CSE508 Information Retrieval

Winter 2024

Report Assignment-1

Ankit Kumar 2021015

> Explainations are already present in the notebook's markdown cells

Question 1: Preprocessing

- Approach:
    - iterate through the documents presented in `text_files` directory and for first five files, do the individual preprocessing. preprocessing explained in Ques 2 in details.
- Methodologies:
    - used `os` and then extracted the file list and then preprocessed in the same sequence mentioned in the question
- Assumptions: NA
- Results:

```
--------------------------------------------------------------------------
file1.txt
Original Text: Loving these vintage springs on my vintage strat. They have a good
tension and great stability. If you are floating your bridge and want the most out
of your springs than these are the way to go.
1. Lower Case:  loving these vintage springs on my vintage strat. they have a good
tension and great stability. if you are floating your bridge and want the most out
of your springs than these are the way to go.
2. tokenized:  loving these vintage springs on my vintage strat . they have a good
tension and great stability . if you are floating your bridge and want the most
out of your springs than these are the way to go .
3. Stopwords Removed: loving vintage springs vintage strat . good tension great
stability . floating bridge want springs way go .
4. Punctuation Removed: loving vintage springs vintage strat  good tension great
stability  floating bridge want springs way go
5. Space Removed: loving vintage springs vintage strat good tension great
stability floating bridge want springs way go
--------------------------------------------------------------------------
file2.txt
Original Text: Works great as a guitar bench mat. Not rugged enough for abuse but
if you take care of it, it will take care of you. Makes organization of workspace
much easier because screws won't roll around. Color is good too.
1. Lower Case:  works great as a guitar bench mat. not rugged enough for abuse but
if you take care of it, it will take care of you. makes organization of workspace
```

much easier because screws won't roll around. color is good too.
2. tokenized: works great as a guitar bench mat . not rugged enough for abuse but if you take care of it , it will take care of you . makes organization of workspace much easier because screws wo n't roll around . color is good too .
3. Stopwords Removed: works great guitar bench mat . rugged enough abuse take care , take care . makes organization workspace much easier screws wo n't roll around . color good .
4. Punctuation Removed: works great guitar bench mat rugged enough abuse take care take care makes organization workspace much easier screws wo nt roll around color good
5. Space Removed: works great guitar bench mat rugged enough abuse take care take care makes organization workspace much easier screws wo nt roll around color good
---
file3.txt
Original Text: We use these for everything from our acoustic bass down to our ukuleles. I know there is a smaller model available for ukes, violins, etc.; we haven't yet ordered those, but these will work on smaller instruments if one doesn't extend the feet to their maximum width. They're gentle on the instruments, and the grippy material keeps them secure.

The greatest benefit has been when writing music at the computer and needing to set a guitar down to use the keyboard/mouse - just easier for me than a hanging stand.

We have several and gave one to a friend for Christmas as well. I've used mine on stage, and it folds up small enough to fit right in my gig bag.
1. Lower Case: we use these for everything from our acoustic bass down to our ukuleles. i know there is a smaller model available for ukes, violins, etc.; we haven't yet ordered those, but these will work on smaller instruments if one doesn't extend the feet to their maximum width. they're gentle on the instruments, and the grippy material keeps them secure.

the greatest benefit has been when writing music at the computer and needing to set a guitar down to use the keyboard/mouse - just easier for me than a hanging stand.

