22-я Международная конференция по компьютерным и информационным технологиям (ICCIT), 18-20 декабря 2019 года

*Система ответов на вопросы Bangla Intelligence*

*На основе математики и статистики*

|  |  |  |
| --- | --- | --- |
| Доктор Коушер  *Кафедра прикладной математики Ноахали Наука и технология*  *Университет, Ноакхали-3814, Бангладеш. ga.kowsher@gmail.com* | М М Махабубур Рахман  *Кафедра CSTE*  *Наука и техника Ноахали*  *Университет Ноахали-3814, Бангладеш toufikrahman098@gmail.com* | Ск Шохораб Ахмед  *Отдел информации и*  *Техника связи*  *Университет Раджшаи, Раджшай-6205,*  *Бангладеш*  *shohorab.ahmed.it@gmail.com* |

Нусрат Джахан Протташа *Отдел CSE*

*Международный университет нарциссов*

*Дакка,Бангладеш*

*nuaratjahan1234561234@gmail.com*

**Информативная система ответов на вопросы Bangla (BIQAS) является важным методом машинного обучения (ML), который помогает пользователю отслеживать соответствующую информацию с помощью обработки бенгальского естественного языка (BNLP). В этой исследовательской работе мы применили три математические и статистические процедуры для BIQAS, основанные на данных ответов на вопросы. Этими процедурами являются косинусное сходство, сходство Жаккара и наивный алгоритм Байеса. Косинусное сходство взаимодействовало с техникой уменьшения размерности SVD на вопросы пользователей и вопросы, отвечающие на данные, чтобы уменьшить пространственно-временную сложность. Эти процедуры данного исследования разделены на две части: предварительная обработка данных и установление взаимосвязи между вопросами пользователя и содержащимися информативными вопросами. Мы получили 93,22% точных ответов, используя косинусное сходство, 84,64% по подобию Жаккара и 91,31% по наивному алгоритму Байеса.**

***Ключевые слова — BIQAS, BNLP, Информационный поиск, Машинное обучение, Математика, Статистика.***

# I. ВВЕДЕНИЕ

Настоящее время – это эра информации. Информация увеличивается с каждым днем, и мир становится все более информативным, поэтому виртуальная информационно-поисковая система сохраняет свою значительную, то есть искусственную систему ответов на вопросы. Пользователи часто имеют в виду конкретные вопросы, поэтому они хотят получить ответы. Они хотели бы, чтобы ответы были простыми и точными, и они всегда предпочитают выражать вопросы на своем родном языке, не ограничиваясь конкретным языком запросов, правилами формирования запросов или даже определенной областью знаний. Новый подход к соответствию потребностям пользователя заключается в том, чтобы провести фактический анализ вопроса с лингвистической точки зрения и попытаться понять, что на самом деле имеет в виду пользователь.

В Bangla NLP BIQAS был сформирован тремя основными модулями: сбор данных, обработка информации и вопросов пользователей, а также установление взаимосвязи между ними. Для предварительной обработки BNLP проводятся различные методы, например, анафора, очистка специальных символов и знаков препинания, удаление стоп-слов, обработка глаголов, лемматизация и обработка синонимов слов. Чтобы получить идеальное разрешение анафоры, передается знаменитый алгоритм Хоббса‟. В действии лемматизации мы описали три процедуры с сильной системой в наименьшей временной и пространственной сложности. Чтобы уменьшить размер вопроса и информации, мы использовали SVD, который также минимизирует время выполнения программы. Это также помогает понять и рассчитать простым способом. TF-IDF используется для выяснения влияния слов в документах и построения идеального вектора.

Для того, чтобы генерировать ответы пользователей‟ вопросы, мы использовали косинусное сходство, сходство Жаккара и наивно-байесовское сходство. Эти методы помогают установить отношения между вопросами пользователей и информацией.

Материалы резюмируются следующим образом:

* Мы внедрили математические и статистические процедуры для BIQAS для поиска информации.
* Для предварительной обработки данных мы применили алгоритм Hobbs‟, Edit Distance, Trie и DBSRA.
* Мы использовали SVD с косинусным сходством для уменьшения сложности времени и пространства, а также мгновенного ответа на вопросы.
* Для того, чтобы сгенерировать ответ BIQAS Cosine аналогично, используются алгоритмы Jaccard и Naïve-Bayes.
* Чтобы упростить этапы предварительной обработки, мы разработали модуль BLTK, который содержит все этапы предварительной обработки и математические методы.

Для самого простого объяснения мы рассматриваем два вопроса пользователей в качестве примера в каждом термине этой статьи, это

Вопрос пользователя 1: Вот

? [Где находится самый большой женский зал в Бангладеш?]

