Gadaa Journal/Barruulee Gadaa

Vol. 3 No. 2 June 2020
https:journals.ju.edu.et

e-ISSN: 2616-3985
p-ISSN: 2616-3977

A Bilingual Journal of the Institute of Oromo Studies
(IOS) Jimma University

A Probabilistic Information Retrieval System for Afaan Oromoo Text

Tolessa Desta (MSc.)*1, Million Meshesha (PhD)2, Workineh Tesema (MSc.)3

[1]Wollega University, Department of Information Science
[2]Addis Ababa University, Department of Information Science
[3]Jimma University, Department of Information Science

Submission Date: May 4, 2020
Acceptance Date: Jun 30, 2020

Full Length Research Paper

Abstract

This work presents a probabilistic information retrieval system for Afaan Oromoo text. As a considerable amount of information is being produced in Afaan Oromoo rapidly and continuously; experimenting on the applicability of information retrieval system for Afaan Oromoo is important. The aim of this work is to design prototype architecture of Afaan Oromoo text retrieval system based on probabilistic model in order to increase its effectiveness per the user's information need.Developing an information retrieval (IR) system for Afaan Oromoo allows searching and retrieving relevant documents that satisfy information need of users. A probabilistic retrieval model that has the capability of reweighting query terms based on relevance feedback could be used and also the potential of the model was investigated. The work presents the design and implementation of a probabilistic model for Afaan Oromoo free-text-documents. Both indexing and searching modules were constructed and text operations were applied. Then, the retrieval system was evaluated using two hundred (200) Afaan Oromoo free-text-documents and using ten (10) queries. Other types of documents like video, images and audio were not included. The systemregistered, after user relevance feedback, an average precision, recall and F- measure of 60%, 91.56% and 72.5% respectively. This result is achieved without controlling the problem of synonyms and polysemous of terms that exist in Afaan Oromoo text. It can be concluded that; when the terms are added to the user query and user relevance feedback is applied; the performance of the retrieval system increases. It is recommended that using other probabilistic models like Bayesian network, Bayesian belief network, and Bayesian inference network model will more enhance the performance.

Keywords: Information Retrieval System, Binary Independent Model, Probabilistic Model, Afaan Oromoo Text, Information Retrieval

* Corresponding author.

Axareeraa

Hojiinkun kan dhiyeessu, barreeffama Afaan Oromoofi sirna yaasa odeeffannoo 'piroobaabilistikiiti'. Akkuma odeeffannoo barbaachisaan ittifufinsaafi ariitiin Afaan Oromoo keessatti uumamu, yaaliinitti fayyadama sirna yaasa odeeffannoo, Afaan Oromoof barbaachisaadha. Kaayyoon hojiikanaa, modeela "piroobaabilistikii" irratti hundaa'uun barbaachisummaa isaa dabaluuf akka fedha odeeffannoo fayyadamtootaattis irna yaasa odeeffannoo Afaan Oromoo bocanii kaa'uudha. Sirna yaasa odeeffannoo Afaan Oromoof guddisuun, fedhao deeffannoo fayyadamtootaa guutuun galmee barbaachisoo ta'an barbaaduufi yaasuuf eyyama. Modeelli yaasa 'piroobaabilistikii' kan dandeettii jechootaa gaaffilee irra deebi'uun ilaalu deebii barbaachisaa irratti hundaa'uun fayyadama. Akkasumas, dandeettiin moodelichaas qoratameera. Hojiin kun kan dhiyeessu, bocaafi raawwii moodela 'piroobaabilistikii' kan galmee barreeffama bilisaa Afaan Oromooti. Moojuliin barbaachaafi tartiibessuus ('indeeksiingiis') kan ijaaramaniidha. Akkasumas, dalagaawwan barreeffamaa hojiirra oolaniiru. Achiibooda, sirniyaasaa, galmee barreeffama bilisaa Afaan Oromoo 200fi gaaffilee 10 fayyadamuun kan madaalamanidha. Galmeewwan kan biraa Kan akka viidiyoo, fakkiifi sagalee qorannoo kana keessatti hinhammatamne. Deebii barbaachisaa fayyadamaa booda, giddu-galeessaan piriisiishin, riikoolfi F-meejeriin walduraaduubaan %6, %91.56fi %72.5 sirnichi galmeesseera. Bu'aan kun kan argame, to'annoo rakkoo walfakkiifi faallaa jechootaa kan barreeffamoota Afaan Oromootiin ala. Kanarraa hubachuun kan danda'amu, gaaffii fayyadamaa sanatti yeroo jechoonni dabalamanfi deebii barbaachisaan fayyadamaa hojiirra yoo oolulu, dandeettiin sirna yaasaa nidabala. Modeelota 'piroobaabilistikii' kanneen akka 'Baayeeshiyaan Neettiwoork', Baayeeshiyaan Billiif Neettiwork'fi 'Baayeeshiyaan Infereensi Neettiwoorki' fayyadamuun irra caalaatti dandeettiin sirnayaasaa akka dabaluuf yaada furmaataati.

Jerchoota Ijoo: Sirnayaasa odeeffannoo, moodeelabaayinariibilisa, modeela piroobaabilistikii, barruu Afaan Oromoo, sirna yaasa odeeffannoo

1. Introduction

Today, in the age of information, people use the Internet over day and night to fulfill their information needs. If information becomes large in size and documents easily available electronically, retrieving relevant documents is difficult. This exponential growth of information records of all kinds' results in the problem of information explosion (Christopher, 2009). The need to store and retrieve written information became increasingly important over centuries, especially with inventions like paper and the printing press. Soon after computers were invented, people realized that they could be used for storing and retrieving large amounts of information (Kocabas, 2011).

According to Atalay (2014), noted that the current models handle documents with complex internal structure and most of them incorporate a relevance feedback component that can improve performance. Among the models, probabilistic model is the most common. It works based on the probability ranking principle. As stated by Robertson (1977), if a reference retrieval systems response is ranked in decreasing order, the overall effectiveness of the system to its user will be the best. Document collections are retrieved based on probability of relevance to the user who submitted the request.

Experimental evidences show that other models like vector space model (VSM) and its variant models such as,Extended Boolean Model (EBM) and Generalized Vector Space Model (GVSM) are notattempted to define uncertainty in IR system (Crestan, 2001). They do not have relevancefeedback and term reweighting mechanism by them-selves to do with the external realities ofusers. There are different IR methods which have a probabilistic basis. The most widely usedones are binary independent model (BIM) and Bayesian network model (BNM). Binary independent model (BIM) worksbased on representation of queries and documents with relevance feedback data (Crestan, 2001).

Predicting relevant documents is one of the core issues in IR system. Probabilistic IR models are based on the probabilistic ranking principle. Binary independent model (BIM) is the most and first influential model used in IR system. As the name implies, in BIM, the index terms exist independently in the documents and we can then assign binary values to these index terms. The terms in the document are distributed independently (Neto, 1999).

