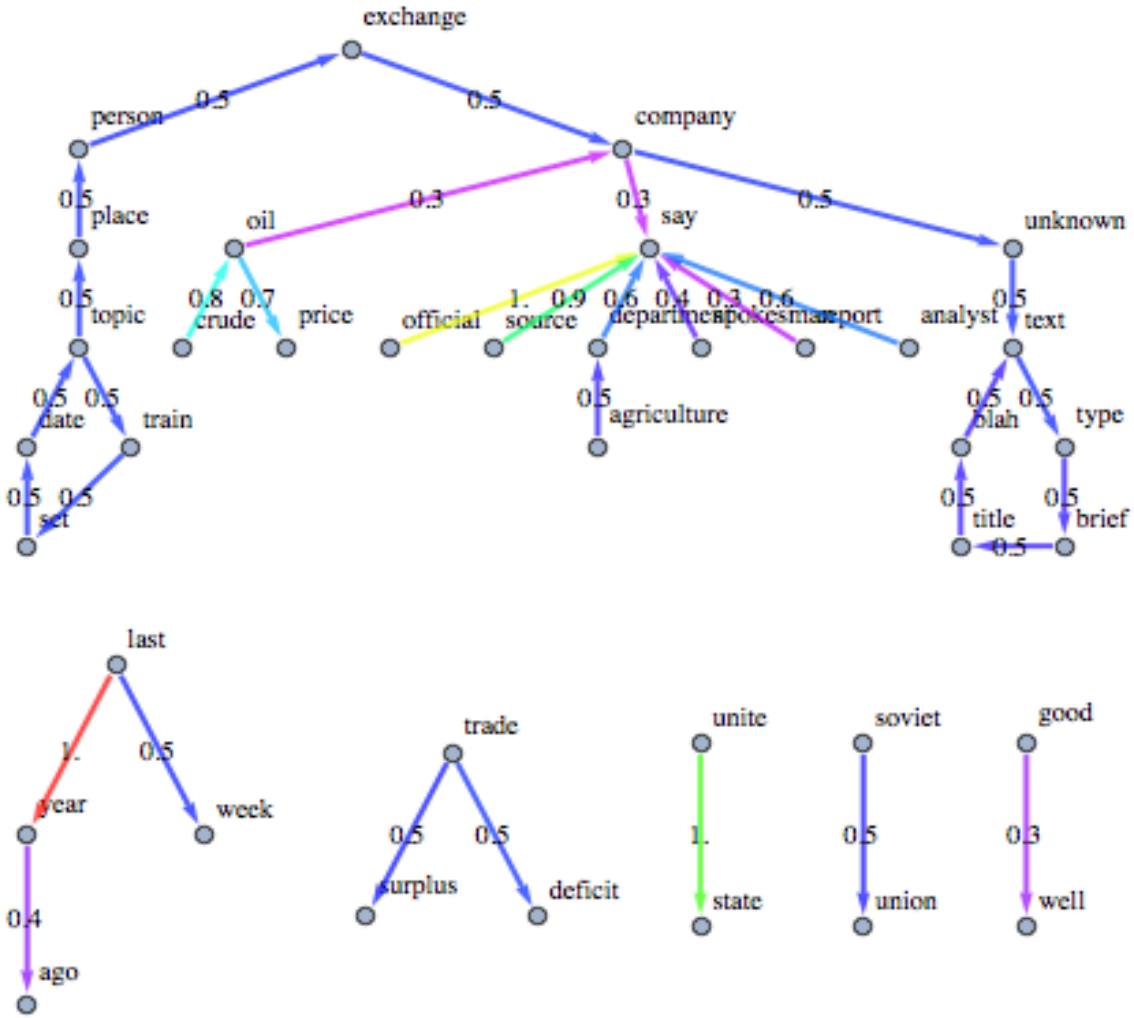
- induced vector space
- semantic relationships.
- male-female,
- verb tense
- even country-capital relationships between words.

# Semantic

- Classify using semantic approach
- Ability to extract concepts from within content and generate the metadata
- Relations between the words are grouped.



- Categorize the words.



# Definition of Problem

- There are wide variety of documents on the web.
- There is a difficulty in intelligently understanding and semantically classifying documents.
- Semantic relationships between words are considered.
- The higher the number of words and phrases with a similar meaning, the cluster grows.

# Analysis

- WordNet relationships.
- Named-entity recognition (NER)-based entity matching techniques.
- Word-embedding.

# Solution

Which technologies will we use in this project?

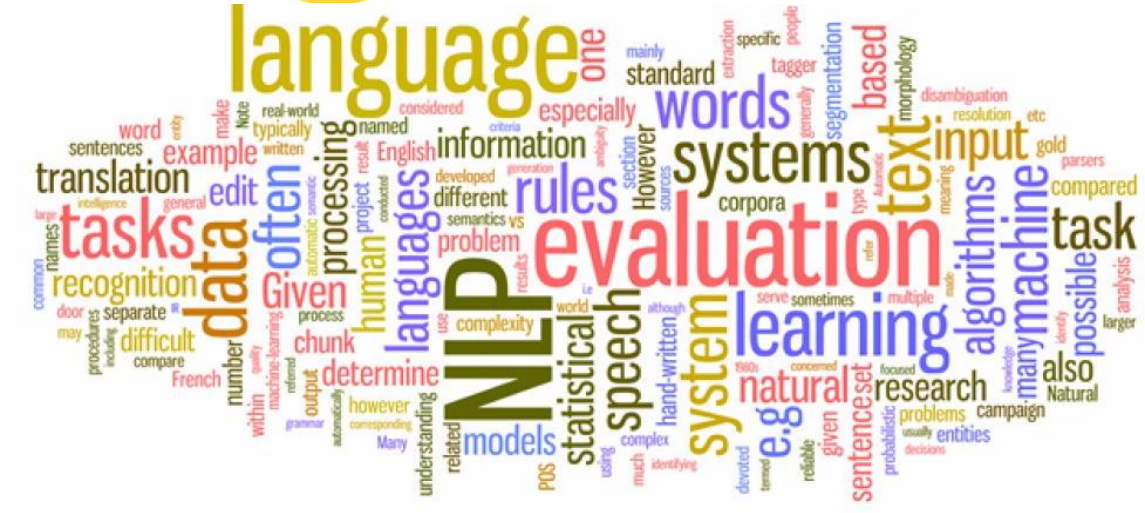
- We use vektor space model and k-means clustering