Вопрос пользователя-2: ? [Как называется его аудитория?]

# См. РЕЛИРОВАННЫЕ СТУДЫ

|  |
| --- |
| 978-1-7281-5842-6/19/$31.00 ©2019 IEEE |

Welbl et al. сформировали набор данных WikiHop, который содержал вопросы, для ответа на которые требуется более одного документа Википедии [1]. Asiaee et al. разработали систему контроля качества на основе онтологии, Onto NLQA. Он состоит из пяти основных частей, таких как лингвистическая предварительная обработка, распознавание сущностей, сопоставление элементов онтологии, обнаружение семантических ассоциаций и формулировка запроса и извлечение ответов [2]. Xie et al. предложили систему ответов на вопросы, которая была основана на онтологии. Из курса «Естественный язык»

Обработка», извлечены данные онтологии [3]. Kowsher, Md, et al. предложил чат-бот Bangla, основанный на обработке языка Bangla. Эта структура состоит из трех основных этапов: обработка вопросов, поиск информации (из Интернета) и извлечение ответов [4]. Ли и др. предложили систему контроля качества, основанную на онтологии. Они определили шестнадцать типов запросов в этой системе. Затем для каждого запроса был определен и реализован соответствующий подход к выводу [5]. Лопес и др. разработали систему ответов на вопросы, основанную на онтологии, AquaLog. Входные вопросы обрабатываются и классифицируются на 23 группы. Если входной вопрос попадает в одну из этих групп, система точно обработает его [6]. Радж предложил систему контроля качества для конкретной области, основанную на онтологической информации. Эта система состоит из четырех основных частей: анализа вопроса, который анализирует вопрос пользователя [7]. Роберт Ф. Симмонс и др. разработали системы ответов на вопросы естественного языка, которые в основном фокусировались на синтаксическом, семантическом и логическом анализе английских строк [8]. Борис Кац изобрел первую в мире ' веб-систему ответов на вопросы под названием START. Он был создан InfoLab Group в Лаборатории компьютерных наук и искусственного интеллекта Массачусетского технологического института, целью которой является предоставление только нужной информации, на его месте предоставление ряда хитов [9]. Moldovan et al. использовали технику понимания естественного языка на основе синтаксиса и технику классификации вопросов, чтобы получить лучшую точность в задаче ответа на вопрос под названием TREC[10] и Kowsher, Md, et al. обнаружил информационную систему автоматического ответа на вопросы Bangla, которая может предоставлять информативные знания от пользователей, задающих [11]. Цай Дунфэн и Цуй Хуань обнаружили китайскую систему автоматического ответа на вопросы, которая использует API веб-сервисов Google [12]. Liu Hongshen, Qin Feng, Chen Xiaoping, Tao Tao и др. предложили программное обеспечение для режима обучения, которое может поддерживать посещаемость студентов, а также может сохранять ответ любого студента [13]. Jeon et al. оценили действие поиска вопроса для четырех известных методов извлечения: модели векторного пространства, модели Окапи, языковой модели и модели трансляции [14]. Вэй Ван, Байчуань Ли, Ирвин Кинг предложили улучшенную модель поиска вопросов, которая может обнаруживать намерения пользователей, связанные с прежним извлечением вопросов [15]. В отличие от этих работ, мы ввели интеллектуальную систему ответов на вопросы Бангла с помощью математики и статистики.

# III.B ACKGROUND STUDY

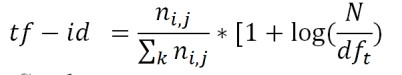
## А. Лемматизация

Лемматизация - это упрощение процесса для выяснения корня экстракта в понимании естественного языка. Лемматизация используется в различных реальных приложениях, таких как интеллектуальный анализ текста, чат-бот, вопросы и ответы и т. Д.

В этом исследовании мы использовали эффективный алгоритм лемматизации для BNLP. Сначала мы немного модифицировали алгоритм Trie на основе префиксов. После этого мы использовали новый алгоритм на основе сопоставления под названием DictionaryBased Search by Removing Affix (DBSRA).