Boolean model and VSM are important in the history of IR, and then probabilistic model came in to take the dominant role in IR system. Vector space model is representation of index terms and query as vectors embeddedin a high dimensional Euclidean space, where each term is assigned as a separatedimension. Boolean model is the oldest model of information retrieval. In the Boolean there are three basic logical operatorsAND, OR and NOT. AND is logical product, OR is logical sum and NOT is logical difference. Since IR system is dealing with free-text-document, there is a need to apply text operations, such as tokenization, normalization, stop word removal and stemming. Stemming is the process of reducing morphological bounding of a given word to their stem. Stemmers are categorized into three subcategories by their stemming method. Dictionary-based stemmers, statistical-based stemmers and affix removal stemmers (Sharifloo, 2008).

Therefore, this work presents Afaan Oromoo text retrieval system by using binary independent model. In this case, the probabilistic approach can improve uncertainty of the retrieval system for users' query and effectiveness of the system, because it overcomes the limitations of vector space model present in the language. A probabilistic approach gives a chance for a user as a relevance feedback to a user query in the retrieval process.

2. Overview of Information Retrieval

Information retrieval is very wide-ranging area of study, with the main aim of searchingrelevant documents from large corpus that satisfies information needs (Zaman, 2010).

In old days, people have become aware about the consequences of archiving and finding information. With arrive of computers, storing the huge amount of information become possible and finding the useful information become necessary. For this purpose, information retrieval becomes a very important research. This information retrieval concerned with searching and retrieving of information from a huge collection of documents (Deep, 2017).

According to Hiemstra (2009), IR technology is a combination of experiments and theory. Experiments are required to assess how the technology deals with the rapid growth of information and documents, and theoretical models help researchers avoid deductive reasoning during such experiments.

IR is the discipline that deals with retrieval of unstructured data, especially textual documents, in response to a query. The need for effective methods of automated IR has grown in importance.

An IR request may specify desired characteristics of both the structured and unstructured components of the documents to be retrieved (Zaman, 2010).

Melkamu (2017), states that, information retrieval (IR) deals with the representation, storage, organization, and access to information items. The representation and organization of the documents should provide the user with easy access to the information in which he/she is interested.

3. Methods and Materials

The goal of IR is to provide users with those documents that will satisfy their information need. IR model is the mechanism of predicting and explain the need of the user given the query to retrieve relevance documents from the collection. The three most widely used information retrieval model that bases on statistical approaches are: Boolean model, vector space model, and probabilistic model (Singhal, 2002).

The Boolean model uses set theory and it is failed to rank the result list of retrieved documents.
According to the Boolean model a document is either relevant or nonrelevantwith respect to a particular query; there is no notion of grading. The similarityof a document dj to a query q is defined in equation 3.1.
simdj, q = 1___, if document satisfies the Boolean query……………………..………. (3.1)
0___, otherwise

The vector space model is a way of representing documents through the words that they contain. The vector model has the disadvantage that index terms are assumed to be mutually independent andcomputationally expensive. The probabilistic model attempts to address the uncertainty problem in IR through the formal methods of probability theory. Unlike in the vector space model, in probabilistic model, the document ranking is based on the probability of the relevance of documents and the query submitted by the user(Singh, 2015).
The similarity function between a document vector Di and query Q is depicted in equation 3.2 as
follows (Manning, 2008): -

$$Cosine\theta = sim(Q, D_i)$$

$$= \frac{\sum_{i=1}^{v} w_{Q,j}\ x\ w_{i,j}}{\sqrt{\sum_{j=1}^{v} w_{Q,j}^2}\ x\ \sqrt{\sum_{j=1}^{v} w_{i,j}^2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ldots.(3.2)$$

Where $w_{Q,j}$is the weight of term j in the query Q, and is defined in similar way as $w_{i,j}$ (that is, $tf_{Q,j}$×idf_j). The term weighting scheme plays an important role for similarity measure.

$$\mathbf{w_{i,j} = tf_{i,j}xidf = tf_{i,j}xlog\frac{D}{df_j}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.3)$$

Where, D is the number of documents in the document collection and IDF stands for inverse document frequency.

Afaan Oromoo is one of the widelyspoken languages in Ethiopia with large number of speakers under Cushitic family. As Afaan Oromoo has a large number of speakers in Ethiopia, a huge amount of information isreleased by this language per day. Those speakers may want to browse by their own language;since, they can easily build their query by their own language. When users lack to clearly definetheir information need, it is difficult to find and get relevant documents. Using probabilisticmodel reduces difficulty of finding relevant documents and uncertainty.

Afaan Oromoo is phonetic language in which its characters sound is the same in every word in contrast to English language. Some of the Afaan Oromoo language specific features are having one or two vowels in between consonants convey different meanings which are called as, '*Jechadheeraa*' and'*jechagabaabaa*' depending on the number of vowel letters used. Afaan Oromoo has own structural morphology which is different from other languages. For example, '*GammachuunJimmaatiidhufe*' means Gemechu comes from Jimma. It disagrees with subject-verb-object (SVO) is in English.But in Afaan Oromoo subject-object-verb (SOV) agreement.

In indexing component, text preprocessing like tokenization, normalization, stop word removal and stemming) techniques were done. One can view an index inverted file as a list of words where each word is followed by the identifier of every text that contains the word. The number of occurrences of each word in a text is also stored in this structure. Hence, the major step in creating an inverted index file is:-

1. Collecting the documents and read all documents to be indexed
2. Tokenizing the document collected
3. Normalizing the tokenized documents in similar case
4. Remove the stop word list from the documents collected
5. Change all terms from the documents collected into their root (stem) words
6. Identify list of tokens to be indexed and create inverted index file which includes vocabulary files and posting files.

In the searching component, a similar text preprocessing (tokenization, normalization, stop word removal and stemming) techniques is followed just as in the indexing part. Then, probability of relevance based on binary independent model techniques used to retrieve from inverted index file and rank relevant documents accordingly. After ranking the relevant retrieved documents, the users gave a feedback by reformulating the query and restart the search for improved results. Again based on binary independent model, searching systems from the inverted index file can be occurred. The users query also expanded to enhance relevant documents retrieval to satisfy information need of the users.Then the documents are re-ranked in decreasing order of their probability of relevance. The user reformulates the query by adding new terms in order to enhance the performance of the system and to satisfy the user information need. Based on the reformulated query, the system searches the documents from inverted index file and give a result in decreasing order again. The system give an option for the user to search relevant documents up to the user can be satisfied. If the user is satisfied to the obtained documents, the system exit and give an acknowledgement for the users. Hence, an activity of a cycle could be applied in the architecture designed. The above descriptions are depicted in figure 1 below.

