CS5560 - Knowledge Discovery and Management
Dr. YugYung Lee

Fri/24/2016

Project Report - 1
Semantic Search Engine

Sudhakar Reddy Peddinti - #33
Rakesh Reddy Bandi - #3

Motivation:

Recent advancements in the world of computer science has led to evolution of Machine Learning algorithms and Data mining capabilities, using which it is possible to build complex cognitive computing agents such as Microsoft's Oxford, Baidu's Minwa, IBM's Watson and Boston dynamics robots. Although each agent works mostly on Artificial Intelligence, each machine is designed to achieve specific goal like Natural Language processing, accurate Image recognition etc., with less or no human involvement. In all these machines, the very common process is the ability to extract information from what humans can perceive - interpret, and cluster them with unsupervised learning, so that machines can also understand and interact using human understandable content. The same can be sentenced in a different way as "machines are achieving capabilities to realize the Vannevar Bush's concept of Memex". For all this to happen, information retrieval and clustering is the key, and achieving good results in both areas is a difficult task, some of the best algorithms used in extracting text information are Word2Vec, WordNet. Self-Organizing Vectors (SOM), K-means are best used to cluster the data.

SigSpace model is developed using SOM clustering algorithm which is aimed at generating feature representation of input data known as signatures. Signatures being smaller in size carries essential information only and then can be used as input for further stages like building knowledge graph or ontology models.

Objectives:

Most of the text information retrieval models like Word2Vec- which represents each term on a axis and the distance between them would represent the similarity, WordNet - which identifies the important concepts present in the document are either based on the term or phrase extraction and the feature data that they generate will be higher in size than the actual data. The novel approach of SigSpace model is designed to output only the necessary information avoiding the unnecessary information. While it is proved that SigSpace will produce almost the same accurate results in matching and classification experiments conducted over image and audio data. This projects aims to test the capabilities over text data and below are some of the objectives:

- Model the SigSpace architecture to generate signatures for Text data.
- Compare and validate the classification results of Signatures generated over text data.
- Efficient modeling and usage of SigSpace concept to generate mutli-level signatures.
- Evaluate the accuracy of classification by updating the signature generation using mediod values instead of mean.
- Using multi-level signatures, build an ontology or Knowledge graph model.

Expected outcomes:

The core idea of this project is to model the SigSpace concept to test the accuracy levels of its usage for text data. While the additional or extended work is to focus on achieving or improving the accuracy results obtained and also to witness the results of using signatures as input and generate multi-level signatures to generate Ontology.

Software Architecture

- **Data set:**

Text data from 20 newsgroups is being used for initial experiments to observe the functioning aspects and to determine the accuracy levels of SigSpace signature generation over text data. The data set consists of 20 newsgroup each containing e-mail communication information. The overall count of documents present is almost equivalent to 20,000 documents, and they are partitioned into 20 different categories. The partition criteria is based on their subject topic, some groups are very similar while some are entirely different from each other.