## В. ТФ-ИДФ

TF-IDF - это аббревиатура от Term Frequency-Inverse Document Frequency. Это численный метод, чтобы выяснить важность слова для предложения и является математически статистически значимым. По сути, TF определяет частоту термина в предложении, а IDF определяет важность слова для своих документов. Математически



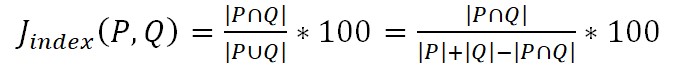
## Сходство косинуса

Косинусное сходство — это мера подобия между двумя ненулевыми векторами внутреннего пространства произведения, которая измеряет косинус угла между ними. Косинус двух ненулевых векторов может быть выведен как:



## Сходство Д. Жаккара

Индекс Жаккара, также называемый сходством Жаккара, является статистическим методом определения сходства различных выборочных наборов. Если P и Q равны двум множествам, то задается формула индекса Жаккара



## Э. Наивно-байесовская теория

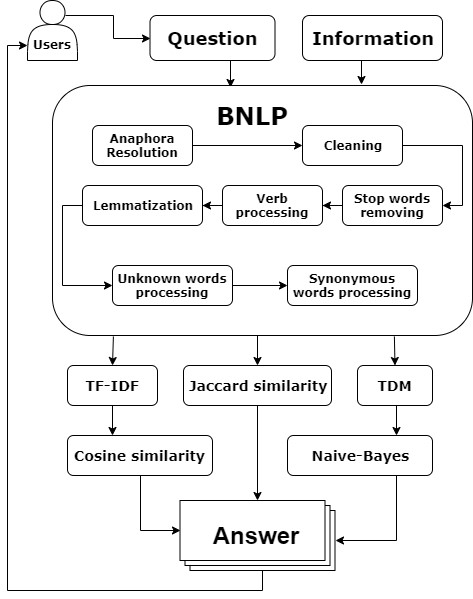
Байесовская гипотеза является наиболее распространенным применением теоремы Байеса, которая является статистической гипотезой. Тестирование лекарств, продуктов и материалов, а также расчет всей продукции компании на основе машин являются одним из видов примеров. Эта теорема математически излагается в следующем уравнении:

M(A│B) = (M(B│A) M(A))/(M(B)), где A и B — события, а M(B) ≠0 и M(A│B) — условная вероятность события A, учитывая, что B истинно. Точно так же M(B│A) также является условной вероятностью, учитывая, что A истинно.

# IV. ПРОПОЗИРОВАННЫЙ ВОРК

В этой исследовательской работе мы представили бенгальскую интеллектуальную систему ответов на вопросы (BIQAS), основанную на математике и статистике с использованием обработки бенгальского естественного языка (BNLP).

Процедура разделена на три части: сбор информативных документов, предварительная обработка данных и отношения между информацией и вопросами пользователей. Корпуса были приложены для предварительной обработки вставленных данных. Действие косинусного подобия, подобия Жаккара и наивных алгоритмов Байеса призвано получить связь между вопросами и ответами. Но косинусное сходство имеет дело с векторами. В этом случае документы и вопросы передаются векторам с помощью модели TF-IDF. Чтобы минимизировать время выполнения и сложность пространства, мы использовали методы SVD.



Стоп-слова относятся к словам, которые не оказывают никакого влияния на документы или предложения. Примерами бенгальских стоп-слов являются (и), (где), (или), (к), (с) и т.д. Поскольку наш BIQAS является данными, основанными на алгоритме, стоп-слова должны быть отклонены.

Здесь мы убрали бенгальские стоп-слова "

(most)‟, " (где)‟ и " (что)‟ из предыдущих вопросов.

В BNLP есть несколько глаголов, которые не могут быть лемматизированы какой-либо системой из-за игнорирования всех видов алгоритмов лемматизации. Например, ( , went) и ( , going) генерируют из корневого слова ( , go). Нет отношений характера между ( , пошел) и ( , go). Так что обработка алгоритмами этих слов не является хорошим выбором. Вот почему эти типы глаголов преобразуются в их корневые глаголы для легкого доступа в качестве лемматизации.

В нашей проектной работе мы использовали различные виды методов лемматизации, такие как DBSRA и Trie. Иногда в языке бангла есть несколько слов, которые не работают в Trie, но работают в DBSRA или работают в Trie, но не в DBSRA. Поэтому мы использовали расстояние Левенштейна, чтобы найти лучшее слово леммы между DBSRA и trie.

Алгоритмы лемматизации не являются хорошим выбором для неизвестных слов. Здесь неизвестные слова относятся к названию места, человеку или названию чего-либо. Расстояние Левенштейна помогает определить, какое слово известно или неизвестно. Подсчитываем вероятность редактирования между нимиn лемма и слово (до лемматизации). Если вероятность P (лемма |word) больше 50% [P (лемма |word) > 50%], то она засчитывается как неизвестные слова.

Рисунок 1. Предлагаемая работа

В.

P

D

-

P

РОКЕССИНГ

Для запуска алгоритма набор данных должен быть предварительным

-

обработанный.