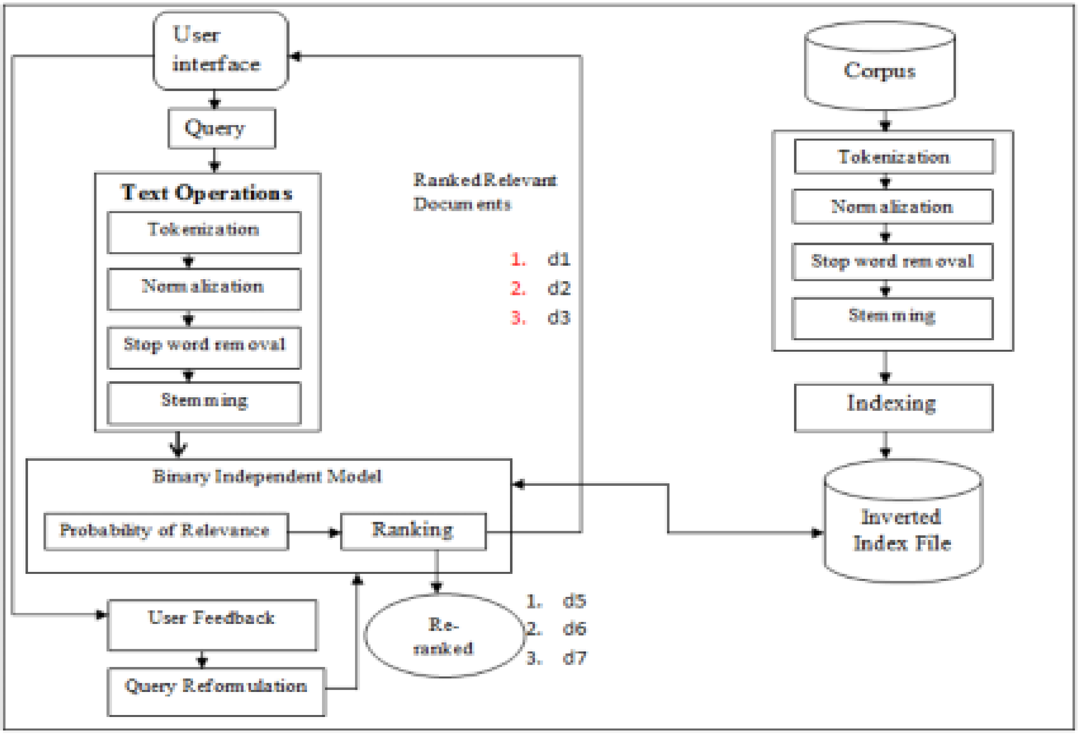


Figure 1: -A probabilistic Based Architecture of Afaan Oromoo Text Retrieval System (Baeta-Yates et al., 1999)

Again, as shown in the figure 1 above, IR process starts with the specifying the problem (user need), then this user need is transformed to some query. The system searches from the inverted index file for such query and get back the result to specify the relevancy of the document to his/her need.

Searching Using Probabilistic Model

The probabilistic model that attempts to simulate the uncertainty nature of an IR system guides the searching processes. Binary independent probabilistic IR model is adopted to search the relevant documents from Afaan Oromoo corpus.In binary independent model, there are three steps to compute term probability.The first step compute terms whenthere is no retrieved document at initial stage. The second step compute terms after documentsare retrieved and feedback is provided by the user. The third step compute terms when partialfeedback is given (Neto, 1999).

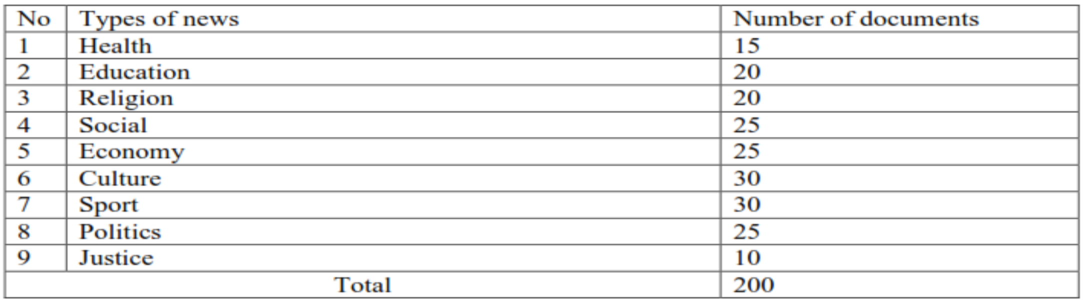
$$P(k_i|R) = 0.5 \; and \; P(k_i|R) = log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . . (3.4)$$

Where, N is the total number of documents in the collection and nis the number of documents which contain the index term ki.

Corpus Acquisition and Preparation

The corpus of Afaan Oromoo documents were selected from different sources for the experimentation. In IR system, the corpus is needed for training of the system. Corpus is a large collection of texts. Document is collected from different news articles and other online resources, including Oromia Broadcasting Network (OBN), Voice of America (VOA), and different websites publishing magazines, newspapers, educational books and fictions to make the corpus variety. We collected some of the corpus from OBN Gazetteers Adama and Finfinnee branch. Others are collected from Crawling, asking the news providers, manually copying texts from the websites. For the sake of this study, two hundred (200) documents with average size of 6 MB and 10 testing queries was collected for training and prepared for testing the evaluation of the system respectively. Newspapers are considered as consisting different issues of the community such as social, political, economic, sport, educational, culture, justice, religion and health issues. They are a potential source for collecting corpus, which is not biased to specific issue. This heterogeneity of the data set help to test the system more generic. Each file is saved under common folder using .txt format. Theexperiments in this study were based on sets of documents and queries set upby the researcher.

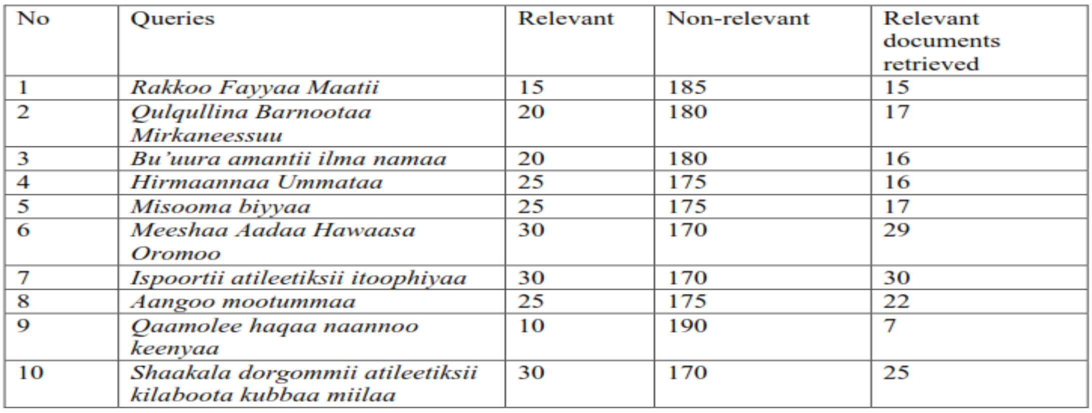
Table 1: - Types of news article used for development of Afaan Oromoo IR system



No	Types of news	Number of documents
1	Health	15
2	Education	20
3	Religion	20
4	Social	25
5	Economy	25
6	Culture	30
7	Sport	30
8	Politics	25
9	Justice	10
Total		200

Each news articles are saved under common file folder using .txt format in notepad, which is supported by most programming languages. Additionally, 10 test queries were selected by the researcher to test the performance of the system after collecting user query from the public via questions from twenty individuals of the native speakers and reviewing the corpus collected.

