IT-550 INFORMATION RETRIEVAL

An Empirical Study of Tokenization Strategies for Biomedical Information Retrieval

`

Presentation By:-
Mallipeddi Akshay
201301216

Mentor :- Jainisha Sankhavara

Challenges Involved

Challenge 1

- Frequent occurrences of gene symbols in the given biomedical data.

Challenge 2

- Use of inconsistent lexical variants of same gene symbols. **Example**

Challenge 3

- The text also contains various names involving genes,proteins and chemicals.

Variation in tokenizers

Variant	Original Text	Tokenized Text			
		Tokenizer 1	Match ?	Tokenizer 2	Match ?
Query	MIP-1-alpha	mip 1 alpha	N/A	mip1alpha	N/A
Variant 1	MIP-1alpha	mip 1alpha	No	mip1alpha	Yes
Variant 2	(MIP)-1alpha	mip 1alpha	No	mip1alpha	Yes
Variant 3	MIP-1 alpha	mip 1 alpha	Yes	mip1 alpha	No

Steps

- Remove non-functional characters by following some **rules**.

- Even after following these rules, there are possibilities of occurrence of non-functional characters. **Example.**

- Finding **hidden places** in the text. Hidden places are places where the text can be further broken down.

Break Points(BP)

- **BP1:-** Contains ( ) [ ] { } - _ /
- **BP2:-**Contains the above characters and . : , ; +
- **BP3:-** Contains all special characters and hidden places as defined in the above slide.

Break Point Normalization

- **H-Norm**- Replace break points by hyphen (H)
  For example, MIP-1-alpha,MIP-1alpha and (MIP)-alpha will change to **MIP-1-alpha**.

- **S-Norm**-Replace break points by space (S).
  For above example, all will change to **MIP 1 alpha.**

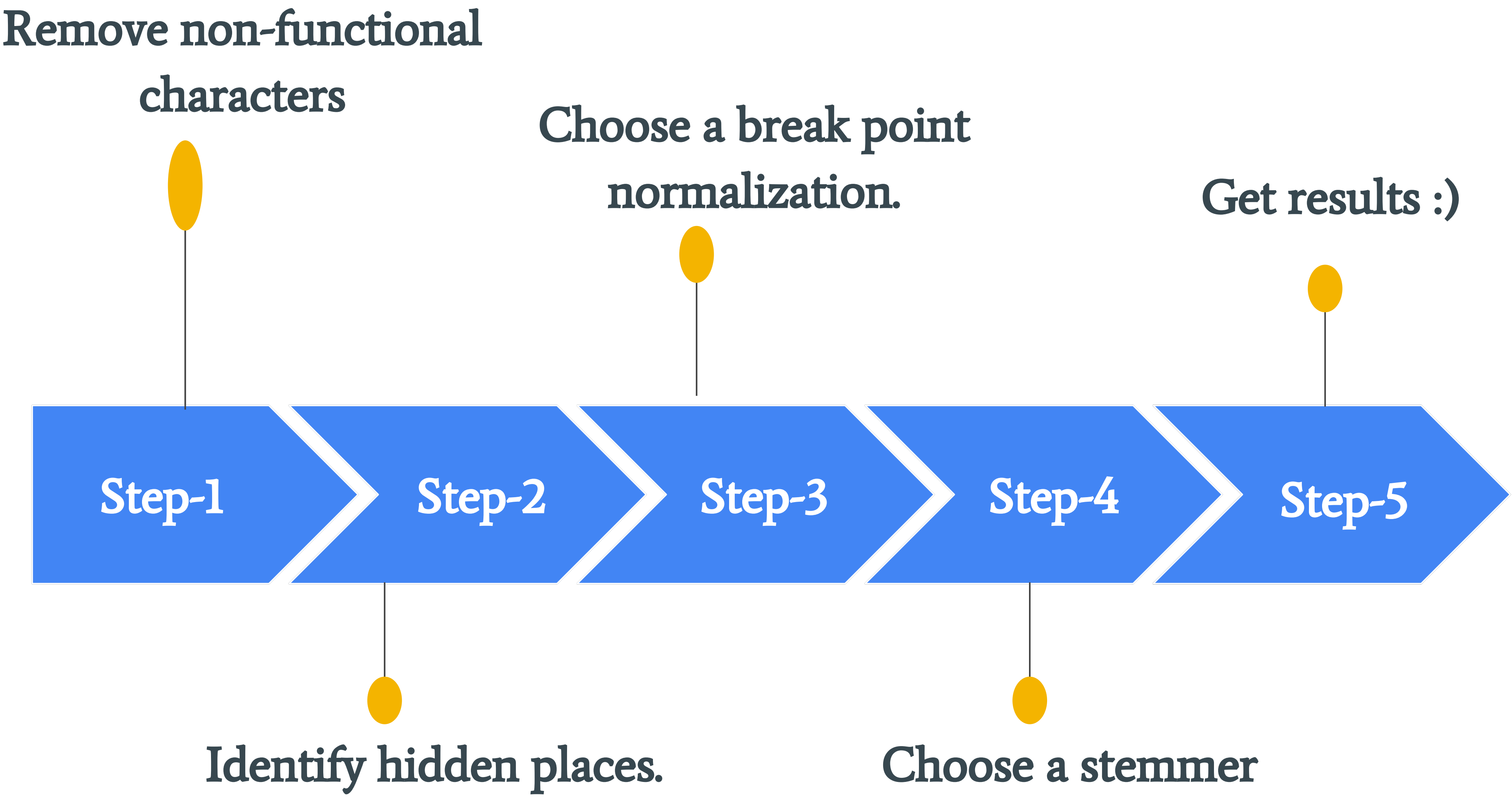
- **J-Norm**-Just apply the break points.