Есть несколько предварительных

-

методы обработки проводятся в

БИКАС. Первый из них - Анафора, которая относится к слову

который используется ранее

в предложении, чтобы избежать повторения, например,

Местоимения. Это требует успешной идентификации и

разрешение НЛП. В предлагаемом BIQAS мы имеем

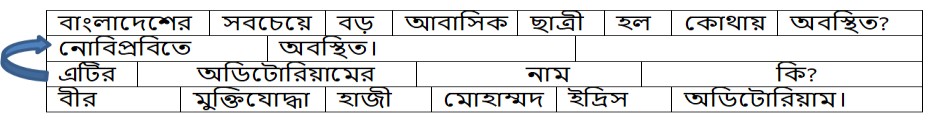
описан обзор работы, проделанной в области анафоры

разрешение, которое оказывает влияние на местоимения, ma

внутренне

личное местоимение с использованием алгоритма Хоббса‟.

Таблица.1: Рабочий процесс разрешения анафоры.



Здесь

"

‟(Ит)

есть

тем

местоимение

из

тем

имя

из

"

но-бибипроби

‟(НГТУ). Поэтому мы использовали "

‟ вместо "

но-бибипроби

‟

из предыдущего примера как Anaphora Resolution.

Очищающее слово относится к удалению нежелательного символа

который не имеет никакого отношения к информативным данным;

для

пример: двоеточие, точка с запятой, запятая, вопросительный знак,

восклицательный знак и другие знаки препинания.

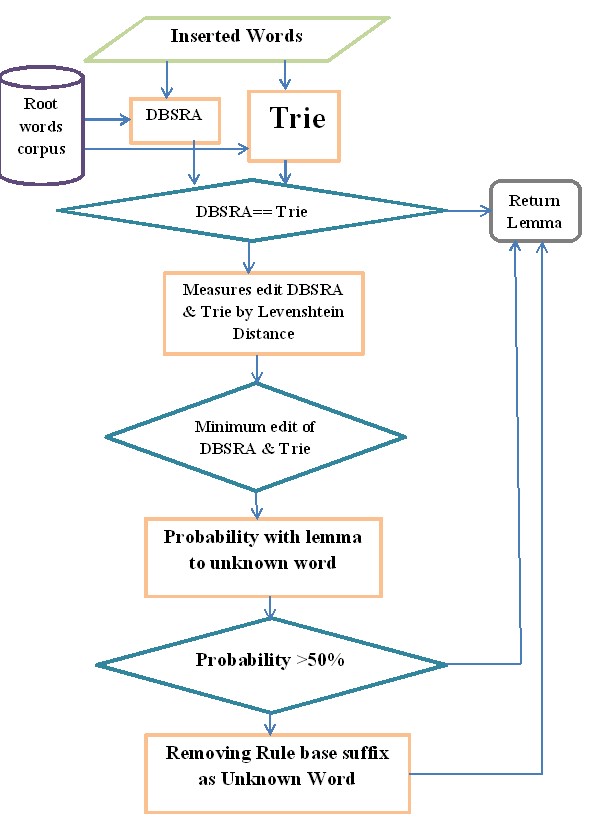


Рисунок 2. Рабочий процесс леммтиазации

Чтобы обработать неизвестные слова, мы установили корпус суффикса бенгальского языка; например:

(te), (che), (yer) и т.д. Самый длинный распространенный суффикс был удален из последней позиции неизвестного слова. Таким образом, мы получаем лемму или корень неизвестного слова.

Синонимы слов указывают на точное или почти одинаковое значение разных слов. Пользователи задают вопросы, имея слова, которые недоступны в информационных данных, но их значения есть. В этом смысле BIQAS может не отвечать правильно. Таким образом, обработка синонимичных слов имеет значительное действие в BIQAS, а также в понимании естественного языка (NLU).

Здесь " হ (большой)‟ является синонимом " ড় (большой)‟ и " হ' (большой) считается общим словом.

После предварительной обработки вопрос будет

Вопрос пользователя 1: Да.

[Где находится самый большой женский зал в Бангладеш?]

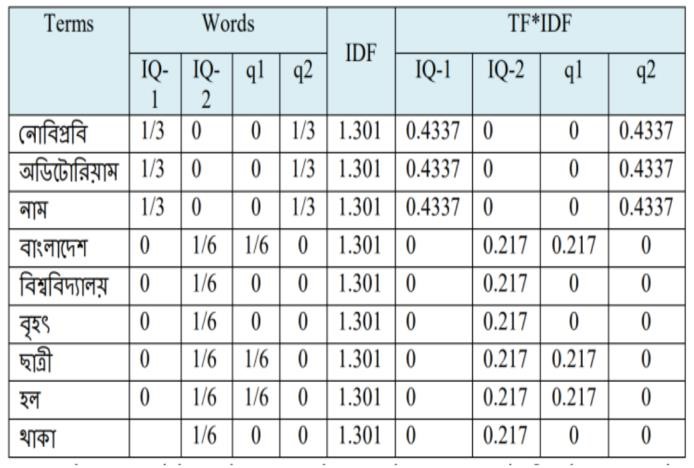
Вопрос пользователя-2: [Как называется его аудитория?]