Table 2: -List of queries with their relevant judgments



No	Queries	Relevant	Non-relevant	Relevant documents retrieved
1	*Rakkoo Fayyaa Maatii*	15	185	15
2	*Qulqullina Barnootaa Mirkaneessuu*	20	180	17
3	*Bu'uura amantii ilma namaa*	20	180	16
4	*Hirmaannaa Ummataa*	25	175	16
5	*Misooma biyyaa*	25	175	17
6	*Meeshaa Aadaa Hawaasa Oromoo*	30	170	29
7	*Ispoortii atileetiksii itoophiyaa*	30	170	30
8	*Aangoo mootummaa*	25	175	22
9	*Qaamolee haqaa naannoo keenyaa*	10	190	7
10	*Shaakala dorgommii atileetiksii kilaboota kubbaa miilaa*	30	170	25

For speeding up searching the document corpus was indexed using inverted index structure. To this end, text operations was applied for identifying content-bearing terms with the help of tokenization, stop word detection, normalization and stemming processes. Given Afaan Oromoo text corpus, the IR system organize them using index file to enhance searching. The first step is tokenization of the text words to identify stream of tokens (or terms). Next, text is normalized in order to bring together similar word written with different punctuation marks and variation cases (UPPER, lower or mixed). The normalized token is checked again as it is not stop word. This is followed by removing stop words from the corpus. Content bearing terms (nonstop words) are stemmed. For all stemmed tokens their weight calculated and then inverted index file was constructed (Debela,2010).

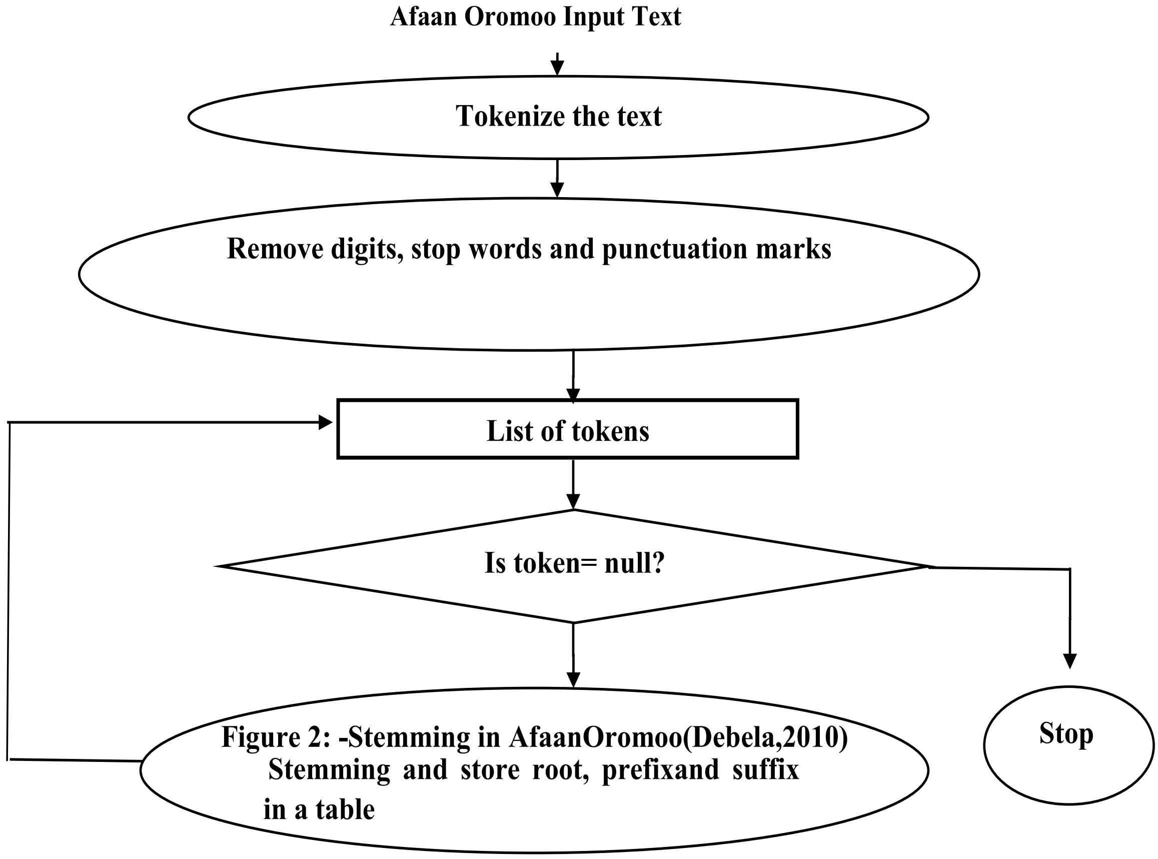

Afaan Oromoo Input Text
Tokenize the text
Remove digits, stop words and punctuation marks
List of tokens
Is token= null?
Stop
Stemming and store root, prefixand suffix in a table

Figure 2: -Stemming in AfaanOromoo(Debela,2010)

Inverted index

The major concept in information retrieval is how the documents/corpus are going to be represented in information retrieval system. This logical representation of documents using its content bearing words is inverted file. An Inverted index always maps back from terms to the parts of a document where they occur (Christopher, 2009).

An inverted index is an optimized data structure that can be used for information retrieval. The basic idea for building an inverted index is to keep a dictionary of the unique terms in the collection. For each term in the collection, we maintain a list of documents (by document IDs) in which the term occurs as well as a number for the term's frequency in the specified document. This list is called a posting list. The posting list is stored in the secondary storage, while the dictionary is stored in main memory (Manning, 2008).

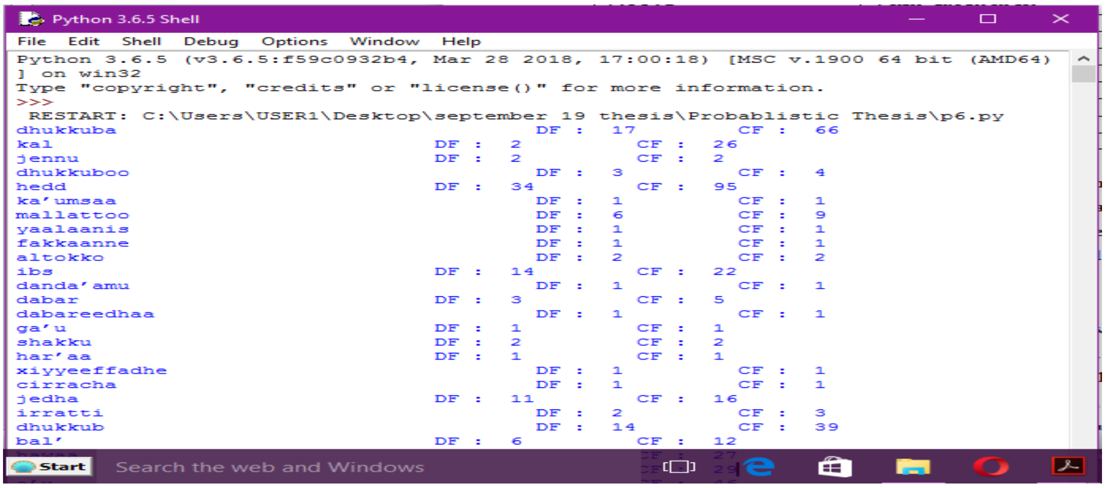
As shown in figure 1 above, the indexing part of the corpus contains vocabulary file and posting file. The text should undergo several preprocessing operations like tokenization, normalization, stop word removal and stemming before it can be stored in an inverted index. The inverted file allows an IR system to quickly determine what documents contain a given set of words, and how often each word appears in the document (Heinz, 2003).

