International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 11 | Nov 2018 www.irjet.net p-ISSN: 2395-0072

Towards Efficient Framework for Semantic Query Search Engine in Large-Scale Data Collection

JANGA NAGARAJU[1], BIMAL KUMAR[2], S AKHILENDRANATH[3], P GANGADHAR[4]

[1]M.TECH CSE SCHOLAR, SSSISE, ANANTAPUR
[2]ASSOCIATE PROFESSOR, CSE DEPARTMENT, SSSISE, ANANTAPUR
[3,4]ASSISTANT PROFESSOR, CSE DEPARTMENT, SSSISE, ANANTAPUR

ABSTRACT - Clustering short texts (such as news titles) by their meaning is a challenging task. The semantic hashing approach encodes the meaning of a text into a compact binary code. Thus, to tell if two texts have similar meanings, we only need to check if they have similar codes. The encoding is created by a deep neural network, which is trained on texts represented by word-count vectors (bag-of-word representation).

Unfortunately, for short texts such as search queries, tweets, or news titles, such representations are insufficient to capture the underlying semantics. To cluster short texts by their meanings, we propose to add more semantic signals to short texts. Specifically, for each term in a short text, we obtain its concepts and co-occurring terms from a probabilistic knowledge base to enrich the short text. Furthermore, we introduce a simplified deep learning network consisting of 3-layer stacked auto-encoders for semantic hashing. Comprehensive experiments show that, with more semantic signals, our simplified deep learning model is able to capture the semantics of short texts, which enables a variety of applications including short text retrieval, classification, and general purpose text processing.

INTRODUCTION:

1.1 Data Mining:

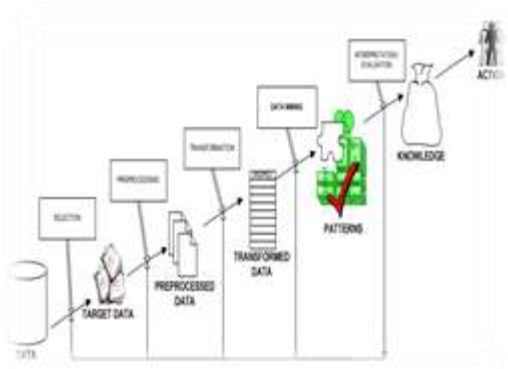


Fig: 1 Structure of Data Mining

Generally, records mining (every now and then known as information or information discovery) is the system of reading records from wonderful views and summarizing it into beneficial facts - statistics that can be used to increase revenue, cuts expenses, or each.

Generally, any of four types of relationships are sought:

Classes: Stored facts are used to find statistics in predetermined businesses. For instance, a restaurant chain could mine patron buy statistics to determine whilst customers go to and what they generally order. These facts can be used to boom traffic by having each day specials.

Clusters: Stored facts are used to find statistics in predetermined businesses. For instance, a restaurant chain could mine patron buy statistics to determine whilst

Customers go to and what they generally order. These facts can be used to boom traffic by having each day specials.

- **Clusters**: Data items are grouped according to logical relationships or patron preferences. For instance, data may be mined to become aware of marketplace segments or consumer affinities.

1.2 Elements of data mining

Data mining consists of five major elements:

1) Extract, transform, and cargo transaction data onto the statistics warehouse machine.

2) Store and manipulate the data in a multidimensional database system.

3) Provide data access to commercial enterprise analysts and statistics era professionals.

4) Analyze the statistics by software.

5) Present the records in a beneficial layout, consisting of a graph or desk.

2. SYSTEM ANALYSIS

2.1 EXISTING SYSTEM:

- Many approaches had been proposed to facilitate brief textual content data by way of enriching the fast textual content.

- More efficiently, a quick text may be enriched with express semantic records derived from external sources consisting of Word Net, Wikipedia, the Open Directory Project (ODP), and many others.

- Salakhutdinov and Hinton proposed a semantic hashing model primarily based on Restricted Boltzmann Machines (RBMs) for lengthy documents, and the experiments showed that their model completed comparable accuracy with the conventional strategies, which include Latent Semantic Analysis (LSA) and TF-IDF.

2.2 DISADVANTAGES OF EXISTING SYSTEM:

- Search-based totally strategies may match nicely for therefore-referred to as head queries, however for tail or unpopular queries, it's miles very likely that some of the top seek effects are irrelevant, which means that the enriched brief text is in all likelihood to contain a lot of noise.