# VI. СОЗДАНИЕ РЕЛАЦИОННОГО КОРАБЛЯ

## А. Косинусное сходство

Для того, чтобы представить слова пользовательских вопросов или информативные вопросы в численном виде, мы использовали метод векторизации TF-IDF модели. Для упрощения нашей задачи мы взяли два примера из рассматриваемого корпуса. Значение TF-IDF информатива

вопросы и вопросы пользователей были показаны в таблице



Зададим t

веса erm

и построить термин

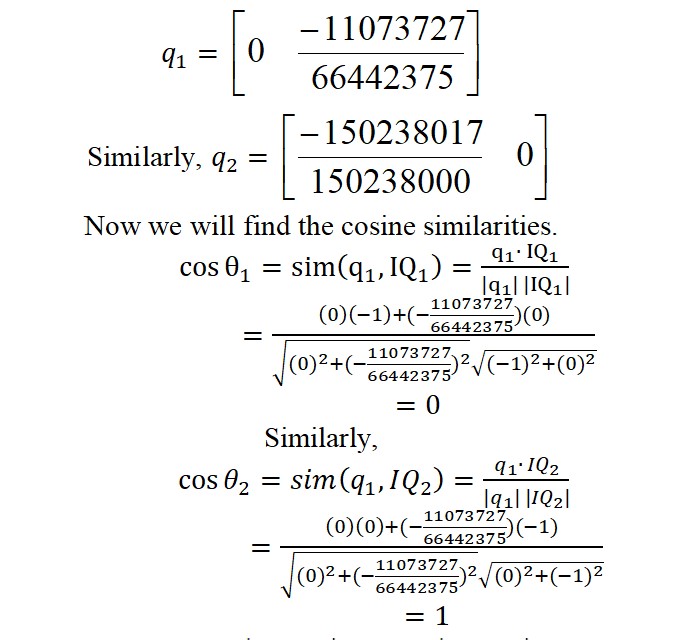
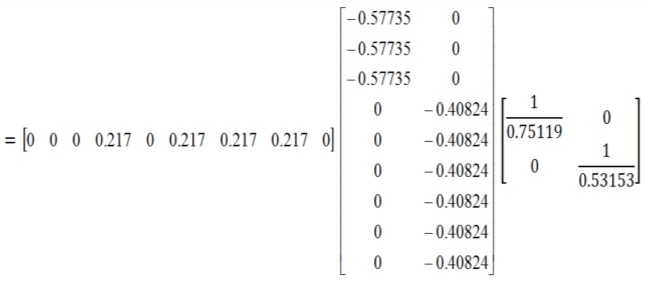
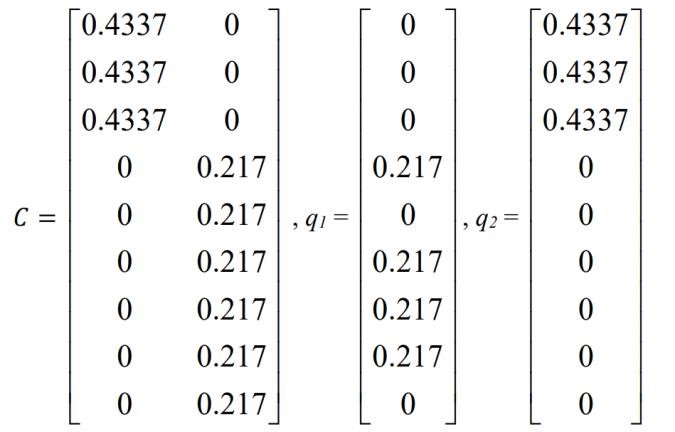
-

документ

матрица



и матрица запросов:



Мы видим, что



. Итак, пользователь

вопрос1 можно найти в информативном вопросе2, т.е.



.

Пользователь

вопрос

-

2

c

найти в информативном вопросе1,

то есть.