4.Results and Discussion

In the experiment, we presented on Afaan Oromoo text retrieval by probabilistic model approach by using ten queries. We calculated the percentage of recall, F-measure and precision for each output. We described the experiment conducted to compare the performance of the system. The result obtained enhances the system performance. The system performance is evaluated before and after user relevance feedback. Evaluations were done by measuring the recall, precision and F-measure.

The results before relevance feedback indicated registered low performances with their registered value for recall, precision and F-measure of 83.5%, 34.81% and 49.14% respectively. The F-measure score registered 49.14% for Afaan Oromoo text, which indicates the performance of the system is not good. This is because; documents containing one of the query terms, but that are not-relevant are retrieved. These documents are irrelevant because, the query term found in those documents does not express the meaning of the query with respect to other terms found in the query. On the other hand, in probabilistic model, the relevant document is based on Boolean expression. Thus, all terms that match one of user queries are retrieved which increases the number of denominator used for calculating precision, thereby decreasing the percentage of precision. Therefore, to increase the performance of the system, the probabilistic model uses user relevance feedback so as to apply query terms re-weighting in order to increase the weight of terms found in relevant documents and decrease the weight of terms found in non-relevant documents.

The inverted file has two separates files vocabulary and posting file; the vocabulary file contains Terms, Document frequency and Collection frequency and illustrated in figure 3 and posting file contains documents ID, Term frequency and terms location.The document frequency andcollection frequency is cross referenced to posting file. From below figure 3, terms are retrieved with their document frequency (DF) and collection frequency (CF).

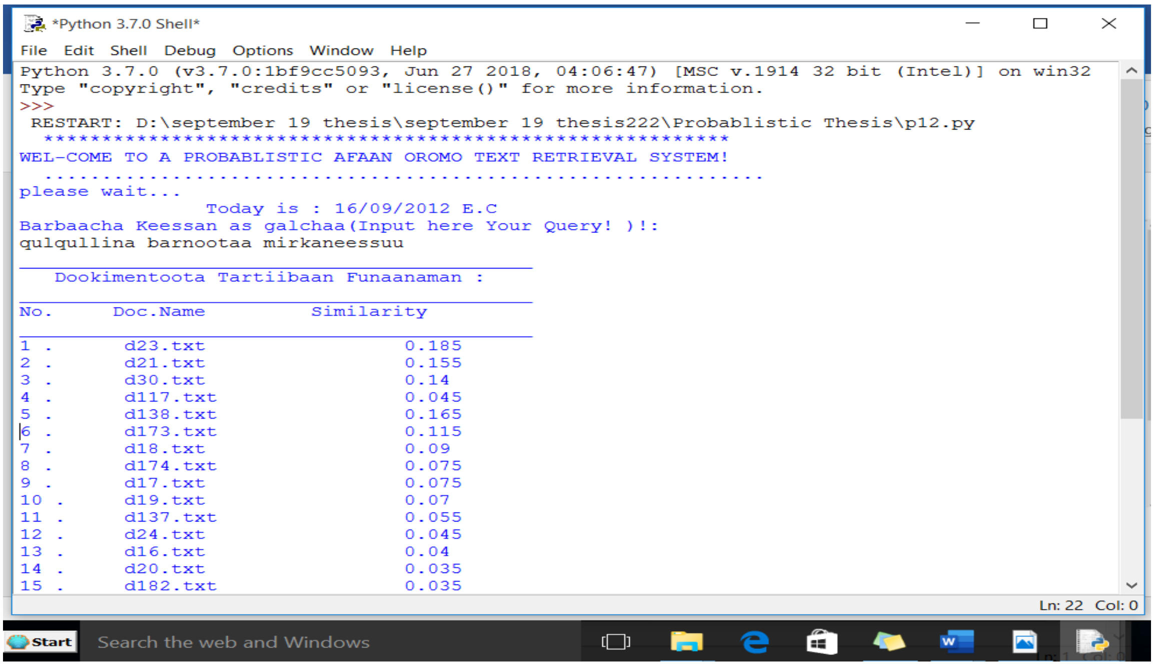


Python 3.6.5 Shell
File Edit Shell Debug Options Window Help
Start
Search the web and Windows

```
Python 3.6.5 (v3.6.5:f59c0932b4, Mar 28 2018, 17:00:18) [MSC v.1900 64 bit (AMD64)
] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
 RESTART: C:\Users\USER1\Desktop\september 19 thesis\Probablistic Thesis\p6.py
dhukkuba                                  DF :  17        CF :  66
kal                               DF :  2         CF :  26
jennu                             DF :  2         CF :  2
dhukkuboo                                 DF :  3         CF :  4
hedd                              DF :  34        CF :  95
ka'umsaa                                  DF :  1         CF :  1
mallattoo                                 DF :  6         CF :  9
yaalaanis                                 DF :  1         CF :  1
fakkaanne                                 DF :  1         CF :  1
altokko                                   DF :  2         CF :  2
ibs                               DF :  14        CF :  22
danda'amu                                 DF :  1         CF :  1
dabar                             DF :  3         CF :  5
dabareedhaa                               DF :  1         CF :  1
ga'u                              DF :  1         CF :  1
shakku                            DF :  2         CF :  2
har'aa                            DF :  1         CF :  1
xiyyeeffadhe                              DF :  1         CF :  1
cirracha                                  DF :  1         CF :  1
jedha                             DF :  11        CF :  16
irratti                                   DF :  2         CF :  3
dhukkub                                   DF :  14        CF :  39
bal'                              DF :  6         CF :  12
```

Figure 3: Vocabulary File

Based on the performance registered, an attempt has been made to compare the result of probabilistic based IR system for Afaan Oromoo with the previously done Afaan Oromoo IR system using vector space model by Gezehagn (2012). The experiment shows that the performance is on the average 0.575(57.5%) precision and 0.6264(62.64%) recall (Gezehagn, 2012).Using the information given, evaluation is done by measuring the recall, precision and F-measure for the initial performance of the system.The first step of this system is to get query from the user.