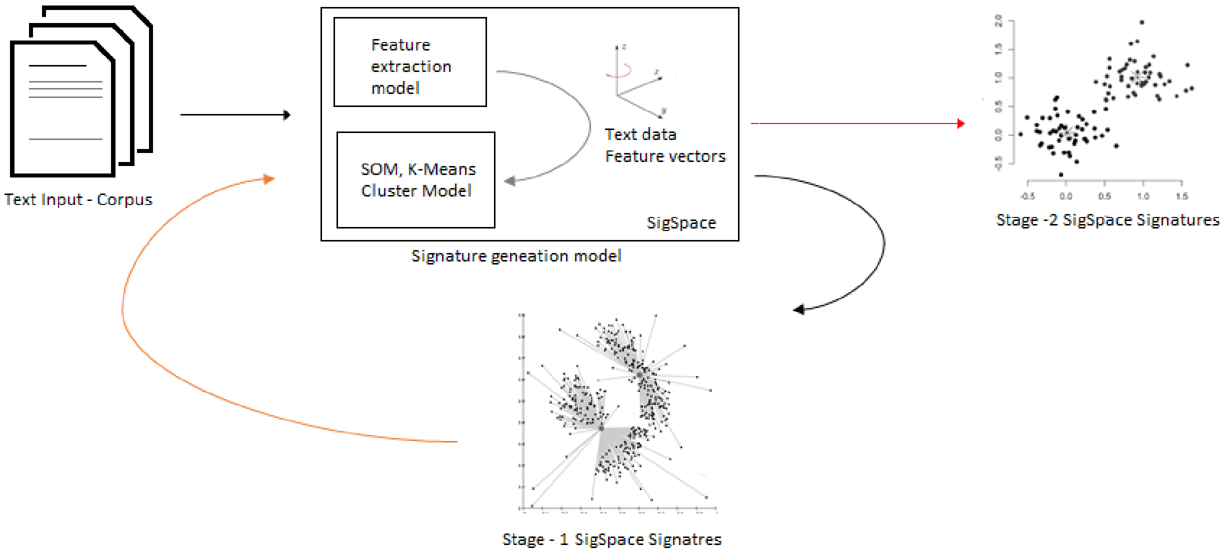

Text Input - Corpus
Feature extraction model
SOM, K-Means Cluster Model
Text data Feature vectors
SigSpace
Signature geneation model
Stage -2 SigSpace Signatures
Stage - 1 SigSpace Signatres

Figure 1: Signature Model Architecture

- **Signature generation model:**

This part includes SigSpace model which accepts input data and generate feature data, In our case, for text input data this will generate text data feature vectors. Later these feature data is clustered using SOM or k-means models and their cluster centers are determined which are the actual stage-1 signatures.

For generating multi-level signature, the signatures generated in the previous step is fed as input and multiple such signatures are again clustered to generate stage-2 signatures.

Tasks Completed:

The SigSpace model was already developed and tested on Image and Audio data. So far the tasks completed include, re-evaluating the image data feature extraction using sift, LBP (Local Binary Pattern) and used this data to realize the SigSpace signatures - cluster centers using both k-means and SOM clustering models. Also text data from 20 newsgroups (http://qwone.com/~jason/20Newsgroups/) is used to extract text features. So far we have extracted the feature data using TF-IDF, NLP based TF-IDF and Sentiment analysis. With this data, we are trying to generate SigSpace signatures and expect to classify the test data by this week.

Feature Extraction:

The primary step is extracting features where features are informative and it doesn't have redundant data. We extract the features from text data to transform it to meaning full and informative. We represent the text documents in vector space.

We represent the textual documents with bag of words. The main point usage of natural language processing here is to identify the boundaries from the document and control the aggressive nature of stopwords on calculating term frequency and inverse document frequency while extracting TFIDF features from the text data. We do lemmatization and stopword removal on the text data.

Lemmatization removes the inflectional ending in the word and returns the conventional base form of that word whereas stemming chops the word to its lexical root. By performing the lemmatization on the text data, the words of same form are treated as single word and prevents to add extra column or dimension in the vector space.

