FAQ RETRIEVAL SYSTEM USING WEIGHTED EDIT DISTANCE FOR NOISY QUERIES

MADE BY:-

DHRUVI MEHTA(202318003)

PRACHI MEHTA(202318008)

RIYA DAVE(202318011)

SATYAM MARAVANIYA(202318026)

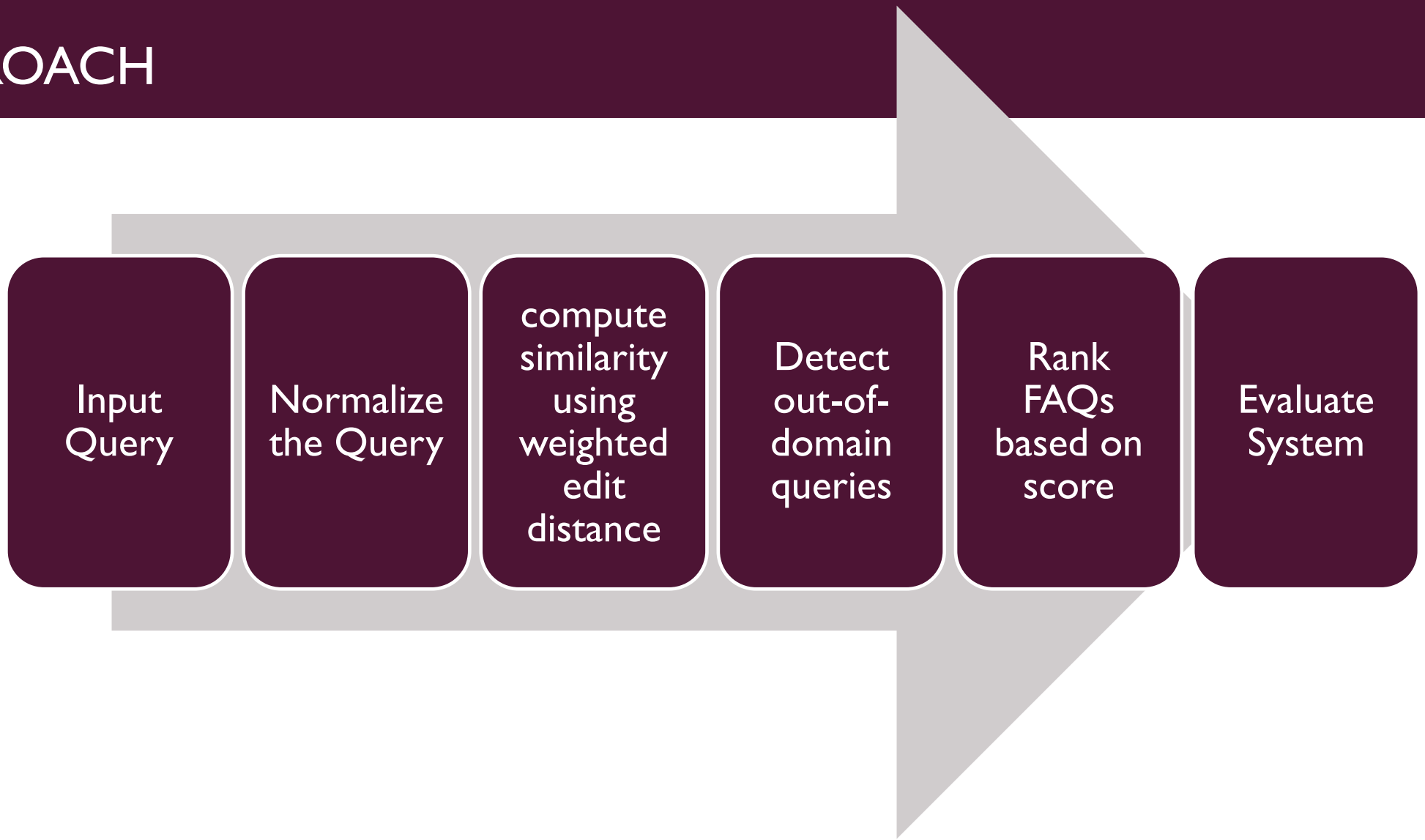
PROBLEM STATEMENT

- The goal of this project is to develop a system that can accurately match user queries to FAQs, even when those queries contain errors like typos or different phrasings.
- This will be achieved using a method called weighted edit distance, which improves the system's ability to handle "noisy" queries by comparing them to a database of FAQs and returning the most relevant match.
- Additionally, the system will identify whether a query is out-of-domain, meaning it cannot be answered by the current FAQs, using a simple classifier.
- The system has shown strong results, correctly matching the majority of in-domain and out-of-domain queries.

YAHOO! ANSWERS DATASET

- **Dataset Size**
  - 1.4 million training samples
  - 60,000 testing samples
- **Categories**
  - Data is categorized into 10 different sections
- **Content Includes**
  - Question titles
  - Question content
  - Best answers

APPROACH


Input Query
Normalize the Query
compute similarity using weighted edit distance
Detect out-of-domain queries
Rank FAQs based on score
Evaluate System

INPUT QUERY

5	why doesn't an optical mouse work on a glass table?
6	What is the best off-road motorcycle trail ?
3	What is Trans Fat? How to reduce that?
7	How many planes Fedex has?
7	In the san francisco bay area, does it make sense to rent or buy ?
5	What's the best way to clean a keyboard?
2	Why do people blush when they are embarrassed?
8	Is Lin Qingxia (aka Brigitte Lin) "the most beautiful woman in Chinese cinema?"
5	What is the origin of "foobar"?
2	How the human species evolved?
4	Who said the statement below and what does it mean?
4	How do I find an out of print book?
7	What are some tips on finding a good mortgage broker?
5	what's the best way to create a bootable windos/dos CD?
9	what is the reason for the increasing divorce percentage in the western world?
2	What is an "imaginary number"?
2	Faxing a pizza
8	What are good sources to find out about new gospel artists?
2	space missions
2	How a black hole is formed?
2	Heavy water

NORMALIZE THE QUERY

- **Input Validation**: Checks if the input is a string. If not, it returns an empty list.
- **Lowercase Conversion**: Converts all characters in the text to lowercase to ensure uniformity.
- **Removal of Non-ASCII Characters and Digits**:
- Removes all numbers and replaces non-ASCII characters with spaces.
- **Tokenization**: Splits the text into individual words (tokens) using word_tokenize