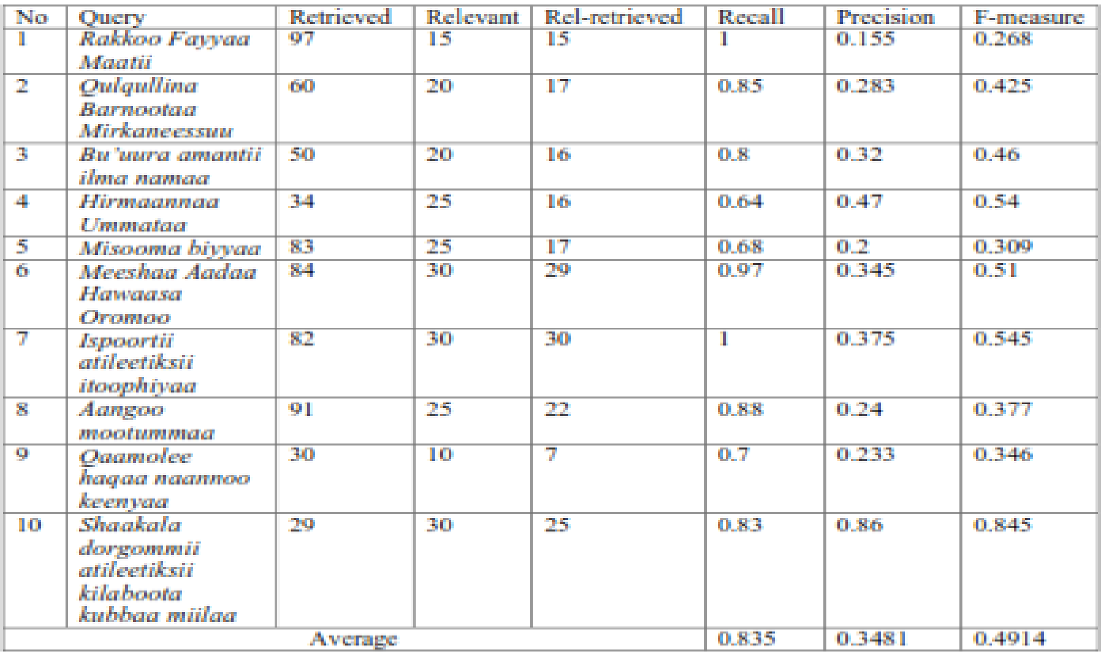
Python 3.7.0 Shell
File Edit Shell Debug Options Window Help
Ln: 22 Col: 0
Start
Search the web and Windows

```
Python 3.7.0 (v3.7.0:1bf9cc5093, Jun 27 2018, 04:06:47) [MSC v.1914 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
 RESTART: D:\september 19 thesis\september 19 thesis222\Probablistic Thesis\p12.py
  ***********************************************************
WEL-COME TO A PROBABLISTIC AFAAN OROMO TEXT RETRIEVAL SYSTEM!
  ..............................................................
please wait...
                Today is : 16/09/2012 E.C
Barbaacha Keessan as galchaa(Input here Your Query! )!:
qulqullina barnootaa mirkaneessuu
____________________________________________
   Dookimentoota Tartiibaan Funaanaman :
____________________________________________
No.      Doc.Name          Similarity
____________________________________________
1 .       d23.txt                 0.185
2 .       d21.txt                 0.155
3 .       d30.txt                 0.14
4 .       d117.txt                0.045
5 .       d138.txt                0.165
6 .       d173.txt                0.115
7 .       d18.txt                 0.09
8 .       d174.txt                0.075
9 .       d17.txt                 0.075
10 .      d19.txt                 0.07
11 .      d137.txt                0.055
12 .      d24.txt                 0.045
13 .      d16.txt                 0.04
14 .      d20.txt                 0.035
15 .      d182.txt                0.035
```

Figure 4: Retrieved documents for a given query '*qulqullinabarnootaamirkaneessuu*'

From the above figure 4, documents are retrieved by BIM based on probability of relevance for the user's information need. For instance; as can be seen in figure 4, **d23.txt** is the first top ranked document retrieved based on the user query. From here, the users enter the query, and then the system retrieves the relevant documents for the users query. Even though all relevant documents are not seen on the snapshot figure, the most top twenty relevant documents retrieved are depicted on the figure. Accordingly, **d23.txt, d21.txt** and **d30.txt** are the first, second and third ranked relevant documents as depicted in figure 4 respectively.

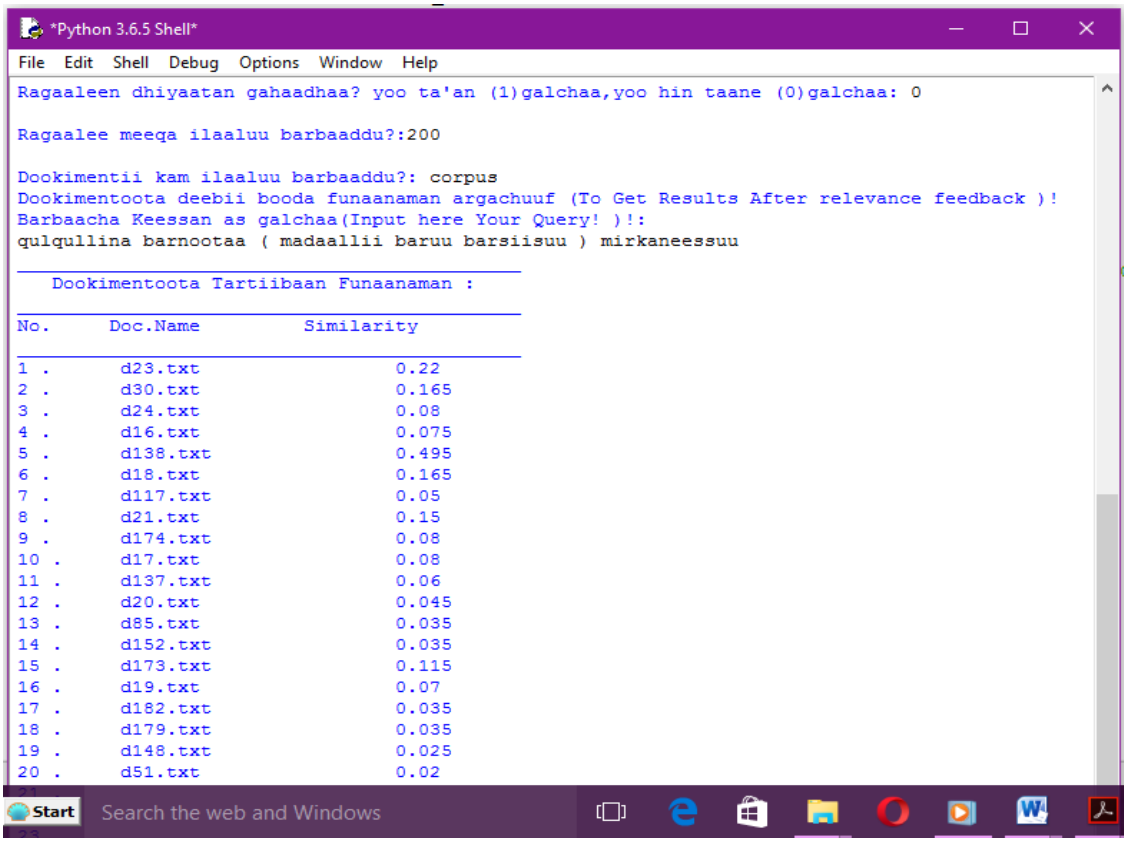
Table 3: - The Initial Performance of the System