DBpedia

python

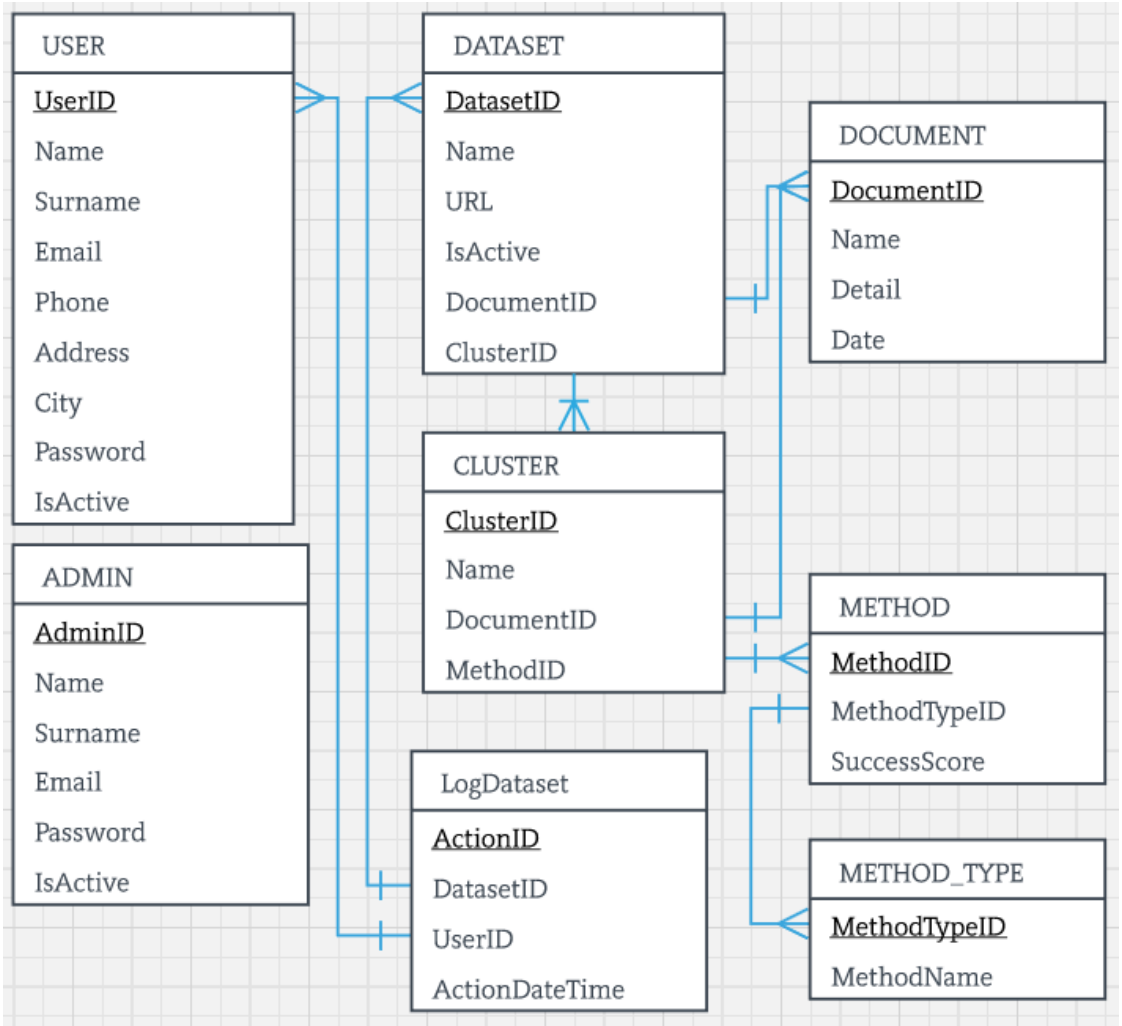
Microsoft SQL Server

Visual C#

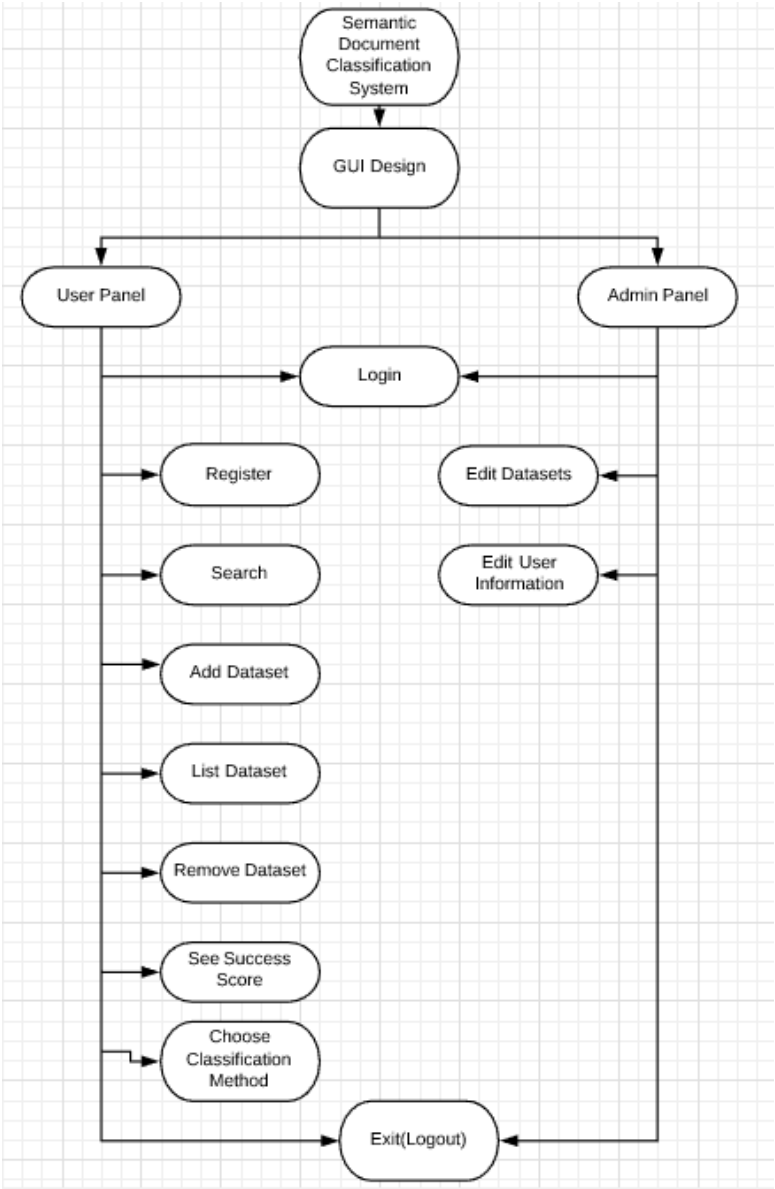


# Database Design

- We will use MSSQL for database in this project with this way, user's information will be stored.