- On the other hand, methods based totally on external sources are restricted through the insurance of those sources. Take Word Net as an instance, Word Net does not incorporate data for correct nouns, which prevents it to recognize entities which include “USA” or “IBM.”

- For ordinary words consisting of “cat”, Word Net includes distinctive statistics approximately its numerous senses. However, lots of the know-how are of linguistic value, and are rarely evoked in day by day utilization. For instance, the feel of “cat” as gossip or female is rarely encountered.

- Unfortunately, Word Net does now not weight senses based on their usage, and these rarely used senses frequently supply upward thrust to misinterpretation of short texts. In precise, without understanding the distribution of the senses, it's far hard to build an inference mechanism to pick out appropriate senses for a word in a context.

2.3 PROPOSED SYSTEM:

- In this project, I endorse a singular approach for know-how short texts.

- This technique A semantic community based totally approach for enriching a brief text.

- I gift a novel mechanism to semantically improve quick texts with both standards and co-taking place phrases, such external information are inferred from a big scale probabilistic expertise base using our proposed thorough techniques.

- For every auto encoder I design a selected and effective mastering method to seize useful features from enters statistics.

- I provide a way to mix know-how statistics and deep neural network for textual content evaluation; in order that it helps machines higher understand quick texts.

2.4 ADVANTAGES OF PROPOSED SYSTEM:

- I carry out enormous experiments on obligations such as records retrieval and classification for brief texts.

- I display considerable upgrades over current strategies, which affirm that

- standards and co-happening phrases efficiently enhance brief texts, and permit higher understanding of them;

- Car-encoder based totally DNN model is capable of seize the summary functions and complex correlations from the enter text such that the learned compact binary codes may be used to represent the means of that text.

3. SYSTEM DESIGN

3.1 SYSTEM ARCHITECTURE

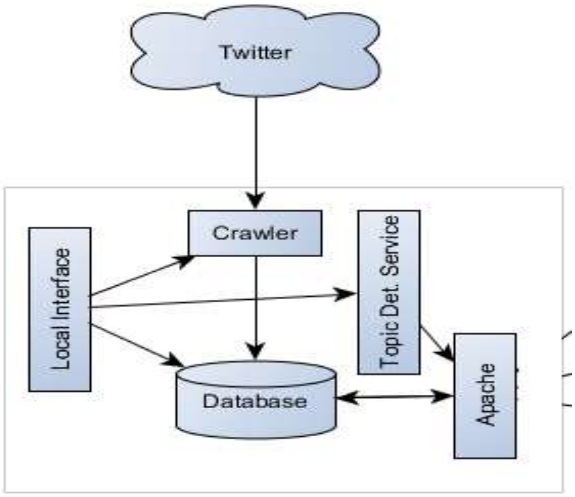


Fig.2: System Architecture

3.2 DATA FLOW DIAGRAM:

1. The DFD is likewise known as bubble chart. It is a simple graphical formalism that can be used to symbolize a machine in phrases of enters facts to the device, various processing achieved in these facts, and the output facts are generated by way of this device.

2. The data flow diagram (DFD) is one of the maximum important modeling gears. It is used to model the machine components. These components are the system procedure, the facts used by the procedure, an outside entity that interacts with the system and the data flows in the machine.

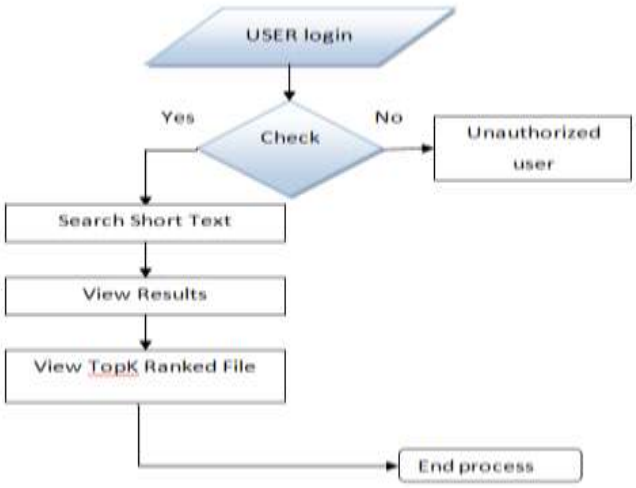


Fig 3: Data Flow Diagram for User Login

3. DFD suggests how the statistics moves via the system and the way it's miles modified through a chain of differences. It is a graphical technique that depicts statistics drift and the differences which might be implemented as facts movements from input to output.