No	Query	Retrieved	Relevant	Rel-retrieved	Recall	Precision	F-measure
1	*Rakkoo Fayyaa Maatii*	97	15	15	1	0.155	0.268
2	*Qulqullina Barnootaa Mirkaneessuu*	60	20	17	0.85	0.283	0.425
3	*Bu'uura amantii ilma namaa*	50	20	16	0.8	0.32	0.46
4	*Hirmaannaa Ummataa*	34	25	16	0.64	0.47	0.54
5	*Misooma biyyaa*	83	25	17	0.68	0.2	0.309
6	*Meeshaa Aadaa Hawaasa Oromoo*	84	30	29	0.97	0.345	0.51
7	*Ispoortii atileetiksii itoophiyaa*	82	30	30	1	0.375	0.545
8	*Aangoo mootummaa*	91	25	22	0.88	0.24	0.377
9	*Qaamolee haqaa naannoo keenyaa*	30	10	7	0.7	0.233	0.346
10	*Shaakala dorgommii atileetiksii kilaboota kubbaa miilaa*	29	30	25	0.83	0.86	0.845
Average					0.835	0.3481	0.4914

As table 3 shows, the retrieval result of the prototype on the average Precision, Recall and F-measure is 0.3481, 0.835 and 0.4914 respectively. Afaan Oromoo terms are highly inflected for number, genders, possession, plural, and conjunctions. There are many terms with the same meanings (synonyms) and a term which has many meanings (polysemous) is another reason for the result. For example, term, '*aangoo*' have the meaning with, '*taayitaa*' which is, Authority. In addition, spelling error is also another factor for being lower performance. For instance; for the word*, 'barnoota'*, if you miss n and write '*baroota'* it is completely changed. 'b*arnoota'* means education, but, 'baroota' means years.

In general, the performance of the evaluation result is low. But in the user relevance feedback, , the queries were allowed to be expanded and re-formulated to get a good result. The result of user relevance feedback for recall and precision was 0.9156 and 0.60 respectively.

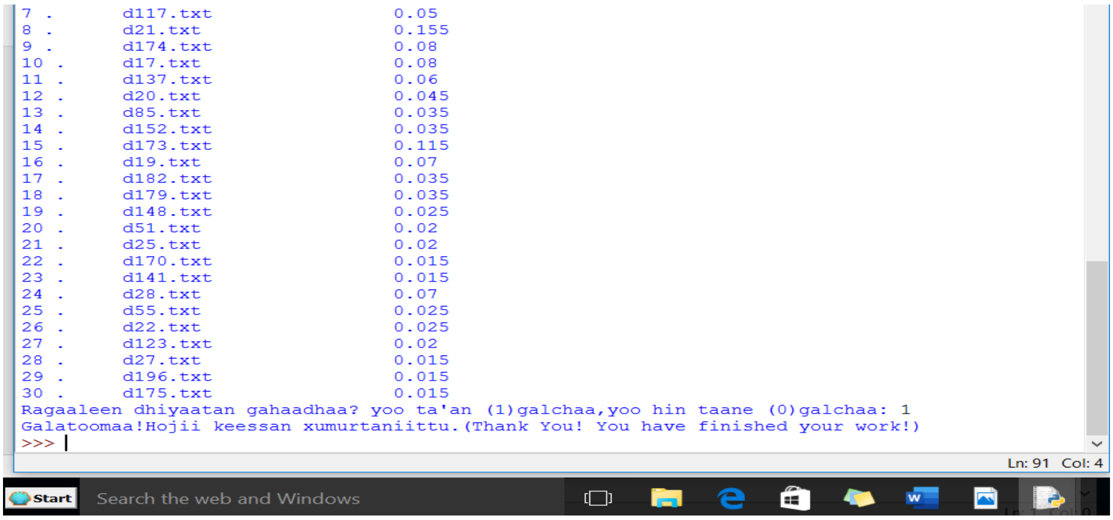

Python 3.6.5 Shell
File Edit Shell Debug Options Window Help
Start
Search the web and Windows

```
Ragaaleen dhiyaatan gahaadhaa? yoo ta'an (1)galchaa,yoo hin taane (0)galchaa: 0

Ragaalee meeqa ilaaluu barbaaddu?:200

Dookimentii kam ilaaluu barbaaddu?: corpus
Dookimentoota deebii booda funaanaman argachuuf (To Get Results After relevance feedback )!
Barbaacha Keessan as galchaa(Input here Your Query! )!:
qulqullina barnootaa ( madaallii baruu barsiisuu ) mirkaneessuu
______________________________________________________________
    Dookimentoota Tartiibaan Funaanaman :
______________________________________________________________
No.        Doc.Name            Similarity
______________________________________________________________
1 .        d23.txt                      0.22
2 .        d30.txt                      0.165
3 .        d24.txt                      0.08
4 .        d16.txt                      0.075
5 .        d138.txt                     0.495
6 .        d18.txt                      0.165
7 .        d117.txt                     0.05
8 .        d21.txt                      0.15
9 .        d174.txt                     0.08
10 .       d17.txt                      0.08
11 .       d137.txt                     0.06
12 .       d20.txt                      0.045
13 .       d85.txt                      0.035
14 .       d152.txt                     0.035
15 .       d173.txt                     0.115
16 .       d19.txt                      0.07
17 .       d182.txt                     0.035
18 .       d179.txt                     0.035
19 .       d148.txt                     0.025
20 .       d51.txt                      0.02
```

Figure 5: -Retrieved Documents after User Relevance Feedback for *'qulqullinabarnootaa Mirkaneessuu'*



```
7 .        d117.txt                     0.05
8 .        d21.txt                      0.155
9 .        d174.txt                     0.08
10 .       d17.txt                      0.08
11 .       d137.txt                     0.06
12 .       d20.txt                      0.045
13 .       d85.txt                      0.035
14 .       d152.txt                     0.035
15 .       d173.txt                     0.115
16 .       d19.txt                      0.07
17 .       d182.txt                     0.035
18 .       d179.txt                     0.035
19 .       d148.txt                     0.025
20 .       d51.txt                      0.02
21 .       d25.txt                      0.02
22 .       d170.txt                     0.015
23 .       d141.txt                     0.015
24 .       d28.txt                      0.07
25 .       d55.txt                      0.025
26 .       d22.txt                      0.025
27 .       d123.txt                     0.02
28 .       d27.txt                      0.015
29 .       d196.txt                     0.015
30 .       d175.txt                     0.015
Ragaaleen dhiyaatan gahaadhaa? yoo ta'an (1)galchaa,yoo hin taane (0)galchaa: 1
Galatoomaa!Hojii keessan xumurtaniittu.(Thank You! You have finished your work!)
>>> |
```

Figure 6: -Exit System

When the user satisfies on his/her needs, he/she exit from the system. The recycle activity can be occurred, until the user satisfied.