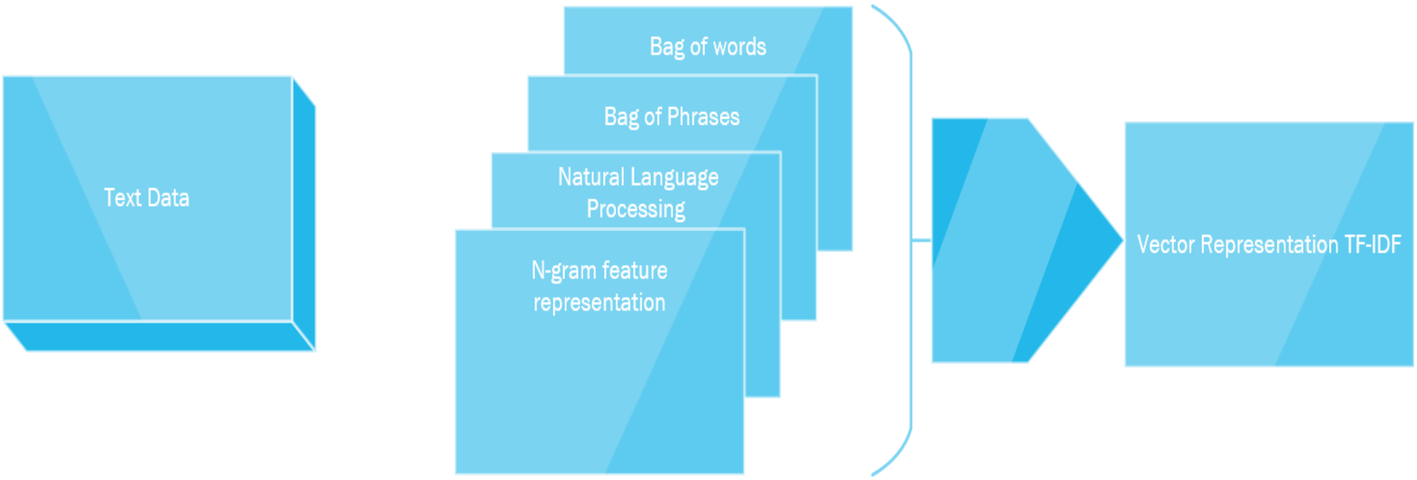

Text Data
Bag of words
Bag of Phrases
Natural Language Processing
N-gram feature representation
Vector Representation TF-IDF

Figure 2: Vector representation of Text Data-Feature Extraction

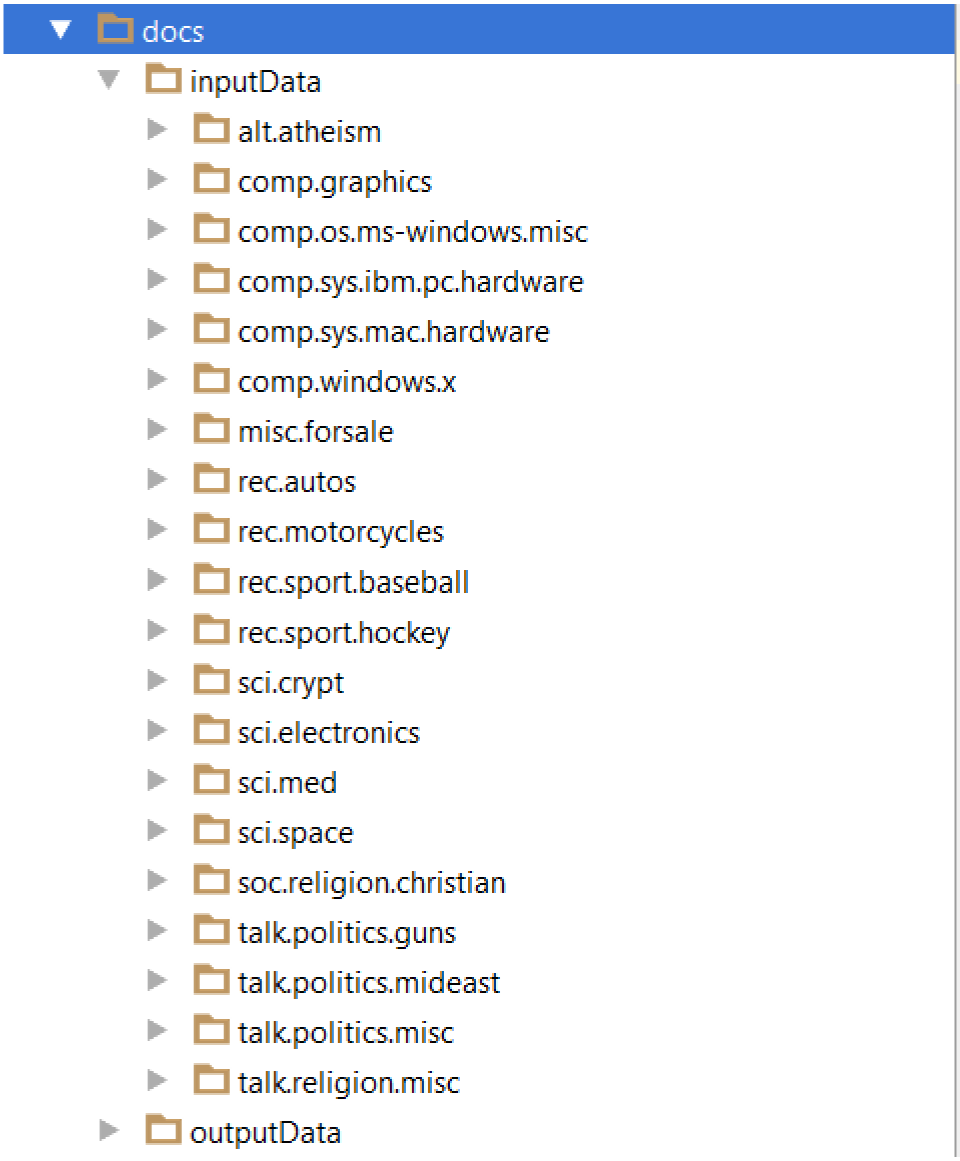
Stop words are common words used. By removing the stop words in the text can focus on understanding the other important words in the document. By using of Stop words removal while extracting the feature vectors from text data prevents the aggressive nature of more frequent stop words.