4. DFD is also known as bubble chart. A DFD may be used to represent a device at any stage of abstraction. DFD can be partitioned into ranges that represent increasing data drift and functional element.

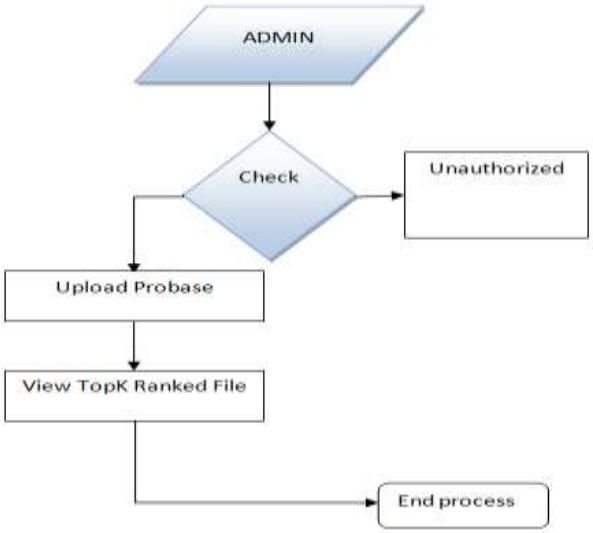


Fig.4 Data Flow Diagram for ADMIN Login

3.3 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized preferred-motive modeling language inside the area of object-orientated software engineering. The general is controlled, and became created by, the Object Management Group.

The goal is for UML to grow to be a commonplace language for creating models of object orientated pc software program. In its modern form UML is constructed from two predominant components: a Meta-model and a notation. In the future, some form of approach or procedure will also be added to; or related to, UML.

GOALS:

Primary goals in the design of the UML are as follows:

1. Provide users a prepared-to-use, expressive visual modeling Language so we can broaden and trade meaningful fashions.

2. Provide extendibility and specialization mechanisms to increase the middle ideas.

3. Be independent of specific programming languages and development system.

3.4 USE CASE DIAGRAM:

A use case diagram inside the Unified Modeling Language (UML) is a form of behavioral diagram defined by means of and created from a Use-case analysis. Its cause is to present a graphical review of the functionality supplied through a machine in phrases of actors, their desires (represented as use cases), and any dependencies between the ones use cases. The main purpose of a use case diagram is to show what system functions are achieved for which actor. Roles of the actors in the machine can be depicted.

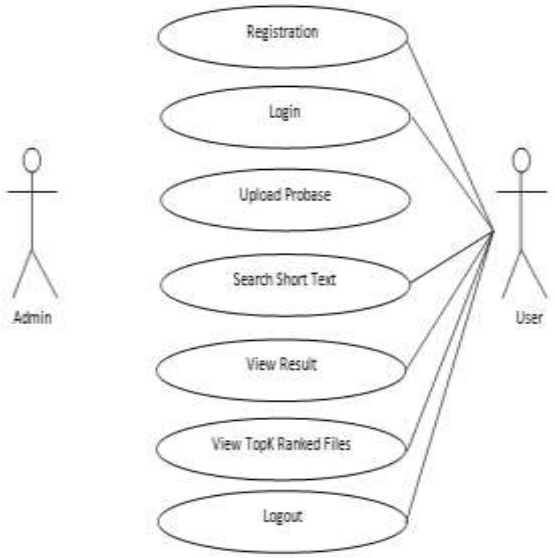


Fig 5 Use Case Diagram for ADMIN and USER Registration

3.5 SEQUENCE DIAGRAM:

A collection diagram in Unified Modeling Language (UML) is a form of interplay diagram that shows how strategies perform with one another and in what order. It is a assemble of a Message Sequence Chart. Sequence diagrams are once in a while referred to as occasion diagrams, occasion scenarios, and timing diagrams.