- **Stopword Removal**: Removes common, non-informative words (e.g., "the", "and") using the stopwords set.
- **Stemming**: Reduces each word to its root form using the PorterStemmer (e.g., "running" becomes "run").
- **Deduplication**: Removes duplicate tokens while preserving their original order.

	class	question_title	question_content	best_answer	question_title_tokens	question_content_tokens	best_answer_tokens
0	1	religious ppl pray! what do they say?	hello!\nif u pray God perhaps u use some sente...	Whenever I pray I just speak from my heart. He...	[religi, ppl, pray, !, say, ?]	[hello, !, \nif, u, pray, god, perhap, use, se...	[whenev, pray, speak, heart, ., alreadi, know,...
1	1	Please honestly share - What was the craziest ...	Okay - it's only fair that I share my experien...	pooped outside in my neb.s lawn my bro locked ...	[pleas, honestli, share, -, craziest, thing, e...	[okay, -, 's, fair, share, experi, first, ., c...	[poop, outsid, neb., lawn, bro, lock, ,]
2	1	Combs of HANNITY AND COMBS is graduate of Hofs...	NaN	well at least I know who Hannity and combs ar...	[comb, hanniti, graduat, hofstra, univers, lon...	[]	[well, least, know, hanniti, comb, !, lol, ...
3	1	Any Polish people here? can u translate this s...	what does it mean in English? "A mam sie ja Cu...	I am Polish. This sentence means: 'Well, I fe...	[polish, peopl, ?, u, translat, sentenc, en, p...	[mean, english, ?, ``, mam, sie, ja, cudowni, ...	[polish, ., sentenc, mean, :, 'well, ,, feel, ...
4	1	Should women work in construction, e.g. masonr...	Hard labor for better pay I am questioning th...	certainly i no lots of woman who are skilled i...	[women, work, construct, ,, e.g, ., masonri, c...	[hard, labor, better, pay, question, like, hoo...	[certainli, lot, woman, skill, construct, trad...
...	...	...	...	...	...	...	...
99995	10	What does sex offender mean in your mind?	What does the question mean in your mind? Why ...	What a sex offender means in my mind is a pers...	[sex, offend, mean, mind, ?]	[question, mean, mind, ?, peopl, think, offend...	[sex, offend, mean, mind, person, made, mistak...
99996	10	where can i find the history of tariff, its or...	NaN	a book about US history, or just contact me; i...	[find, histori, tariff, ,, origin, ?]	[]	[book, us, histori, ,, contact, ;, studi, school]
99997	10	Does anyone see any similarities between Bush ...	http://www.opposingdigits.com/openeyes/	None whatsoever.	[anyon, see, similar, bush, hitler, ?]	[http, :, //www.opposingdigits.com/openeyes/]	[none, whatsoev, .]
99998	10	How can we make chat rooms safer from sexual p...	NaN	By actually spending time with your children w...	[make, chat, room, safer, sexual, predat, ?]	[]	[actual, spend, time, children, onlin, ., othe...