Stemming and Stop Word Removal

- **S stemmer :-** Removes common word endings.

- **Lovins Stemmer:-** Removes longest possible string of characters from a word.(Using a external stop word list).

- **Porter stemmer:-**Checks at vowels and consonant level.



Remove non-functional characters

Choose a break point normalization.

Get results :)

Step-1

Step-2

Step-3

Step-4

Step-5

Identify hidden places.

Choose a stemmer

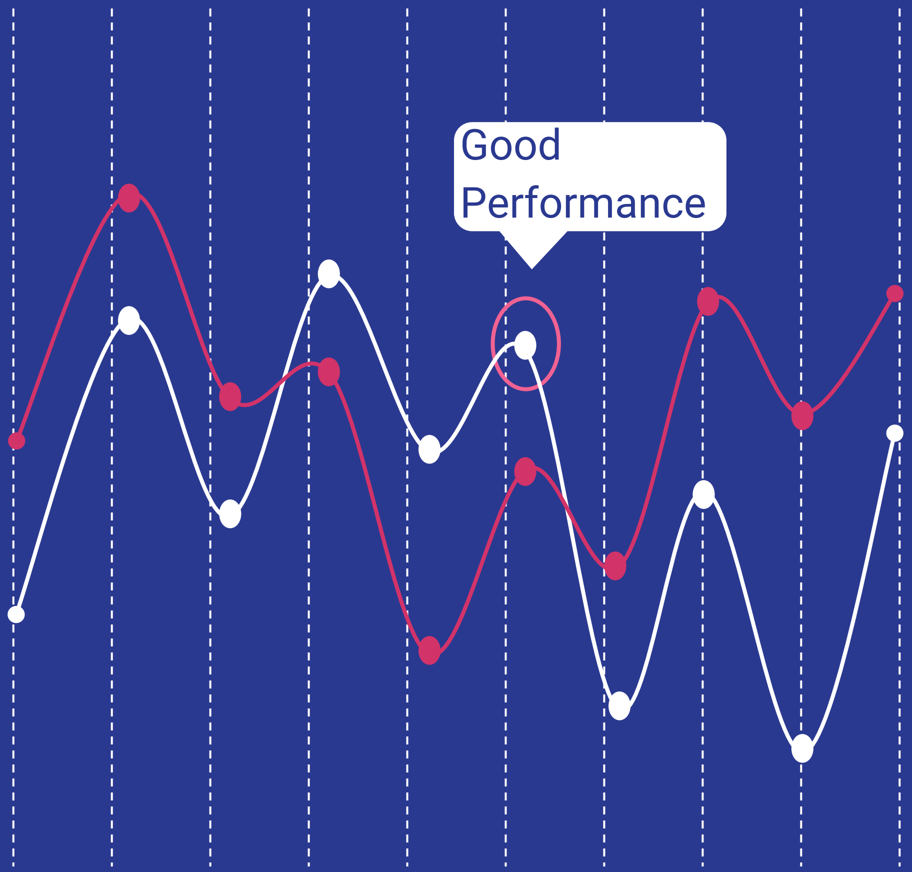
Data Set Used

- The data is taken from **CDS**(Clinical Decision Support) that was introduced in TREC 2014.

- The data contains 7,33,138 documents and 30 topics.

- Each topic consists a case report and one of three clinical question types (‘diagnosis’,’treatment’ and ‘test’).
- Used **tf-idf** retrieval model with terrier.

Results



Results

		NORMALIZATION METHODS		
STEMMER USED		H-NORM	S-NORM	J-NORM
PORTER	AVG PRECISION	0.0953	0.0954	0.0940
	R PRECISION	0.1531	0.1527	0.1565
LOVINS	AVG PRECISION	0.0673	0.0632	0.0613
	R PRECISION	0.1236	0.1242	0.1209
S	AVG PRECISION	0.1001	0.1000	0.0976
	R PRECISION	0.1590	0.1578	0.1581

THANK YOU

Example

- **MIP-1-alpha**
- **MIP-1 alpha**
- **(MIP)-1 alpha**
- **MIP-1alpha**

Removal of Non-Functional Characters

- Replace **! “ # $ % & * < = > \ | ~** with space
- Remove **. : ; ,** if followed by a space
- Remove the following pair of brackets if the open bracket is preceded by a space and the close bracket is followed by a space: **() [] {}**
- Remove single quotation (‘) if followed by a space
- Remove **‘s** and **‘t** if they are followed by space
- Remove **/** if it is followed by a space.

Characters occurring even after Step-1

- (MIP)-1alpha

Special character set 1	Special character set 2
( ) [ ] { } - _ /	. : ; ‘ +

Hidden Places

- Places between an alphabetical character(on left) and numerical character(on right).For example, between **akshay** and **216** in **akshay216.**
- Places between a lower case(on left) and an upper case(on right).For example, between **Lung** and **Cancer** in **LungCancer**.
- Places between an upper case letter(on left) and a lower case letter(on right) unless the upper case is preceded by a space or by a numeric.For example, between **MIP** and **alpha** in **MIPalpha**, but not between **E** and **b** in **TrpEb_1**.