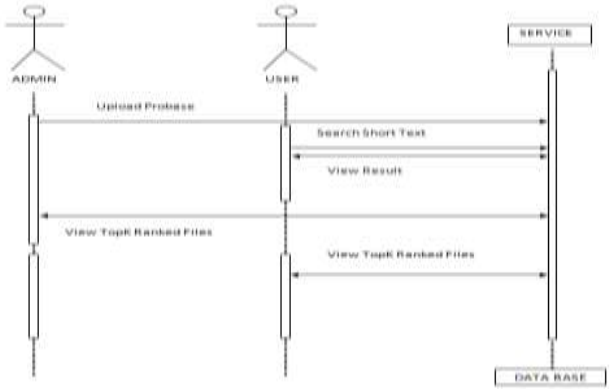


Fig.6 Sequence Diagram for ADMIN and USER

3.6 ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise sports and moves with support for preference, generation and concurrency. In the Unified Modeling Language, hobby diagrams may be used to describe pastime diagram indicates the overall flow of control

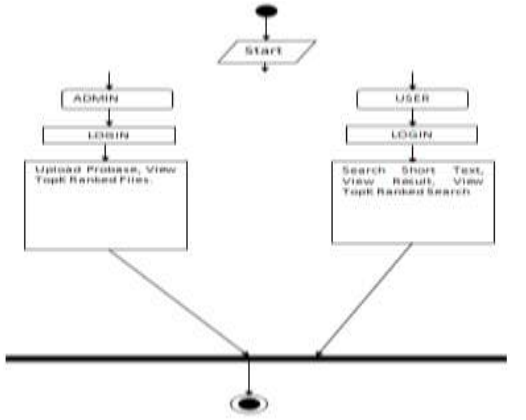


Fig 7 Activity Diagram for ADMIN and USER

4. SYSTEM TESTING

4.1 INTRODUCTION:

The reason of testing is to discover mistakes. Testing is the technique of trying to discover each doable fault or weakness in a work product. There are numerous varieties of take a look at. Each check type addresses a selected testing requirement.

4.2TYPES OF TESTS

4.2.1 UNIT TESTING

Unit trying out entails the layout of test cases that validate that the internal application good judgment is functioning properly, and that application inputs produce legitimate outputs. All selection branches and internal code flow have to be proven. It is the trying out of individual software system of the utility. It's far achieved after the final touch of an character unit earlier than integration. This is a structural trying out, that is based on understanding of its creation and is invasive. Unit exams perform fundamental tests at issue level and test a specific enterprise manner, software, and/or device configuration. Unit assessments make sure that every unique course of a enterprise technique performs correctly to the documented specifications and carries truly defined inputs and predicted consequences.

4.2.2 INTEGRATION TESTING

Integration assessments are designed to check included software program additives to decide in the event that they truly run as one software. Testing is event pushed and is extra worried with the basic outcome of screens or fields. Integration checks exhibit that despite the fact that the components were in my opinion pride, as shown with the aid of effectively unit testing, the aggregate of components is accurate and constant. Integration trying out is in particular geared towards exposing the problems that arise from the combination of additives.

4.2.3 Functional Test

Functional assessments offer systematic demonstrations that capabilities tested are available as designated via the enterprise and technical necessities, machine documentation, and person manuals.

Functional trying out is centered on the following items:

Valid Input: diagnosed classes of valid enter ought to be customary.

Invalid Input: identified training of invalid enters must be rejected.

Functions: identified functions must be exercised.

Output: recognized classes of software outputs have to be exercised.

Systems/Procedures: interfacing systems or methods ought to be invoked.

4.2.4 SYSTEM TESTING

System checking out ensures that the whole integrated software device meets necessities. It assessments a configuration to ensure recognized and predictable consequences. An instance of system checking out is the configuration orientated system integration take a look at. System trying out is based totally on procedure descriptions and flows, emphasizing pre-pushed process hyperlinks and integration factors.

4.2.5 White BOX AND black box testing

White box testing

White Box Testing is a checking out in which in which the software tester has knowledge of the inner workings, shape and language of the software, or at the least its

cause. It is cause. It is used to test areas that cannot be reached from a black box stage.

Black Box Testing

Black Box Testing is testing the software with none understanding of the inner workings, structure or language of the module being tested. Black container assessments, as maximum other types of exams, should be written from a definitive source report, which includes specification or requirements file, together with specification or requirements document. It is a trying out wherein the software program under test is treated, as a black container .You cannot "see" into it. The test presents inputs and responds to outputs without considering how the software works.