we have several and gave one to a friend for christmas as well. i've used mine on stage, and it folds up small enough to fit right in my gig bag.
2. tokenized: we use these for everything from our acoustic bass down to our ukuleles . i know there is a smaller model available for ukes , violins , etc . ; we have n't yet ordered those , but these will work on smaller instruments if one does n't extend the feet to their maximum width . they 're gentle on the instruments , and the grippy material keeps them secure . the greatest benefit has been when writing music at the computer and needing to set a guitar down to use the keyboard/mouse - just easier for me than a hanging stand . we have several and gave one to a friend for christmas as well . i 've used mine on stage , and it folds up small enough to fit right in my gig bag .
3. Stopwords Removed: use everything acoustic bass ukuleles . know smaller model available ukes , violins , etc . ; n't yet ordered , work smaller instruments one n't extend feet maximum width . 're gentle instruments , grippy material keeps secure . greatest benefit writing music computer needing set guitar use keyboard/mouse - easier hanging stand . several gave one friend christmas well . 've used mine stage , folds small enough fit right gig bag .
4. Punctuation Removed: use everything acoustic bass ukuleles know smaller model

```
available ukes  violins  etc   nt yet ordered  work smaller instruments one nt
extend feet maximum width  re gentle instruments  grippy material keeps secure
greatest benefit writing music computer needing set guitar use keyboardmouse
easier hanging stand  several gave one friend christmas well  ve used mine stage
folds small enough fit right gig bag
5. Space Removed: use everything acoustic bass ukuleles know smaller model
available ukes violins etc nt yet ordered work smaller instruments one nt extend
feet maximum width re gentle instruments grippy material keeps secure greatest
benefit writing music computer needing set guitar use keyboardmouse easier hanging
stand several gave one friend christmas well ve used mine stage folds small enough
fit right gig bag
--------------------------------------------------------------------------
file4.txt
Original Text: Great price and good quality.  It didn't quite match the radius of
my sound hole but it was close enough.
1. Lower Case:  great price and good quality.  it didn't quite match the radius of
my sound hole but it was close enough.
2. tokenized:  great price and good quality . it did n't quite match the radius of
my sound hole but it was close enough .
3. Stopwords Removed: great price good quality . n't quite match radius sound hole
close enough .
4. Punctuation Removed: great price good quality  nt quite match radius sound hole
close enough
5. Space Removed: great price good quality nt quite match radius sound hole close
enough
--------------------------------------------------------------------------
file5.txt
Original Text: I bought this bass to split time as my primary bass with my Dean
Edge. This might be winning me over. The bass boost is outstanding. The active
pickups really allow you to adjust to the sound you want. I recommend this for
anyone. If you're a beginner  like I was not too long ago, it's an excellent bass
to start with. If you're on tour and/or music is making you money, this bass will
be beatiful on stage. The color is a bit darker than in the picture. But, all
around, this is a great buy.
1. Lower Case:  i bought this bass to split time as my primary bass with my dean
edge. this might be winning me over. the bass boost is outstanding. the active
pickups really allow you to adjust to the sound you want. i recommend this for
anyone. if you're a beginner  like i was not too long ago, it's an excellent bass
to start with. if you're on tour and/or music is making you money, this bass will
be beatiful on stage. the color is a bit darker than in the picture. but, all
around, this is a great buy.
2. tokenized:  i bought this bass to split time as my primary bass with my dean
edge . this might be winning me over . the bass boost is outstanding . the active
pickups really allow you to adjust to the sound you want . i recommend this for
anyone . if you 're a beginner like i was not too long ago , it 's an excellent
bass to start with . if you 're on tour and/or music is making you money , this
bass will be beatiful on stage . the color is a bit darker than in the picture .
but , all around , this is a great buy .
3. Stopwords Removed: bought bass split time primary bass dean edge . might
winning . bass boost outstanding . active pickups really allow adjust sound want .
recommend anyone . 're beginner like long ago , 's excellent bass start . 're tour
and/or music making money , bass beatiful stage . color bit darker picture . ,
around , great buy .
4. Punctuation Removed: bought bass split time primary bass dean edge  might
```

```
winning  bass boost outstanding  active pickups really allow adjust sound want
recommend anyone  re beginner like long ago  s excellent bass start  re tour andor
music making money  bass beatiful stage  color bit darker picture   around  great
buy
5. Space Removed: bought bass split time primary bass dean edge might winning bass
boost outstanding active pickups really allow adjust sound want recommend anyone
re beginner like long ago s excellent bass start re tour andor music making money
bass beatiful stage color bit darker picture around great buy
```