Я

𝑄

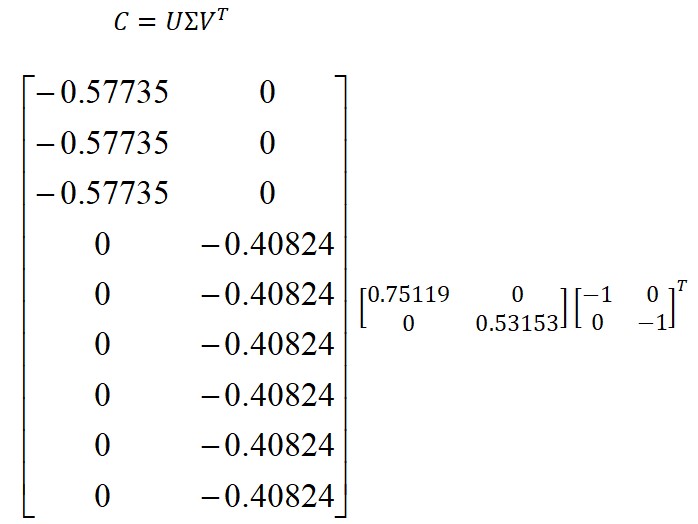
1.

Таблица 2: Расчет косинусного сходства

Теперь мы будем использовать SVD в матрице и найдем и



матрицы, где



Строка V содержит значение собственного вектора. Это координаты отдельных векторов документа. Следовательно



Чтобы найти новые векторные координаты запроса, у нас есть

 Теперь к первому вопросу:

## Б. Сходство Жаккара

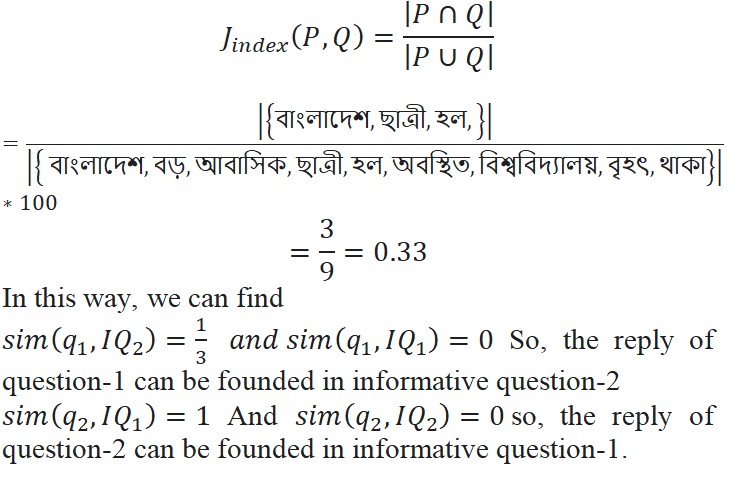
После предварительной обработки вопросы и данные могут быть выявлены в виде наборов

Теперь рассмотрим пример с использованием множественной нотации и диаграммы Венна

P = {yongoddadas, ri, aobic, chotri, hall, ribit}.

Q = {йонгоддадас, универсал, рихат, чотри, холл, тококос}

VII. EXPERIMENTS



*C.*

*Наивные байесовские эксперименты*

После предварительного

-

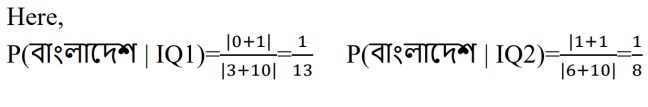
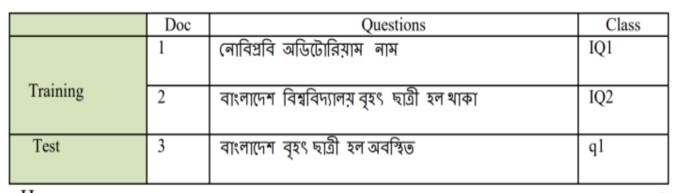
обрабатывая данные, мы преобразовали каждое слово в

срок

-

Матрица документа, затем вычисляемая вероятность

Таблица 3: Наивный расчет Байеса



Мы описали ряд экспериментов

o измерьте наши

предложена модель математических и статистических процедур

для БИКАС. В этом разделе, во-первых, мы представим вопросы

что мы нацелены на то, чтобы ответить на эксперименты и описать

экспериментальная установка. Затем мы обсуждаем спектакль и

результат

нашей предполагаемой работы.

*A*

*.*

*Корпус*

Для подтекста BIQAS мы описываем в основном пять

типы корпусов. В первом корпусе насчитывается 28 324 бенгальца.

корневые слова. Основной целью этого корпуса является лемматизация

Бенгальские слова. Второй, который содержит 382

Бенгальский

Stop words — это удаление стоп-слов из вставленных

документы и вопросы. Мы собрали 74 темы как

информативные документы, такие как информация о зале, отдел

информация, информация для преподавателей, библиотека, природа НГТУ, автобус

расписания и т.д.