Unit Testing:

Unit testing is usually performed as part of a mixed code and unit take a look at section of the software program lifecycle, although it is not unusual for coding and unit checking out to be performed as two awesome levels.

Integration Testing:

Software integration checking out is the incremental integration checking out of or more included software components on a single platform to supply failures as a result of interface defects.

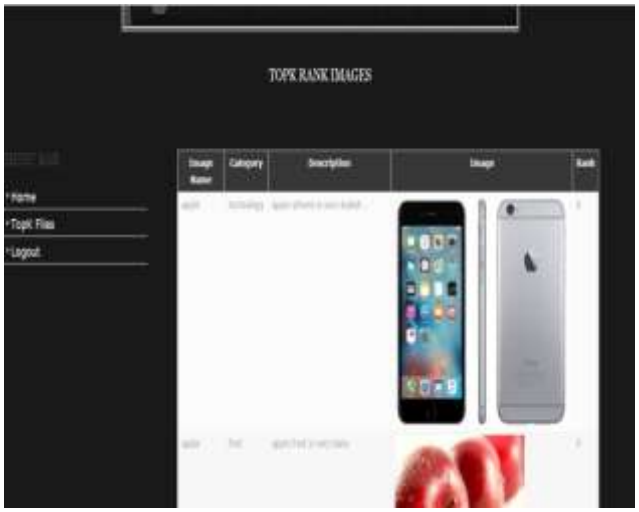
The undertaking of the mixing check is to test those components or software packages, e.g. Components in a software program system or – one step up – software program packages at the employer degree – engage without errors.

Acceptance Testing:

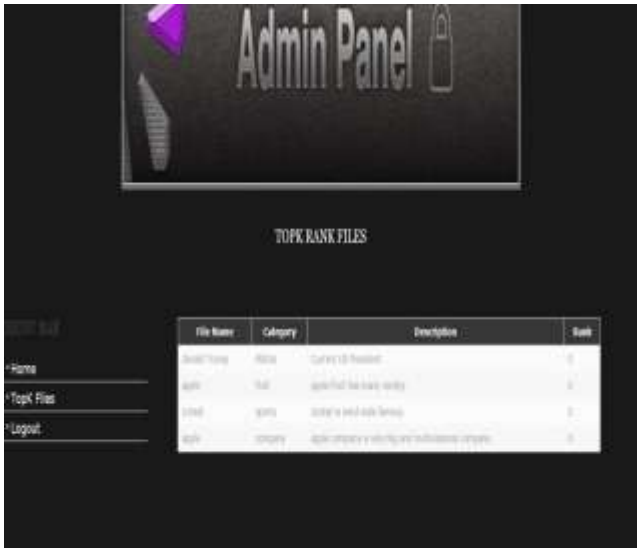
User Acceptance Testing is an essential phase of any undertaking and calls for huge participation via the stop person. It additionally ensures that the machine meets the functional requirements.

APPENDIX

Description: The above figure shows the ADMIN PANEL where ADMIN can Login and Reset.

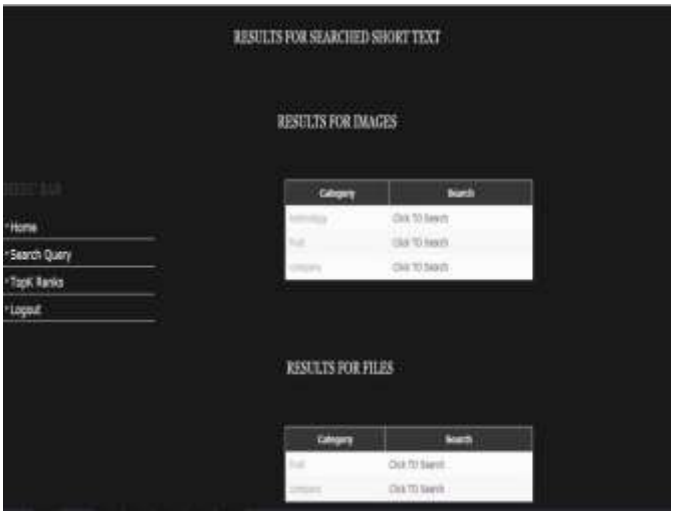


Description: The above figure shows the TOPK images.