Question 2: Unigram Inverted Index and Boolean Queries

- In-Between Step- Preprocessing complete dataset after 5 files. Each file is lowercased, tokenized, removed stop words like `this`, `that` etc, then punctuations are removed and after everything, blank spaces are removed.
- Approach:
    - created a dictionary datastructure using nested for loop where words are keys and values are all the documents they are in.
    - in a for loop queries are solved and stored in a temparory set resolving from left to right without considering traditional priorities.
- Methodologies:
    - brute force method by iterating through each doc
- Assumptions: words are already processed
- Results: A pickle file is saved which is a serialized version of the inverted_index dictionary.

```
Query 1: vintage OR love
Number of documents retrieved for query 1 : 126
Names of documents retrieved for query 1 : file1.txt, file23.txt, file29.txt,
file31.txt, file37.txt, file46.txt, file48.txt, file51.txt, file55.txt,
file57.txt, file64.txt, file77.txt, file111.txt, file114.txt, file119.txt,
file121.txt, file122.txt, file140.txt, file146.txt, file148.txt, file150.txt,
file156.txt, file171.txt, file178.txt, file194.txt, file197.txt, file209.txt,
file212.txt, file214.txt, file223.txt, file240.txt, file243.txt, file247.txt,
file269.txt, file273.txt, file278.txt, file279.txt, file280.txt, file289.txt,
file304.txt, file305.txt, file314.txt, file324.txt, file331.txt, file333.txt,
file340.txt, file349.txt, file369.txt, file384.txt, file396.txt, file397.txt,
file400.txt, file403.txt, file411.txt, file413.txt, file415.txt, file422.txt,
file435.txt, file439.txt, file445.txt, file457.txt, file467.txt, file471.txt,
file473.txt, file474.txt, file478.txt, file483.txt, file494.txt, file506.txt,
file517.txt, file519.txt, file520.txt, file525.txt, file530.txt, file533.txt,
file536.txt, file541.txt, file549.txt, file551.txt, file554.txt, file586.txt,
file587.txt, file590.txt, file593.txt, file597.txt, file612.txt, file621.txt,
file624.txt, file630.txt, file635.txt, file637.txt, file638.txt, file663.txt,
file674.txt, file675.txt, file681.txt, file689.txt, file725.txt, file729.txt,
file737.txt, file744.txt, file755.txt, file770.txt, file798.txt, file807.txt,
```

```
file812.txt, file819.txt, file827.txt, file847.txt, file851.txt, file880.txt,
file890.txt, file894.txt, file895.txt, file907.txt, file908.txt, file924.txt,
file934.txt, file936.txt, file939.txt, file947.txt, file961.txt, file968.txt,
file973.txt, file982.txt, file990.txt
```

Question 3: Positional Index and Phrase Queries

- Approach:
    - The function initializes an empty set named `docs`. For the first word in the query, it retrieves the set of document identifiers (docIDs) where this word appears from the positional index and stores them in `docs`. This set represents the potential documents that could contain the phrase.
    - Then iterates over each word in the phrase query starting from the second word. For each word, it performs the following steps:
        - It intersects the current set of candidate documents (docs) with the set of documents containing the current word. This is done to narrow down the list of documents to those that contain all words seen so far.
        - For each document in this intersection, it checks if the current word follows the previous word in the sequence. This is determined by checking if any position of the previous word (`p`) plus one is a position listed for the current word in the same document. This step ensures that the words are not just present in the document but also follow each other in the correct order.
- Methodologies:
    - A temporary set `temp_docs` is used to collect documents where the current word follows the previous word in sequence.
    - For each document in the narrowed down set, it retrieves the positions of the previous word and checks if adding one to any of these positions matches any position of the current word in the same document. If this condition is met, it means the current word directly follows the previous word in this document, so the document is added to `temp_docs`.
    - After checking all documents, docs is updated to be `temp_docs`, effectively filtering down the set of documents to those that maintain the word order up to the current word in the iteration.
- Assumptions: The query will be searched for exact match and if any word(s) appear in between the original text, that result in different adjacent position but same relative position, the search query will be treated as unsuccessfull
- Results:

```
great buy
Number of documents retrieved for query 1 using positional index: 7
Names of documents retrieved for query 1 using positional index: file5.txt,
file105.txt, file167.txt, file214.txt, file283.txt, file712.txt, file906.txt
```