Наука и техника Ноаджали

Университет (НГТУ), который является нашим третьим корпусом в качестве вопросов с

соответствующий ответ на вопрос о документе. В этой работе

Мы создали 3127 вопросов из наших вставленных

документы как наш четвертый корпус. Каждый вопрос co

ntains его

соответствующий ответ. Чтобы проверить наши данные, мы создали

вопросы по соответствующим 74 темам.

2852

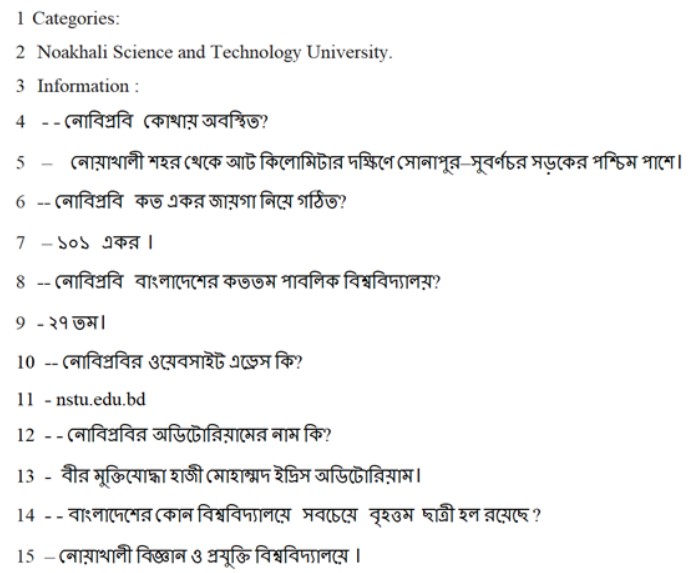
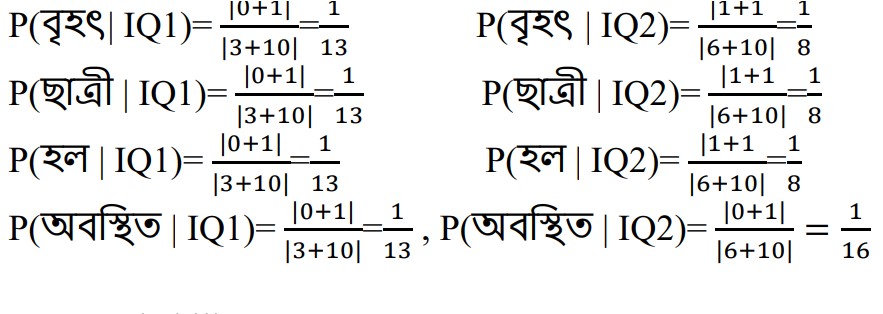


Рис.3: Сходство Жаккара

Рисунок 4. Вопросы и ответы (Данные)



*B. Экспериментальная установка*

Мы реализовали нашу предложенную модель и дистрибутив Anaconda с языком программирования Python 3.7 и выполнили их на ПК с Windows 10 с Intel Core i7, процессором (3,20 ГГц) и 8 ГБ памяти.

Python - это высокоуровневый объектно-ориентированный язык (ООП), который подходит для научной экспертизы и разработки инструментов. Мы использовали Анаконду в качестве распределения Python. Anaconda создает лучшую сцену для науки о данных с открытым исходным кодом, которая работает на Python. Для предварительной обработки мы использовали инструмент BLTK [3, 11], который является инструментарием бенгальского языка. BLTK также предоставляет TF-IDF, SVD, косинусное сходство, сходство Jaccard и наивный алгоритм Байеса.

## C.Результаты и анализ

Чтобы проверить предложенный нами BIQAS, мы создали 2852 вопроса в качестве тестовых данных из выбранных 74 тем Научно-технического университета Ноакхали (НГТУ) и получили 93,22% точности в косинусном подобии, 82,64% по Сходству Жаккара и 91,34% по наивному байесовскому классификатору.

## D. Сравнение английского чат-бота и BIQAS

Поскольку на бенгальском языке нет чат-бота, такого как BIQAS, есть и другой бенгальский чат-бот для сравнения с нашей работой. Наша работа в настоящее время является современной для бенгальского разведывательного бота. Итак, мы сравниваем BIQAS с двумя английскими чат-ботами, которые являются Neural Conversational Machine (NCM) и Cleverbot.

Какой у вас номер мобильного телефона? ( ?) Мицуку: Эта информация является конфиденциальной.

БиКАС: ।

।

Сколько Вам лет? ( ?) Мицуку: Мне 18 лет.

БИКАС: B .