Description: The above figure shows TOPK Rank Files, file name such as Donald, apple…

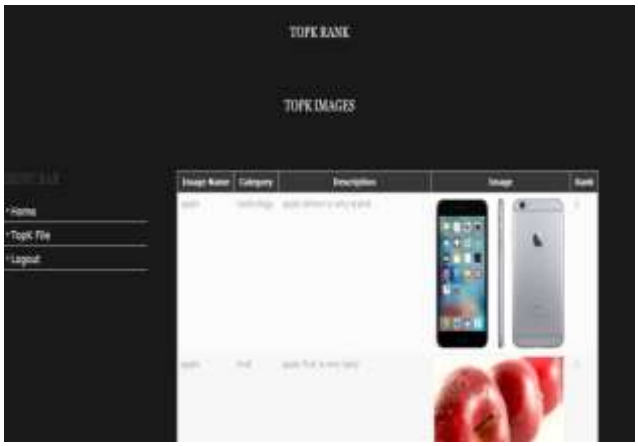
Description: The above figure shows the Search Short Text Query, enter apple in search query.



Description: The above figure shows the results for apple files and images.

Description: The above figure shows results for short text and result for images, image name apple.

Description: The above figure shows the Results for the searched file apple.



Description: Results for TOPK images apple.

CONCLUSION:

In this paper, a novel approach for understanding short texts is proposed. First, I introduce a mechanism to enrich brief texts with concepts and co-happening phrases which might be extracted from a probabilistic semantic network, called Probase. After that, each brief textual content is represented as a 3,000- dimensional semantic function vector. I then layout an extra efficient deep gaining knowledge of model, which is stacked by way of 3 auto-encoders with specific and effective learning functions, to do semantic hashing on those semantic function vectors for quick texts. A -degree semi-supervised education method is proposed to optimize the model such that it could capture the correlation ships and abstract features from brief texts. When training is performed, the output is threshold to be a 128-dimensional binary code that is seemed as a semantic hashing code for that enter text. I perform comprehensive experiments on quick text focused responsibilities which include data retrieval and type. The great improvements on each obligations show that our enrichment mechanism ought to efficaciously enhance brief textual content representations and the proposed car-encoder primarily based deep getting to know version is able to encode complicated features from input into the compact binary codes

FUTURE WORK:

In our future work, I will explore the following points. Double approach for know-how short texts**.** Future work will attempt to enhance security while ensuring reasonable quality of service even with multiple users logged on the system at the same time.

REFERENCES

[1] M. Sahami and T. D. Heilman, “A web-based kernel function for measuring the similarity of short text

snippets," in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 377–386.

[2] W. tau Yih and C. Meek, "Improving similarity measures for short segments of text," in Proc. 22nd Nat. Conf. Artif. Intell., 2007, pp. 1489–1494.

[3] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang, "Query enrichment for web-query classification," ACM Trans. Inf. Syst., vol. 24, no. 3, pp. 320–352, 2006.

[4] C. Fellbaum, WordNet: An Electronic Lexical Database. Cambridge, MA, USA: MIT Press, 1998.

[5] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in Proc. 18th ACM Conf. Inf. Knowl. Manage., 2009, pp. 919–928.

[6] S. Banerjee, K. Ramanathan, and A. Gupta, "Clustering short texts using wikipedia," in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2007, pp. 787–788.

[7] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," in Proc. 20th Int. Joint Conf. Artif. Intell., 2007, pp. 1606–1611.

[8] E. Gabrilovich and S. Markovitch, "Feature generation for text categorization using world knowledge," in Proc. 19th Int. Joint Conf. Artif. Intell., 2005, pp. 1048–1053.

[9] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in Proc. Int. Conf. Manage. Data, 2012, pp. 481–492.

AUTHORS:

JANGA NAGARAJU, M.TECH CSE SCHOLAR, SSSISE, ANANTAPUR.

BIMAL KUMAR, ASSOCIATE PROFESSOR, CSE DEPARTMENT, SSSISE, ANANTAPUR.

S. AKHILENDRANATH, ASSISTANT PROFESSOR, CSE DEPARTMENT, SSSISE, ANANTAPUR..

P GANGADHAR, ASSISTANT PROFESSOR, CSE DEPARTMENT, SSSISE, ANANTAPUR.