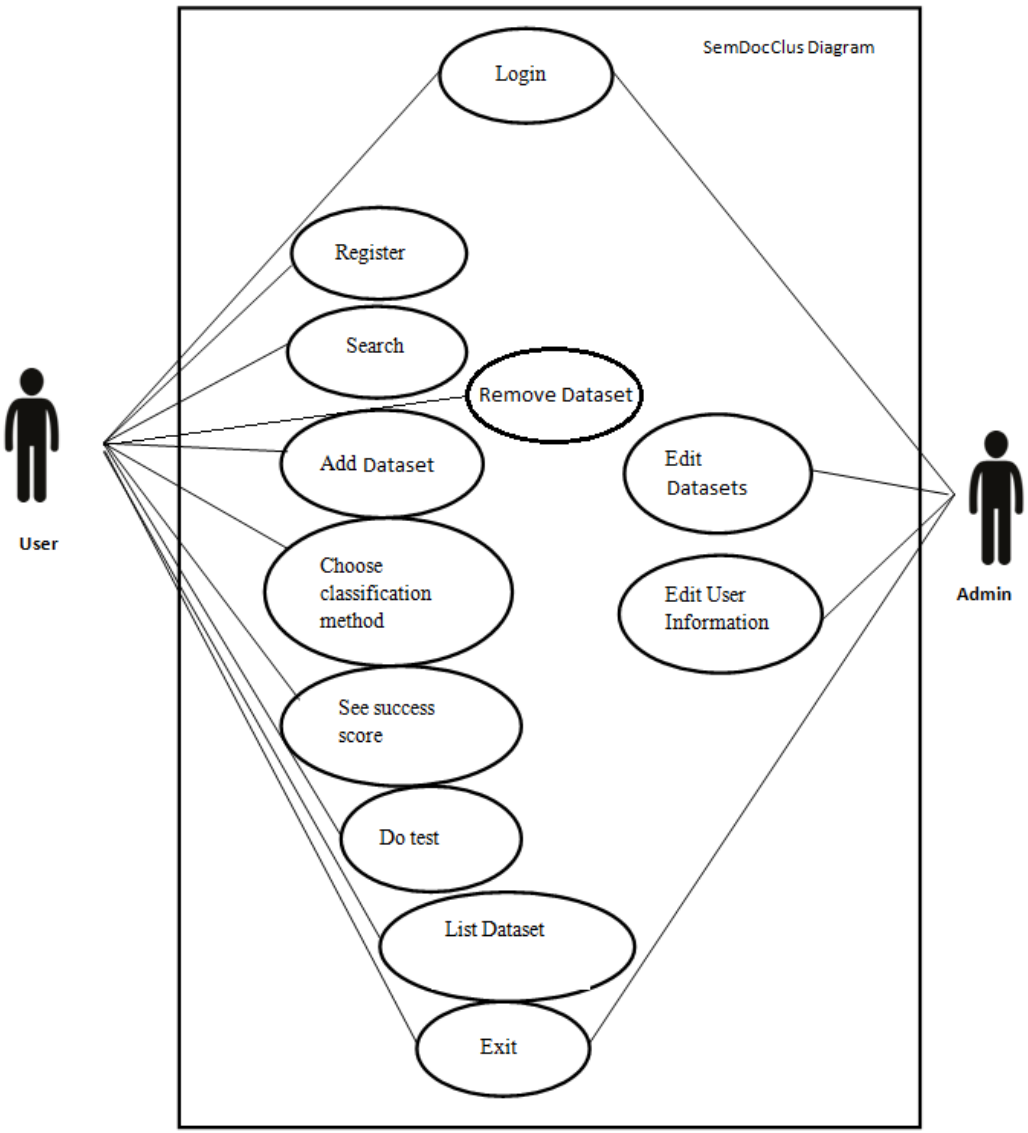


# Block Diagram

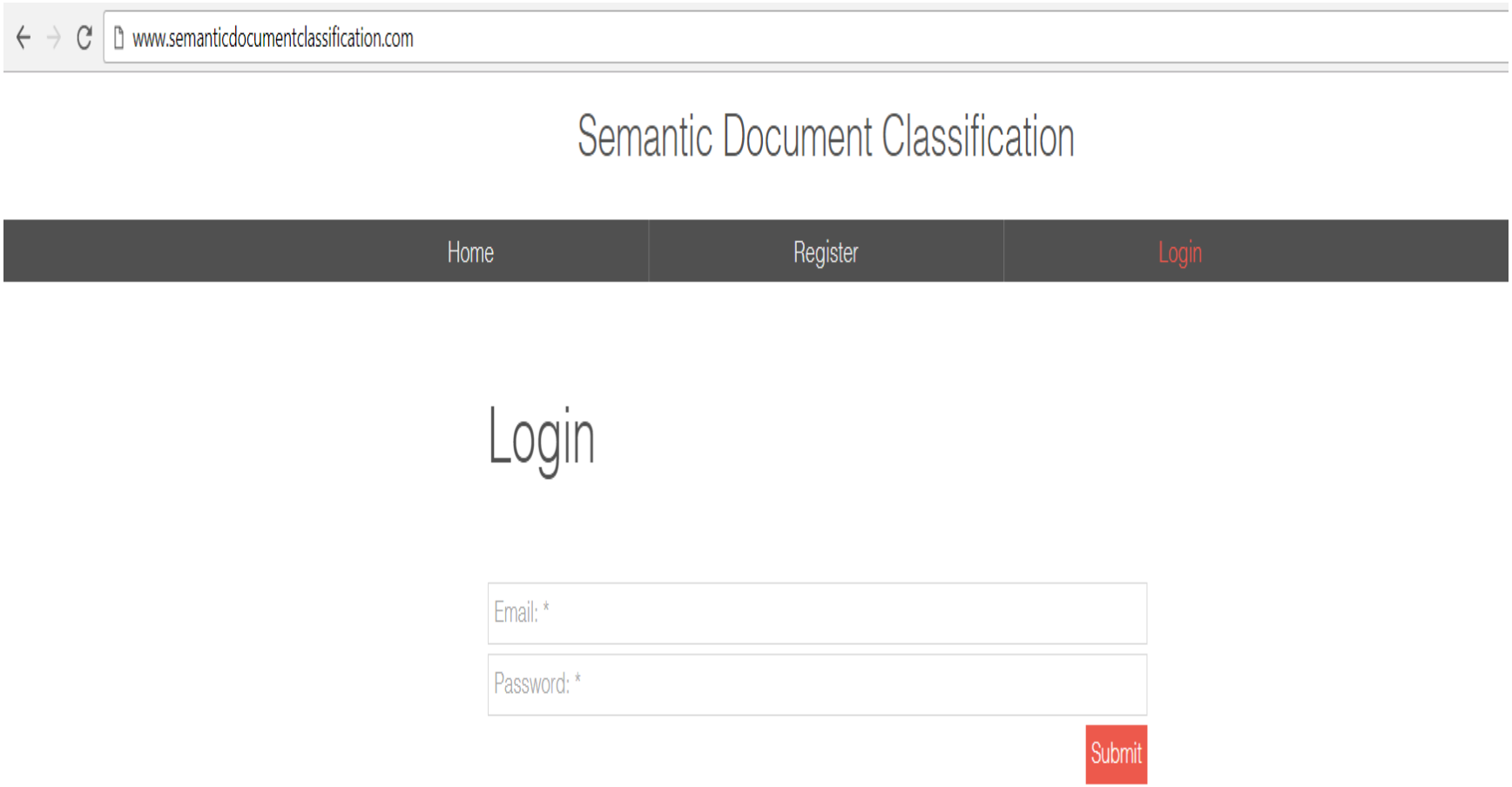


# Use Case Diagram