Каков ваш адрес? ( ?) Мицуку: Я в Лидсе.

БИКАС:

।

Какой у вас номер мобильного телефона? ( ?) Мицуку: Эта информация является конфиденциальной.

БиКАС: ।

।

Сколько Вам лет? ( ?) Мицуку: Мне 18 лет.

БИКАС: B .

Каков ваш адрес? ( ?) Мицуку: Я в Лидсе.

БИКАС:

।

# VIII. CONCLUSION & FUTURE WORKS

Основной задачей этого проектного документа является внедрение бенгальского разведывательного бота для поиска информации. Мы показали теоретическую и экспериментальную методологию предлагаемой нами работы. В этой научной статье мы описали три процедуры с использованием машинного обучения, математики и статистики. Чтобы установить полную методологию, мы следовали некоторым процедурам, таким как предварительная обработка, сокращение времени и пространства, и установили связь между информацией и вопросами.

В будущем предлагаемая система BIQAS может быть включена для целей образования, промышленности, бизнеса и личных задач с системой голосового ответа. Прогресс может быть сформирован с помощью алгоритмов глубокого обучения, таких как рекуррентная нейронная сеть (RNN), путем обработки BNLP.

# РЕФЕРЕНЦИИ

1. Кларк, Питер и др. «Думаете, вы решили вопрос, ответив на вопрос? попробуй дугу, вызов рассуждения ai2." arXiv препринт arXiv:1803.05457 (2018).
2. Азиат, Амир Хосейн. Основа для ответов на вопросы на основе онтологии с применением к данным паразитов. Дисс. Университет Джорджии, Афины, Джорджия, США, 2013.
3. Kowsher, Md, et al. "Doly: Bengali Chatbot for Bengali Education." 2019 1-я Международная конференция по достижениям в области науки, техники и робототехники (ICASERT). IEEE, 2019.
4. Шах, Урви и др. «Поиск информации в семантической паутине». Материалы одиннадцатой международной конференции по управлению информацией и знаниями. АКМ, 2002.
5. Абди, Асад, Норисма Идрис и Захра Ахмад. «QAPD: система ответов на вопросы, основанная на онтологии, в области физики». Soft Computing 22.1 (2018): 213-230.
6. Lopez, Vanessa, et al. "AquaLog: Ontology-based question answering system for organizational semantic intranets" («АкваЛог: онтологическая система ответов на вопросы для организационных семантических интрасетей»). Веб-семантика: наука, сервисы и агенты во всемирной паутине 5.2 (2007): 72-105.
7. Радж, П.C. "Архитектура основанной на онтологии доменной системы ответов на вопросы естественного языка." arXiv препринт arXiv:1311.3175 (2013).
8. Симмонс, Роберт Ф. «Системы вопросов и ответов на естественный язык: 1969». Сообщения ACM 13.1 (1970): 15-30.]
9. Yu, Zheng-Tao, et al. «Извлечение ответов для китайской системы ответов на вопросы, основанной на латентном семантическом анализе».

КИТАЙСКИЙ ЖУРНАЛ КОМПЬЮТЕРОВ-КИТАЙСКОЕ ИЗДАНИЕ29.10

(2006).

1. Уиттакер, Эдвард, Садаоки Фуруи и Дитрих Клаков. «Подход статистической классификации к ответам на вопросы с использованием веб-данных». 2005 Международная конференция по кибермирам (CW'05). IEEE, 2005. 11.
2. Kowsher, Md, Imran Hossen и SkShohorab Ahmed. «Бенгальская информационно-поисковая система (BIRS)». Международный журнал по вычислениям на естественном языке (IJNLC) 8.5 (2019).
3. Dongfeng, Cai, et al. «Китайская система автоматических ответов на вопросы на основе Интернета». Четвертая Международная конференция по компьютерным и информационным технологиям, 2004 год. ЦИТ'04.. IEEE, 2004.
4. Сталин, Шалини, Раджив Пандей и Раджу Барскар. "Веб-приложение для системы ответов на вопросы на хинди." Международный журнал электроники и компьютерных наук 2.1 (2012): 72-78.
5. Чон, Дживун, У. Брюс Крофт и Джун Хо Ли. «Поиск семантически похожих вопросов на основе их ответов». Материалы 28-й ежегодной международной конференции ACM SIGIR по исследованиям и разработкам в области информационного поиска. АКМ, 2005.
6. Ван, Вэй, Байчуань Ли и Ирвин Кинг. «Улучшение поиска вопросов в сообществе вопросов, отвечающих на вопросы с помощью ранжирования меток». Международная совместная конференция по нейронным сетям 2011 года. IEEE, 2011.