In our pipeline to extract the features from the test data with Lemmatization and Stop word removal, we have used Standford core NLP.

Additionally, another representation which is based on the “n-gram” model. n-gram is the probabilistic language model which is sequential list of n-words such as words, letter. Using the n-gram vector representation, we can capture the dependencies and word ordering in the documents.

Feature Vector Output Screen shots:

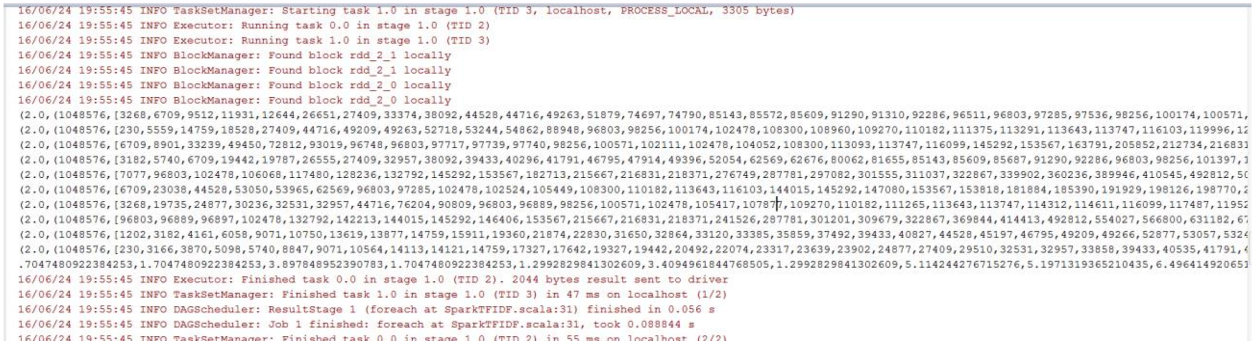
1. We have used 20 news group data set.



```
▼ docs
    ▼ inputData
        ▶ alt.atheism
        ▶ comp.graphics
        ▶ comp.os.ms-windows.misc
        ▶ comp.sys.ibm.pc.hardware
        ▶ comp.sys.mac.hardware
        ▶ comp.windows.x
        ▶ misc.forsale
        ▶ rec.autos
        ▶ rec.motorcycles
        ▶ rec.sport.baseball
        ▶ rec.sport.hockey
        ▶ sci.crypt
        ▶ sci.electronics
        ▶ sci.med
        ▶ sci.space
        ▶ soc.religion.christian
        ▶ talk.politics.guns
        ▶ talk.politics.mideast
        ▶ talk.politics.misc
        ▶ talk.religion.misc
    ▶ outputData
```