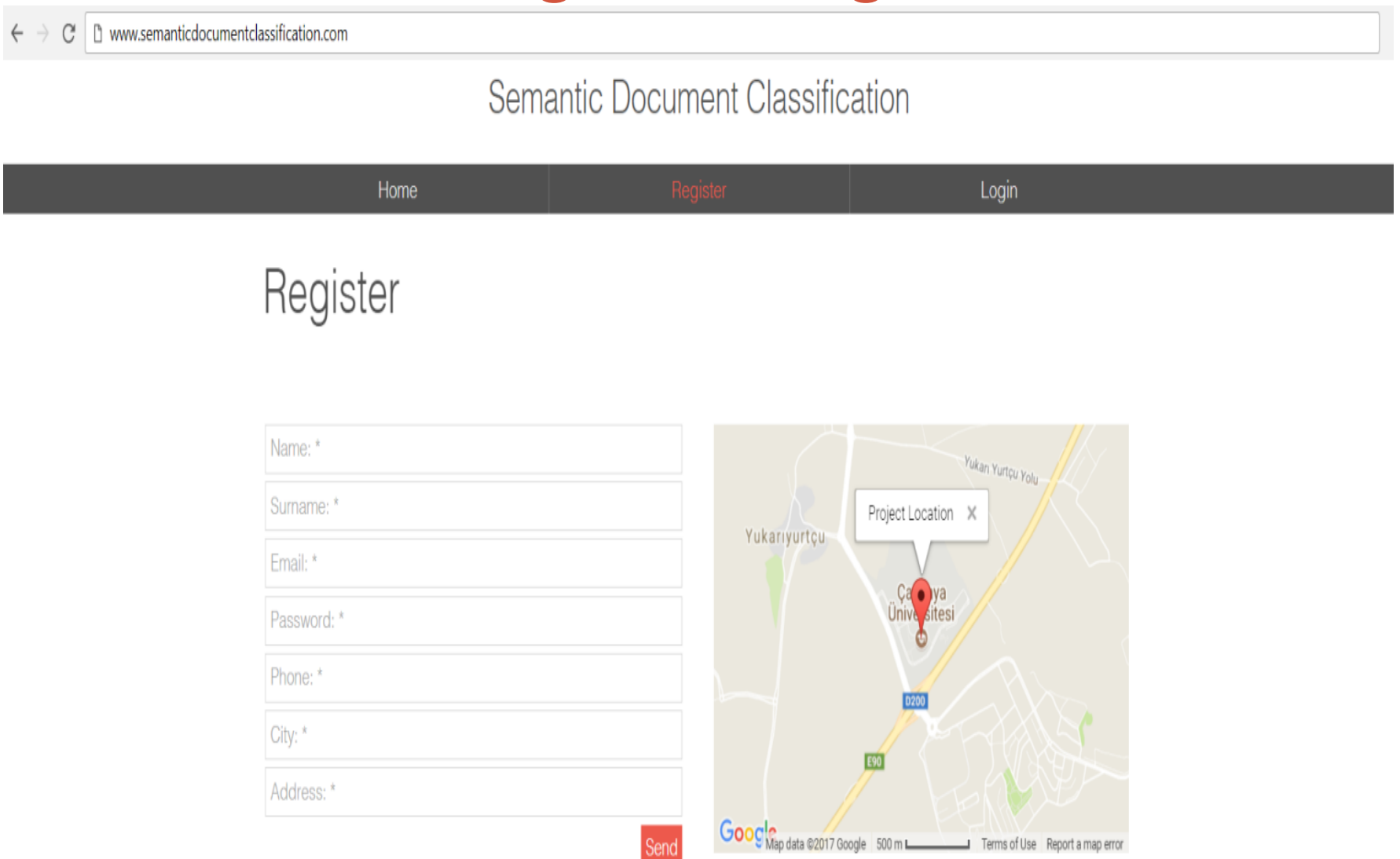
**USE CASE DIAGRAM:**



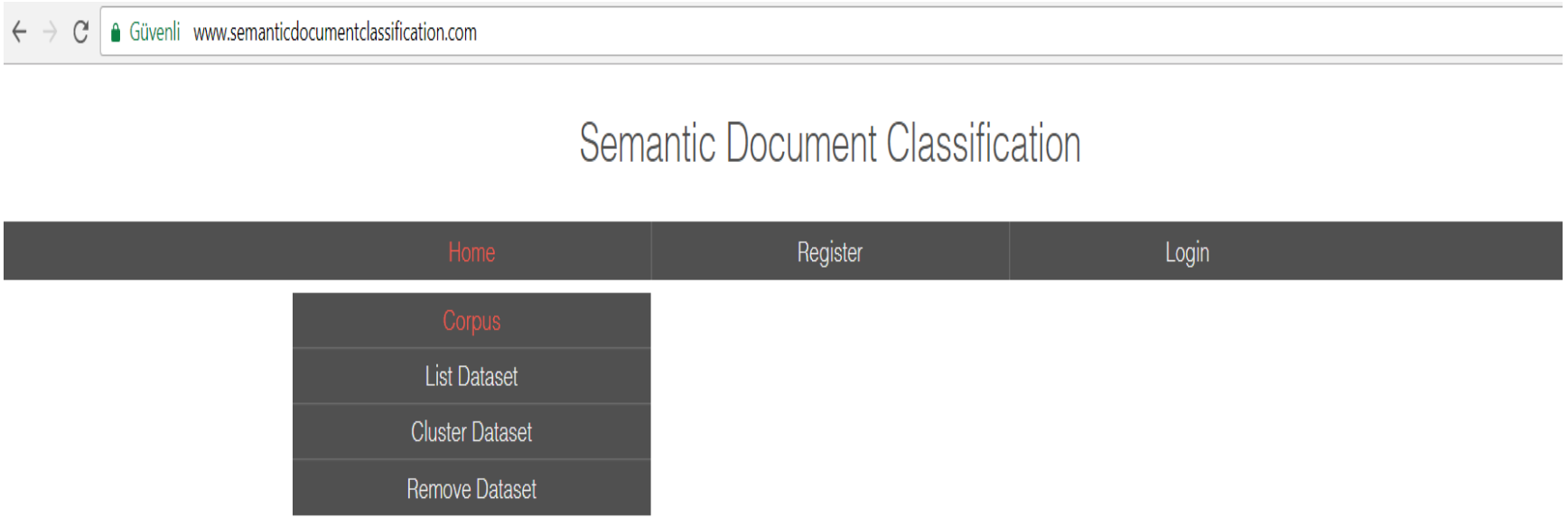
# Login Page

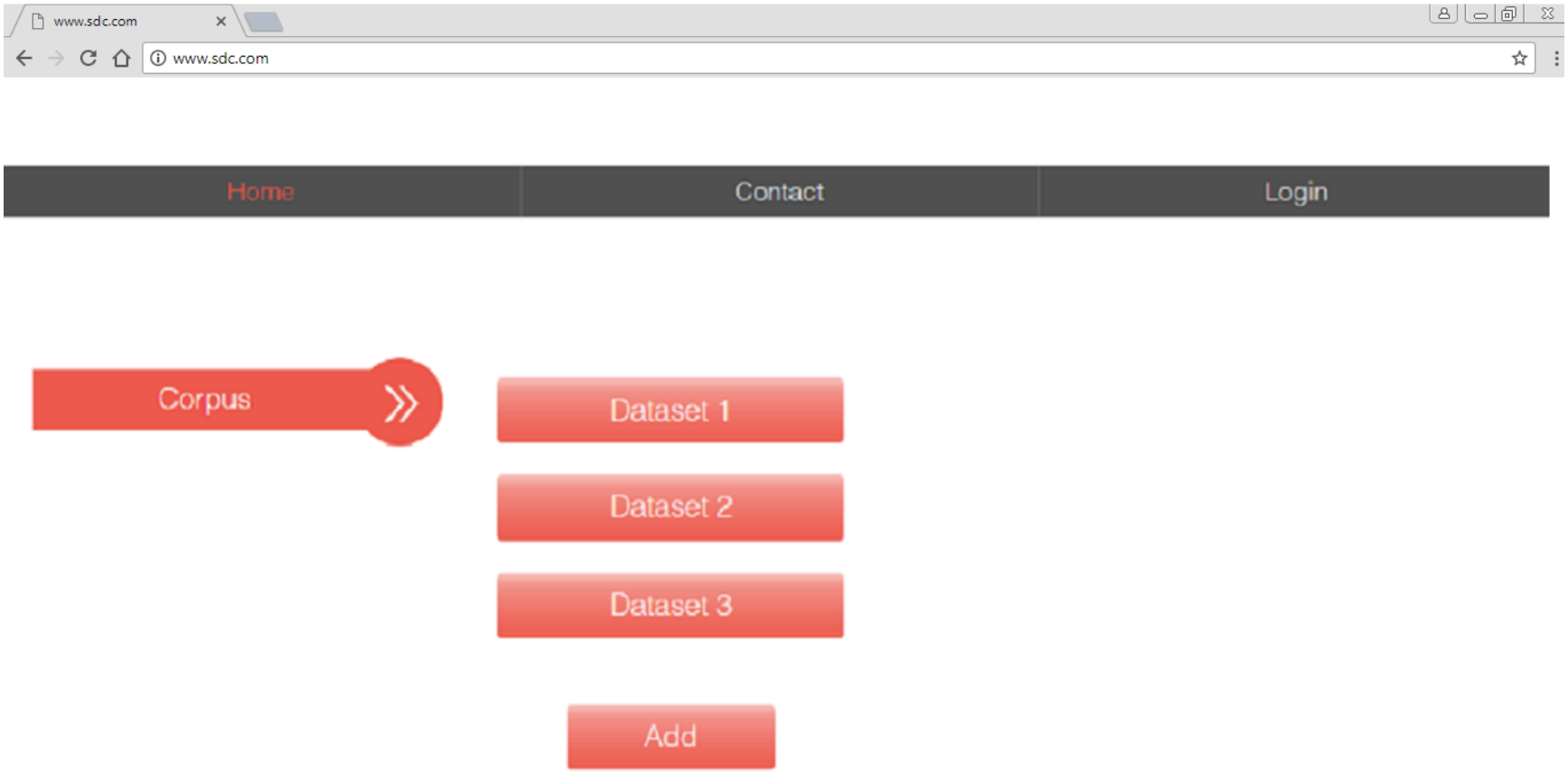

www.semanticdocumentclassification.com
Semantic Document Classification
Home
Register
Login
Login
Email: *
Password: *
Submit

