

2024 Information Retrieval and Extraction

HW 1

Task introduction

- Measure document relevance to a query
 - Implement **vector model** and **BM25** using only **numpy** to compare similarity of questions and documents
 - pandas, csv and matplotlib are available for data analysis and pre-processing
- Requirement
 - Upload your submission to Kaggle
 - Submit a report and your source code to E3
- Deadline is 10/25 (Fri.) 23:59, no late submission

Dataset

- documents_data.csv
 - Documents which contain website information and corresponding Document IDs
 - [link](#)
- train_question.csv
 - Contains questions and corresponding question IDs
 - [link](#)
- test_question.csv
 - Contains questions that need to be used for prediction
 - [link](#)

Training Data

Question ID	Question	Answer ID
1	which is the most common use of opt-in e-mail marketing	1
2	how i.met your mother who is the mother	2
3	what type of fertilisation takes place in humans	3
4	who had the most wins in the nfl	4
5	what happened to the lost settlement of roanoke	5
6	what are the different regions of africa and how do they differ	6
7	who played mantis guardians of the galaxy 2	7
8	who did the voice of the magician in frosty the snowman	8
9	what indian tribe did the acadians form friendships and alliances with	9
10	what is considered the outer banks in north carolina	10
11	who is deputy cm of j and k	11
12	bangko sentral ng pilipinas (central bank of the philippines)	12
13	actual time taken from start to finish to produce one unit of value	13
14	i had trouble in getting to solla sellew	14
15	ed sheeran i ' m in love with your body	15
16	the nashville sound brought a polished and cosmopolitan sound to country music by	16
17	the new atlantis a journal of technology and society	17
18	list the seven wonders of the modern world	18
19	rolling stone top bands of all time list	19
20	what channel is the premier league on in france	20

Testing Data

Question ID	Question
1	when was bank of america founded in the us
2	who plays vanessa baxter on last man standing
3	what is the big cross in effingham illinois
4	which statement accurately describes eastern jerusalem according to the article
5	who is buried in the queen vic eastenders
6	who was the winner of super dancer chapter 1
7	smoking in a car with a child law alberta
8	who put together the list of amendments that became the bill of rights
9	what clues to culture can be traced in treatments of the human figure in prehistoric art
10	when was the last time green bay packers won a superbowl
11	when did croatia finish 3rd in the world cup
12	who is the leading scorer in the nba
13	who has scored the most goals in the english premier league
14	where was the author when he wrote the star spangled banner
15	how many assist does ozil have in his career
16	how much money does argentina make from tourism
17	is the game show the chase still on tv
18	who is the voice for clifford the big red dog
19	who sang the song life in the fast lane
20	when did the name great britain come about

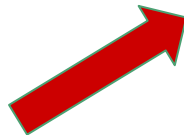
Requirements & Scoring Metrics

Please implement both the **vector model** and **BM25** as the scoring basis to find the most similar Documents, and output the **first three** most similar Document IDs.

The final scoring result will be conducted as **Recall@3**. That is, if the top three most similar document IDs have a hit, it will be 1, if they have no hits, it will be 0, and finally the average will be the result.



Kaggle Submission

- [Kaggle link](#)
- Display team name : <student ID>
- Submission format
 - A 401*2 .csv file, first row is for the column name and the last 400 rows for your result.
 - Column name must be **index** and **answer**.
- There is one simple baseline and one strong baseline. Beat them to achieve a higher score.



index	answer
1	0 0 0
2	0 0 0
3	0 0 0
4	0 0 0
5	0 0 0
6	0 0 0
7	0 0 0

Three numbers
separated by
spaces

#	Team	Members	Score	Entries	Last
	Strong Baseline		0.71000		
	Simple Baseline		0.63500		

Kaggle Submission

- The scoring metric is **Recall@3**.
- You can submit at most 5 times each day.
- You can choose 3 of the submissions to be considered for the private leaderboard, or will otherwise default to the best public scoring submissions.
You can only view your private leaderboard score after the competition has ended.
- Public leaderboard is calculated with 50% of the test data, and private leaderboard is calculated with other 50% of the test data, so the final standings may be different.
- Please **tune your model parameters using your own validation set** instead of adjusting parameters based on the public leaderboard. Otherwise, it's easy to overfit, leading to poor performance on the private leaderboard.

Change your team name

2024 Information Retrieval & Extraction Homework1

Measure document relevance to a query



Settings Overview Data Discussion Leaderboard Rules Team

Remember to change the team name to <student ID>, or there will be a deduction of 5 points for HW 1.

Your Team

Everyone that competes in a Competition does so as a team - even if you're competing by yourself. [Learn more.](#)

General

TEAM NAME

Team Name



Report Submission

Answer the following 3 questions:

1. What kind of pre-processing did you apply to the document data or question text? Additionally, please discuss how different preprocessing methods affected the performance of the models?
2. Please provide details on how you implemented the vector model and BM25.
3. Compare the strengths and weaknesses of the vector model and BM25. What factors might account for the differences in their performance?

Please answer the questions in detail to receive full points for each question.

Grading policy

- Kaggle (70%)
 - 30% based on the public leaderboard score and 70% based on the private leaderboard score
 - Leaderboard score consists of basic score and ranking score
 - Basic score :
 - Over strong baseline : 55
 - Over simple bassline : 40
 - Under simple baseline : 25
 - Ranking score:
 - $15 - (15/N) * (\text{ranking} - 1)$, N=numbers of people in the interval
- Report (30%)
 - 10 for each quesiton

E3 Submission

Submit your source code and report to E3 before 10/25 (Fri.) 23:59.

No late submission !

- Format

- source code : HW1_<student ID>.py or HW1_<student ID>.ipynb
- report : HW1_<student ID>.pdf

If you have any question about HW 1, please feel free to contact with TA : YEN-CHUN HUANG
through email yenchun.cs12@nycu.edu.tw

Have Fun !