COMPUTE SIMILARITY USING WEIGHTED DISTANCE

- **LCSRatio (Longest Common Subsequence Ratio)**:
- Measures how much of one term (tq) is sequentially matched within another (tf).
- Formula:

$$\text{LCSRatio}(tq,tf) = \frac{\text{length of LCS}(tq,tf)}{\text{length } (tf)}$$

- **Weighted Edit Distance**
- Introduces different costs for string operations based on their likelihood queries:
  WeightedEditDistance(tq,tf)=0.3×(vowel insertions)+0.5×(consonant insertions)+0.3×(end insertions)+0.5×(vowel substitutions)+2.0×(deletions)

CALCULATE SCORE FOR FAQ MATCHING

- **Idf(tf)**
- IDF: Reduces the weight of terms commonly found across many documents.

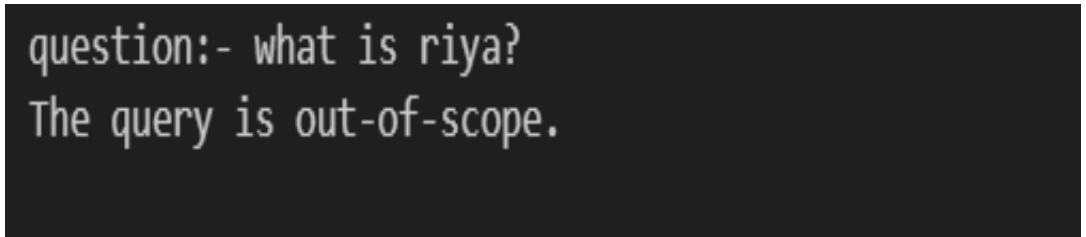
$$\mathrm{IDF}(t) = \log\left(\frac{|\mathcal{D}|}{1 + |\{d \in \mathcal{D} : t \in d\}|}\right)$$

|D|: Total number of documents in the corpus
|{d∈D:t∈d}|: Number of documents containing the term t

- **Similarity Metric**
- Combines **LCSRatio**, **Weighted Edit Distance**, and **IDF (Inverse Document Frequency)**:
- $\mathrm{Sim}(tq,tf) = \mathrm{ID}\frac{\mathrm{IDF}(tf) \times \mathrm{LCSRatio}(tq,tf)}{\mathrm{WeightedEditDistance}(tq,tf)}$

DETECT OUT-OF-DOMAIN QUERIES

- Each term (token) in the query is assigned a score based on its importance using **Inverse Document Frequency (IDF)** values.
- IDF values prioritize **rare terms** over **common ones**, meaning that terms that appear less frequently in a larger dataset are considered more significant.
- If a token is not found in the IDF dictionary, it contributes a score of **0**.
- A predefined **threshold** is set to determine the classification of the query.
- If the total in-scope score is **below the threshold** the query is classified as **out-of-scope**.
- If the total score **meets or exceeds the threshold** the query is classified as **in-scope**.