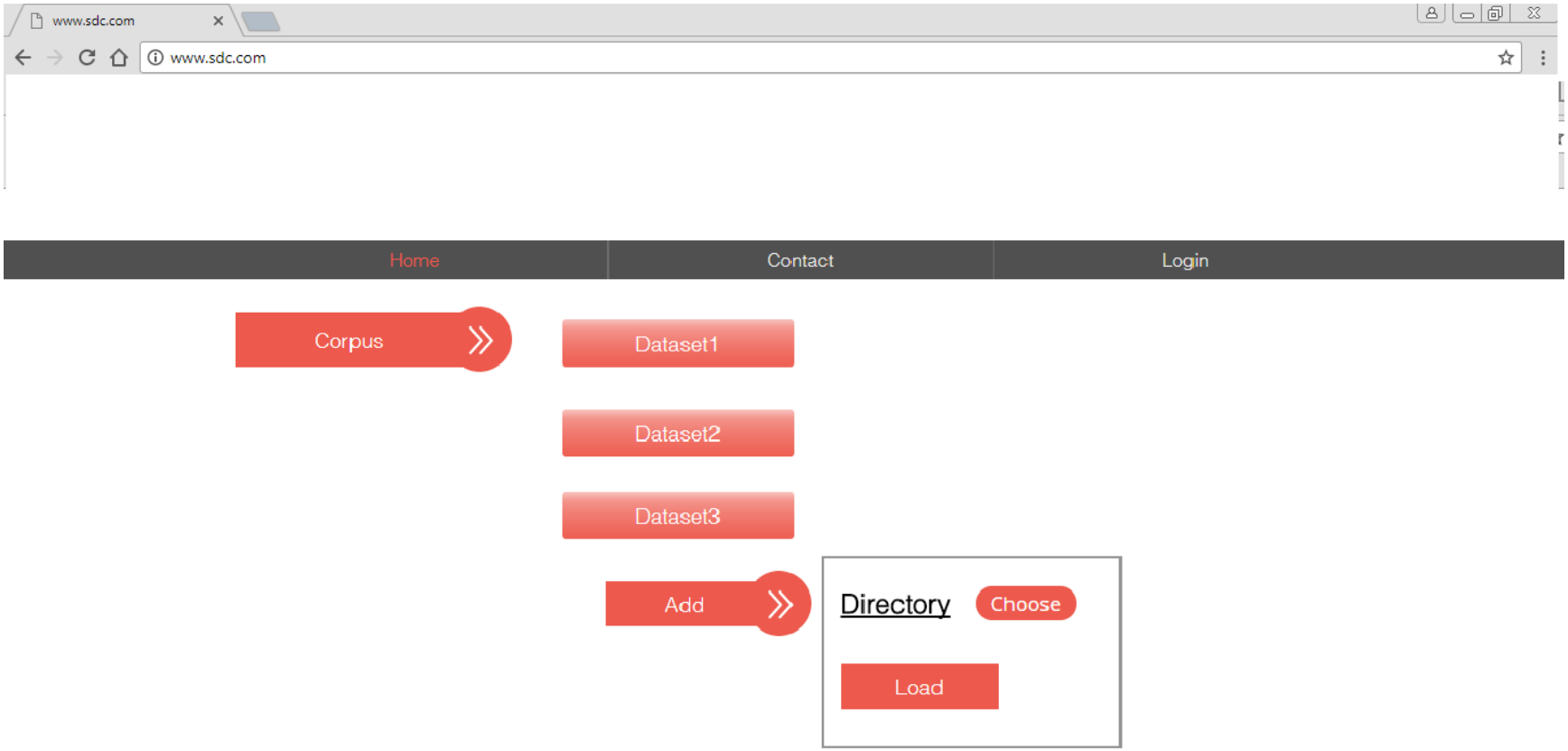
# Register Page

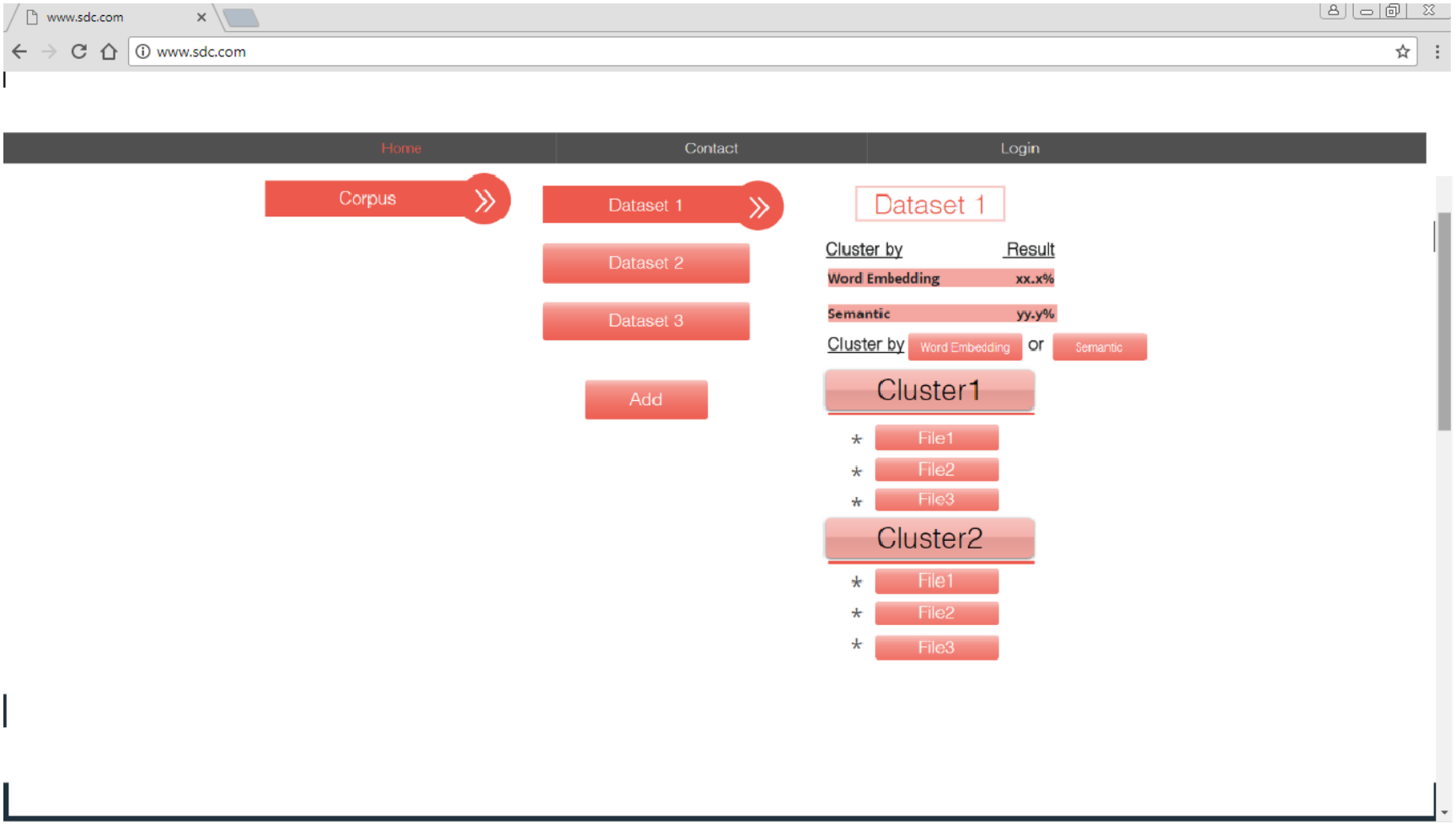


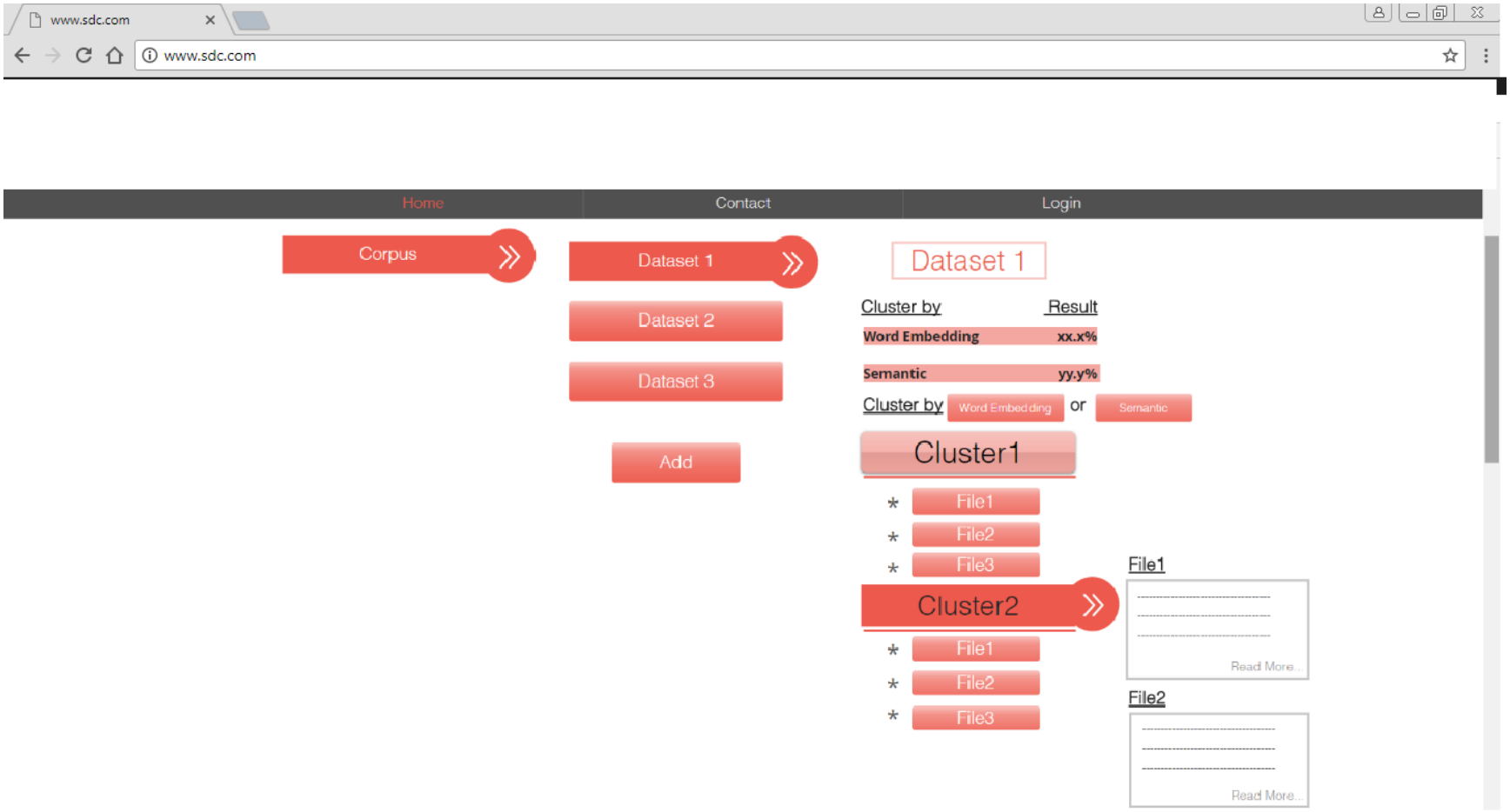
# HomePage

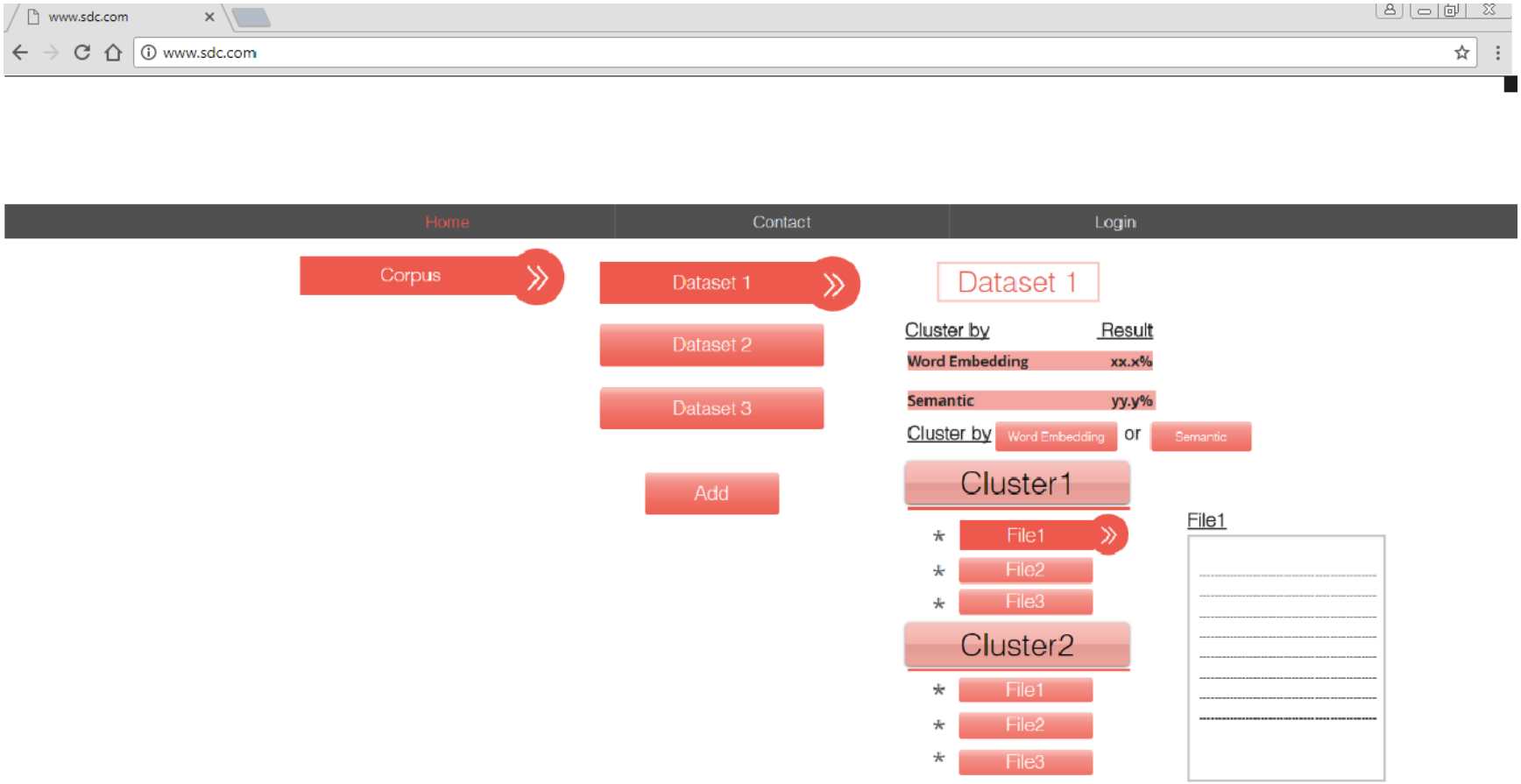

Güvenli www.semanticdocumentclassification.com
Semantic Document Classification
Home
Register
Login
Corpus
List Dataset
Cluster Dataset
Remove Dataset


www.sdc.com
www.sdc.com
Home
Contact
Login
Corpus
Dataset 1
Dataset 2
Dataset 3
Add


www.sdc.com
www.sdc.com
Home
Contact
Login
Corpus
Dataset1
Dataset2
Dataset3
Add
Directory
Choose
Load


www.sdc.com
www.sdc.com
Home
Contact
Login
Corpus
Dataset 1
Dataset 2
Dataset 3
Add
Dataset 1
