CS F469 INFORMATION RETRIEVAL

Domain Specific Retrieval System

“F.R.I.E.N.D.S” Dialogues Information Retrieval System

TEAM MEMBERS

Raaed Ahmed

Ritika Reddy

Anurag Aribandi

Anvitha Nallan

Dataset

F.R.I.E.N.D.S TV SHOW CORPUS

A collection of all the conversations that occurred over 10 seasons of Friends, a popular American TV sitcom that ran in the 1990s.

Across the 10 seasons there are 236 episodes, 3,107 scenes (conversations), 67,373 utterances, and 700 characters (users).

The available metadata varies by seasons, but can include: character entities, emotion, a tokenized version of the text, caption information, and notes about the transcript.

PRE-PROCESSING

TOKENIZATION

We have used the RegexpTokenizer which is imported from nltk.tokenize. RegexpTokenizer splits a string into substrings using a regular expression. We have used the regular expression "[\w']+".

STOP WORDS

Stop words are generally the most common words in a language. We have used stopwords from nltk.corpus package to construct our stopset.

STEMMING

Stemming is the process of producing root/base word of the given word in the query. We have used Porter Stemmer which is imported from nltk.stem.

LEMMATIZATION

Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form. We have used the WordNetLemmatizer from nltk.stem package.

RUNNING TIMES

Pre-Processing documents :	36.26308250427246 s
Pre-Processing Query :	0.000998258590698 s
Building Inverted Index :	0.4946365356445312 5 s
Retrieving Documents :	0.02593207359313965 s

DATA STRUCTURES USED

The Data structures used in the assignment are **Dictionary, Lists and Data Frame**

- **df** - Data Frame which contains the entire show's dialogue corpus
- **dialogue** - Dictionary of lists which contains the scene_ids mapped to transcripts of dialogues in that scene.
- **docs** - Dictionary which maps scene_ids to the normalized tokens derived from the transcripts.
- **inverted index** - Dictionary which contains the presence of each preprocessed word which is mapped to each occurrence of the word.
- **score_matrix** - Dictionary of tuples which contains the number of words from the query which occurs in the document followed by the tf-idf score
- **sorted_indices** - contains the list of documents sorted according to their TF-IDF scores.

ORIGINALITY

Different options of selecting the way the term frequency should be calculated :

- ★ Natural Term Frequency
- ★ Logarithm Term Frequency
- ★ Augmented Term Frequency
- ★ Boolean Term Frequency

N (Natural)	$tf_{t,d}$
L (Logarithm)	$1 + \log(tf_{t,d})$
A (Augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{\mathrm{Max}_t (tf_{t,d})}$
B (Boolean)	1 if $tf_{t,d}>0$ 0 otherwise