2. Feature Vectors of each document in the set of class comp.graphics

The below feature vector of each document is extracted by using the TF-IDF using NLP.



```
16/06/24 19:55:45 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3, localhost, PROCESS_LOCAL, 3305 bytes)
16/06/24 19:55:45 INFO Executor: Running task 0.0 in stage 1.0 (TID 2)
16/06/24 19:55:45 INFO Executor: Running task 1.0 in stage 1.0 (TID 3)
16/06/24 19:55:45 INFO BlockManager: Found block rdd_2_1 locally
16/06/24 19:55:45 INFO BlockManager: Found block rdd_2_1 locally
16/06/24 19:55:45 INFO BlockManager: Found block rdd_2_0 locally
16/06/24 19:55:45 INFO BlockManager: Found block rdd_2_0 locally
(2.0,(1048576,[3268,6709,9512,11931,12644,26651,27409,33374,38092,44528,44716,49263,51879,74697,74790,85143,85572,85609,91290,91310,92286,96511,96803,97285,97536,98256,100174,100571,
(2.0,(1048576,[230,5559,14759,18528,27409,44716,49209,49263,52718,53244,54862,88948,96803,98256,100174,102478,108300,108960,109270,110182,111375,113291,113643,113747,116103,119996,12
(2.0,(1048576,[6709,8901,33239,49450,72812,93019,96748,96803,97717,97739,97740,98256,100571,102111,102478,104052,108300,113093,113747,116099,145292,153567,163791,205852,212734,216831
(2.0,(1048576,[3182,5740,6709,19442,19787,26555,27409,32957,38092,39433,40296,41791,46795,47914,49396,52054,62569,62676,80062,81655,85143,85609,85687,91290,92286,96803,98256,101397,1
(2.0,(1048576,[7077,96803,102478,106068,117480,128236,132792,145292,153567,182713,215667,216831,218371,276749,287781,297082,301555,311037,322867,339902,360236,389946,410545,492812,50
(2.0,(1048576,[6709,23038,44528,53050,53965,62569,96803,97285,102478,102524,105449,108300,110182,113643,116103,144015,145292,147080,153567,153818,181884,185390,191929,198126,198770,2
(2.0,(1048576,[3268,19735,24877,30236,32531,32957,44716,76204,90809,96803,96889,98256,100571,102478,105417,107877,109270,110182,111265,113643,113747,114312,114611,116099,117487,11952
(2.0,(1048576,[96803,96889,96897,102478,132792,142213,144015,145292,146406,153567,215667,216831,218371,241526,287781,301201,309679,322867,369844,414413,492812,554027,566800,631182,67
(2.0,(1048576,[1202,3182,4161,6058,9071,10750,13619,13877,14759,15911,19360,21874,22830,31650,32864,33120,33385,35859,37492,39433,40827,44528,45197,46795,49209,49266,52877,53057,532
(2.0,(1048576,[230,3166,3870,5098,5740,8847,9071,10564,14113,14121,14759,17327,17642,19327,19442,20492,22074,23317,23639,23902,24877,27409,29510,32531,32957,33858,39433,40535,41791,
.7047480922384253,1.7047480922384253,3.897848952390783,1.7047480922384253,1.2992829841302609,3.4094961844768505,1.2992829841302609,5.114244276715276,5.1971319365210435,6.49641492065
16/06/24 19:55:45 INFO Executor: Finished task 0.0 in stage 1.0 (TID 2). 2044 bytes result sent to driver
16/06/24 19:55:45 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 47 ms on localhost (1/2)
16/06/24 19:55:45 INFO DAGScheduler: ResultStage 1 (foreach at SparkTFIDF.scala:31) finished in 0.056 s
16/06/24 19:55:45 INFO DAGScheduler: Job 1 finished: foreach at SparkTFIDF.scala:31, took 0.088844 s
16/06/24 19:55:45 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 55 ms on localhost (2/2)
```

Referecences:

1. “Self Organizing Map -based Document Clustering Using WordNet Ontologies”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012
2. “A SOM-based document clustering using phrases” J. Bakus ; Dept. of Syst. Design Eng, Waterloo Univ., Ont., Canada
3. “Concept-based clustering of textual documents using SOM”, Abdelmalek Amine ; EEDIS Laboratory, Department of computer science, Djillali Liabes University, Sidi Belabbes - Algeria
4. “A spatial user interface to the astronomical literature” P. Poincot, S. Lesteven, and F. Murtagh”, Faculty of Informatics, University of Ulster, Londonderry BT48 7JL, Northern Ireland
5. “Ontology in Text Mining To Cluster Research Project Proposals”, Prof. Pankaj Chandr, Bharat Vishe, Hemant Vishe, Pralhad Lengule, Ankush Shah, Department Of Computer Engineering, Sharadchandra Pawar College of Engineering, Otur, Pune

6. “A Scalable and Dynamic Self-Organizing Map for Clustering Large Volumes of Text Data”, Sumith Matharage, Hiran Ganegedara and Damminda Alahakoon
7. http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html