**2IMM15 Web Information Retrival & Data Mining**

Email Filtering and Quering

GROUP 35
Chun Li
Alex Anthis
Kostantinos Messanakis

Project Overview: Using Information Retrieval and Machine Learning techniques to categorize the spam and non-spam emails, query through email dataset, obtain a cluster analysis of the dataset

Data Source: The dataset is formed by AUEB professor I. Androutsopoulos and contains a body of email messages in preprocessed txt format.

Information Retrieval Part

1)Boolean IR
Goal
- Query data
- Indexing
Method
- Store in SQLite

	word	mail	frequency
	Filter	Filter	Filter
1	keplerstrasse	9-314msg1.txt	0
2	schrodt	9-314msg1.txt	0
3	pittner	9-314msg1.txt	0
4	beninca	9-314msg1.txt	0
5	padua	9-314msg1.txt	0

2)Word2vec
Goal
- Produce word embeddings
Method
- Skip-Gram model
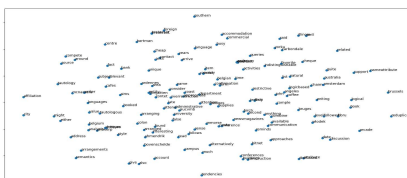
Stop Words & Lemmatization
Goal
- clean data
Method
- NLTK

Workflow
Source Data (2700 mails)

Stop Words & Lemmatization

Boolean IR

TF.IDF Vectorization

Word2vec

RNN

Naive Bayes   Clustering   Logistic Regression
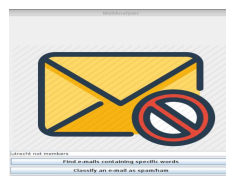
Java Application

Machine Learning Part

1)Naïve Bayes
Goal
- Classify Email spam or ham
- Word frequency
Method
- Term frequency analisys

2)Clustering
Goal
- Group into clusters
- Specify characteristics
Methods
- TF.IDF Vectorization
- K-Means algorithm