RANK FAQS BASED ON SCORE

- Rank all FAQs by their computed similarity scores in descending order.
- Select the top 5 FAQs with the highest scores as the most relevant to the query.

```
Your Query: where can i fid the surce of virus?

Top Relevant FAQs:
Category: 5, Score: 4614095937.6687, Reciprocal_Rank: 1.0000, Answer: Acctually you can't
Category: 5, Score: 2603050998.7073, Reciprocal_Rank: 0.5000, Answer: Trojan Horses are impostor files that claim to be something desirable but, in fact, are malicious. Rather than ins
Category: 5, Score: 2603050998.4084, Reciprocal_Rank: 0.3333, Answer: CMOS viruses are the worst. They get into your BIOS which means you have to replace the hardware. The second worst
Category: 5, Score: 2603050998.3351, Reciprocal_Rank: 0.2500, Answer: ur IP address has been noted down \nthe next time u talk abt creating a virus ur system would be blown away
Category: 5, Score: 2603050998.2603, Reciprocal_Rank: 0.2000, Answer: mostly companies of anti-virus software \nto sell their software better
```

```
Your Query: What does RCAC mean?

Top Relevant FAQs:
Category: 6, Score: 2641578858.1962, Reciprocal_Rank: 1.0000, Answer: that - pron.\n\n   1.\n        a. Used to refer to the one designated, implied, mentioned, or understood: What kin
Category: 10, Score: 2094385904.6558, Reciprocal_Rank: 0.5000, Answer: ROYAL CANADIEN AIR CORPS
Category: 8, Score: 1761052573.9221, Reciprocal_Rank: 0.3333, Answer: male orgasm= 20 seconds of feeling like a God.\nfemale orgasm = earth moving, body shaking, summit climbing, inten
Category: 1, Score: 1761052573.8889, Reciprocal_Rank: 0.2500, Answer: Merriam webster online dictionary:\nEtymology: Middle English, from Late Latin ethnicus, from Greek ethnikos natio
Category: 4, Score: 1761052573.7793, Reciprocal_Rank: 0.2000, Answer: moron means idiot or dumbo.
```

```
Your Query: What's Cricket?

Top Relevant FAQs:
Category: 6, Score: 4148999851.6611, Reciprocal_Rank: 1.0000, Answer: A PASSION.\nA MIND BLOWER.\nA SEDATOR.\nA TREASURE.\nA KILLER.\nA ENTERTAINER.\nA HOBBY.\nA FRIENDSHIP MAKER.\nA F
Category: 6, Score: 2765999901.5997, Reciprocal_Rank: 0.5000, Answer: Type "history" and "cricket" into the UK or Australian based yahoo. What are you, stoopid?\n\nCheck it out, first
Category: 6, Score: 2765999901.5599, Reciprocal_Rank: 0.3333, Answer: Cricket began in Northern Europe sometime after the Roman Empire and before the Normans invaded England.\nAll rese
Category: 6, Score: 2765999901.5406, Reciprocal_Rank: 0.2500, Answer: TO MAKE YOU ASK SUCH A QUESTION.
Category: 6, Score: 2765999901.5294, Reciprocal_Rank: 0.2000, Answer: Me....cant even hold a cricket bat at the right angle...duh....i LUV football
```

EVALUATE SYSTEM

1. **Ground Truth Creation**:
   - Define a set of queries and their correct answers (ground truth).
   - Label the FAQs in the database as **relevant** or **irrelevant** for each query.
2. **Compare Results**:
   - Retrieve the top 5 results from your system for each query.
   - Check these results against the ground truth to determine relevance.
3. **Compute Metrics**:
   - **Accuracy**: How accurate is your model.
   - **MRR**: Rank of the first relevant result.

```
Mean Reciprocal Rank (MRR): 0.6667
Accuracy: 0.567
```

THANK YOU