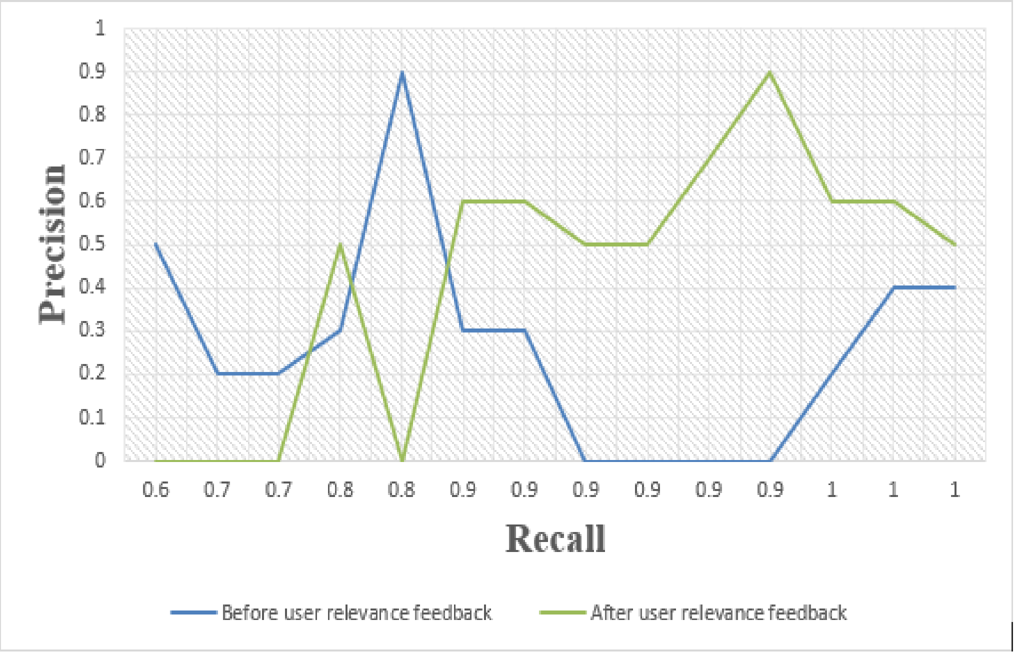


Figure 7: Precision/Recall curve before and after user relevance feedback

As shown in the Figure 7 above, the performance of the system registers better performance when the user provides relevant feedback. Vertical and horizontal line of the figure 7 above shows that, the result of precision and recall respectively. The curve at the upper side of the graph in which recall and precision reaches maximum point indicates, the highest performance registered by the system. The maximum precision registered is 0.9, before and after relevance feedback at recall level of 0.8 and 0.9 respectively. From this point, the value of maximum precision is same before and after feedback. But the value of recall level is increased by 0.1. This represents the average performance registered by the system. From this figure, the minimum precision value before and after user relevance feedback is 0.2 and 0.5 respectively. Hence, the minimum value of precision after user relevance feedback is increased by 0.3.

5. **Conclusion**

Text retrieval system is very important for retrieval of textual documents. The study attempts to develop a probabilistic IR system for Afaan Oromoo text (Debela, 2010). Our hypothesis was enabling accessing unstructured Afaan Oromoo free-text-documents from the system using Afaan Oromoo queries and increasing the performance of the system by using user relevance feedback. The result we have obtained shows significant improvement over the previous runs. We feel that, this is relatively good improvements due to the enhancement of our performance and refinements of model. For this study, two hundred (200) different textual documents and ten (queries) were used for doing the experimentation. In this study, the Binary Independent Model (BIM) is chosen and implemented. At first step when the search component initiated the system generates the first ranked list of relevant documents then using terms from the initial guess made the system also searches again using user relevance feedback. Finally, based on the user relevance feedback, the system improves its performance. This leads us to conclude that, user relevance feedback is useful for the improvement of an IR system. However, as the researcher observed from literatures, since probabilistic model used Boolean expression for initial guess of relevant document, it does not consider the importance of the document based on the frequency of the terms in the document. Because of this, sometimes those documents having query terms with highest frequency than others could be ranked lately. In this case, users faced with the problem of having to choose the appropriate words that are also used in the relevant documents. Hence, poor result could be displayed when the system retrieve documents after user relevance feedback. The stemming technique significantly increases the number of documents that match a user‘s query. We used a preprocessing technique, in which the corpus were preprocessed using the tasks such as tokenization, case normalization, stop word removal, stemming, and indexing allows us to have the same standard between query terms and index terms.

Acknowledgement

We would like to thank Jimma University, College of Natural Sciences for financial contribution to conduct this research.

References

Atalay Luel. (2014).A Probabilistic Information Retrieval System for Tigrinya.MSc Thesis, School of Information Science, Addis Ababa University, Ethiopia.

B. R. Ribeiro Neto. (1999). Modern Information Retrieval, 2nd Edition, Addison Wesley Longman Publishers, New York, USA.

C. D. Manning, et al. (2008). Introduction to information retrieval, Cambridge University Press.

Debela Tesfaye. (2010).Designing a Stemmer for Afan Oromo Text: A Hybrid Approach "MSc Thesis, school of information science, Addis Ababa University, Ethiopia.

Hiemstra, D. (2009). "Information retrieval models," Information Retrieval: searching in the 21st Century, (pp. 1-19).

H.S. Christopher D.Manning, Prabhakar Raghavan. (2009). An Introduction to Information Retrieval, 1st Edition, Cambridge University Press, Cambridge England.

Kocabas, et al. (2011). Investigation of Luhn"s claim on information retrieval. *Turkish Journal of Electrical Engineering & Computer Sciences,* pp. 993-1004.

Melkamu Abetu. (2017). Query Expansion for Afan Oromo Information Retrieval Based On Wordnet. Msc Thesis, school of graduate studies, Haramaya University, Haramaya.

Sharifloo, Amir Azim, and MehrnoushShamsfard (2008). A Bottom Up approach to Persian Stemming." In *IJCNLP*, pp. 583-588.

S. Heinz (2003). Efficient single-pass index construction for text databases, ‖ *Journal of the American Society for*, vol. 54, no. 8, pp. 713-729.

Sneha Deep (2017). A Review of Information Retrieval System Using Relevance Feedback Algorithm.*International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882.*

Zaman, A. (2010).Study of Document Retrieval Using Latent Semantic Indexing (LSI) on a Very Large Data Set.

Zhu, R. (2016). Improvement in Probabilistic Information Retrieval Model - Rewarding Terms with High Relative Term Frequency.