Cluster by
Result
Word Embedding
xx.x%
Semantic
yy.y%
Cluster by
Word Embedding
or
Semantic
Cluster1
File1
File2
File3
Cluster2
File1
File2
File3


www.sdc.com
www.sdc.com
Home
Contact
Login
Corpus
Dataset 1
Dataset 2
Dataset 3
Add
Dataset 1
Cluster by
Result
Word Embedding
xx.x%
Semantic
yy.y%
Cluster by
Word Embedding
or
Semantic
Cluster1
File1
File2
File3
Cluster2
File1
File2
File3
File1
Read More...
File2
Read More...


www.sdc.com
www.sdc.com
Home
Contact
Login
Corpus
Dataset 1
Dataset 2
Dataset 3
Add
Dataset 1
Cluster by
Result
Word Embedding
xx.x%
Semantic
yy.y%
Cluster by
Word Embedding
or
Semantic
Cluster1
File1
File2
File3
Cluster2
File1
File2
File3
File1

# Results and Conclusions

- Categorize the document set.
- Test and classify documents.
  - *Advantages*
    - Grouping similar features
    - Quick access to classified documents
    - Millions of dataset classifications
  - *Disadvantages*
    - Lack of good test dataset/benchmark

# References


1. M. M.Yücesan and E. Doğdu.(n.d.) New Clustering Using Linked Data Resources and Their Relations

2. 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

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

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


THANK YOU
FOR
LISTENING


Any
Question?